

# Modeling Continuous Self-Report Measures of Perceived Emotion Using Generalized Additive Mixed Models

Gary J. McKeown and Ian Sneddon  
Queen's University Belfast

Emotion research has long been dominated by the “standard method” of displaying posed or acted static images of facial expressions of emotion. While this method has been useful, it is unable to investigate the dynamic nature of emotion expression. Although continuous self-report traces have enabled the measurement of dynamic expressions of emotion, a consensus has not been reached on the correct statistical techniques that permit inferences to be made with such measures. We propose generalized additive models and generalized additive mixed models as techniques that can account for the dynamic nature of such continuous measures. These models allow us to hold constant shared components of responses that are due to perceived emotion across time, while enabling inference concerning linear differences between groups. The generalized additive mixed model approach is preferred, as it can account for autocorrelation in time series data and allows emotion decoding participants to be modeled as random effects. To increase confidence in linear differences, we assess the methods that address interactions between categorical variables and dynamic changes over time. In addition, we provide comments on the use of generalized additive models to assess the effect size of shared perceived emotion and discuss sample sizes. Finally, we address additional uses, the inference of feature detection, continuous variable interactions, and measurement of ambiguity.

**Keywords:** emotion, mixed models, continuous variables, GAMMs, dynamic stimuli

For many years the predominant experimental paradigm within emotion research has been what Russell (1994) has referred to as the *standard method*—the use of static pictures of people displaying posed or acted facial expressions of emotion. The benefit of this approach is that it permits a high level of control over the emotional expression at the focus of investigation. However, this comes at a heavy price in terms of the level of ecological validity offered by such stimuli. It is true that many studies have shown high levels of recognition for emotions expressed in static images (Elfenbein & Ambady, 2002). Yet, these stimuli are far from the dynamic real time emotional stream of information that we experience in everyday interaction with other people, and that is likely to have formed the basis upon which the machinery of our emotional perception has evolved.

There is now considerable evidence that our perception of facial expressions of emotion is influenced by the presence or absence of dynamic information (Tcherkassof, Bollon, Dubois, Pansu, & Adam, 2007). Researchers have used a range of methods to compare dynamic and static images that has included dynamically morphing a sequence of still images (Kamachi et al., 2001),

animating synthetic images (Wehrle, Kaiser, Schmidt, & Scherer, 2000), and presenting film of spontaneous responses to emotional slides (Wagner, MacDonald, & Manstead, 1986). All of these studies have employed a recognition paradigm in which observers are asked, after viewing the stimulus, to judge what emotion was being displayed. Additionally, Biele and Grabowska (2006), using morphed sequences of still images, have asked observers to rate the intensity of emotion displayed, again after viewing the stimulus. Research that allows the observer to respond while viewing the stimulus, however, has been less commonly reported. Tcherkassof et al. (2007), using a forced choice paradigm, have recorded judgments of the presence and duration of emotions while facial expressions are being viewed, and, more recently, Sneddon, McKeown, McRorie, and Vukicevic (2011) have described results based on a method that permits continuous rating of intensity of emotion dimensions.

Continuous rating of the facial expression of emotion may have appeared rarely in the psychological literature, but its importance has been recognized in engineering and human computer interaction contexts during attempts to construct emotion recognition devices (Gunes & Schuller, 2013). Also in musical disciplines and the psychology of music, continuous self-report rating of the musical expression of emotion has been much more widely adopted. Continuous self-report measurement requires a participant to watch or listen to a stimulus and continuously report on the nature of their perception, feeling, or understanding of that stimulus as they are perceiving it—providing real-time feedback on their current state in response to the stimulus.

Rozin, Rozin, and Goldberg (2004) have emphasized the importance of continuous response in understanding the emotion induced by music and provide evidence that summative judgments

---

This article was published Online First November 11, 2013.

Gary J. McKeown and Ian Sneddon, School of Psychology, Queen's University Belfast, Belfast, Northern Ireland.

We thank Chris Edgar for permitting us to use data he collected for the continuous interactions section of the article.

Correspondence concerning this article should be addressed to Gary J. McKeown, School of Psychology, Queen's University Belfast, University Road, Belfast, Northern Ireland, BT7 1NN, United Kingdom. E-mail: g.mckeown@qub.ac.uk

(taken after the stimulus has finished) differ from those taken during the stimulus. However, Schubert (2010) discussed the difficulties that researchers face in developing statistical techniques that can cope with the fluctuations over time and inherent nonlinearity that characterize continuous self-report data. The kinds of data produced in these experiments involve changes over time, but they differ from more traditional longitudinal measurement such as that found in panel surveys and clinical, developmental, or educational studies, which typically have small numbers of trials or measurements over a long time period. Increased computational power and the pervasive nature of modern computing have led to new data-gathering techniques that can take many more meaningful measurements, often over short time spans, which has led to these kinds of data being labeled *intensive longitudinal data* (Walls & Schafer, 2006). These kinds of data, which have been common in affect-related research and in research where biosensor data are collected, are sometimes known as functional data (Ramsey & Silverman, 2005) and are likely to become increasingly common in other areas of psychology as psychologists begin to utilize modern smartphones as data gathering devices (Miller, 2012).

Some recent efforts to account for temporal issues in emotion research have aimed to enable new types of research questions to be addressed using intensive longitudinal data (Oravecz, Tuerlinckx, & Vandekerckhove, 2011; Tan, Shiyko, Li, Li, & Dierker, 2012). A range of techniques that can be used to address issues of nonlinearity and that can fit nonlinear regression models or provide alternatives to simple and multiple linear regressions include the following: classification and regression trees (CART; Finch, Chang, & Davis, 2011; Strobl, Malley, & Tutz, 2009), neural network modeling techniques (Marshall & English, 2000), and functional data analysis (FDA; Levitin, Nuzzo, Vines, & Ramsay, 2007). Careful consideration should be given to the choice of technique. Often an inertia driven by the investment in time required to learn new techniques means researchers stay with familiar, tried and tested techniques; yet, there are many techniques that offer better performance and greater statistical power. Care must also be taken concerning the appropriateness of a given technique for the data; a good comparison of some of these techniques in application to the prediction of intelligence is given in Finch, Chang, Davis, Holden, et al. (2011).

Studies incorporating continuous self-report measures often involve recruitment of a number of participants to observe or listen to some stimuli; typically they move a mouse/slider/stylus in a one or two dimensional space to report on their perceived emotion at a given instance in time. There is, however, an important statistical issue that stands in the way of broad adoption of the techniques; the issue concerns the non-independence of the data—this violates the basic assumption of generalized linear models that the residuals of a fitted model will be independent and randomly distributed. It occurs in two important and related ways in continuous self-report measures: First, the data are correlated within-participant—they are equivalent to a series of rapid repeated measures; second, they are autocorrelated (sometimes referred to as serial correlation)—the measurement at one point in time strongly depends on the measurement at the previous point in time. In addition to these statistical issues, we would like to be able to include information about the individual differences between participants in any model in a satisfactory manner. Both the within-participant

correlation and individual differences can be accommodated by moving to a mixed model framework and treating the participant measurements as random effects components in a model. This views the participants as a sample from a population—an assumption that is usually valid. It also ensures within-participant information is incorporated in the model, and variability due to individual differences is estimated as a variance component. Autocorrelation occurs as the data are collected in time order without resetting to the same initial position for each measurement. The placement of a measured point at time  $t$  will in part depend on the placement of the previous data point  $t - 1$  and in part depend on responses to new information available at time  $t$ . This is obvious in the emotion perception example given above where the physical placement of a mouse or slider cannot be returned to the initial or zero point used in the first measurement when measurements are taken 10 times a second. Thus, we require analysis techniques that can account for individual differences, within participant correlation, and autocorrelation inherent in data gathered by the same participants over time.

Standard linear regression techniques are well known and understood by most behavioral researchers, and they are easy to fit to the kinds of data sets gathered in behavioral research paradigms. There are many cases where it is unsafe to assume linearity, but often these cases can be handled by data transformations or polynomial regression models that can easily deal with a subset of curvilinear relationships within the linear model framework. Where these are applicable they are the preferred option. However, when changes over time become more complex, models that can provide the flexibility to adequately describe these often nonlinear relationships are required. Returning to our example of continuous self-report measures in emotion research, if the instructions to a participant explicitly tell them to move a slider back and forth in response to a dynamic stimulus the resulting data may take the form of complex dynamic trajectories over time that require more flexible modeling techniques. There are certainly occasions when linear relationships occur, for example, observing a steadily increasing smile. However, in most emotional expressions there is an onset, emotional apex and offset, and often more than one apex. In these situations, a parametric model would be better replaced with nonparametric regression models that offer more flexibility to capture dynamic changes over time.

Simple nonparametric regression techniques use *loess* (local regression) or *lowess* (locally weighted scatterplot smoothing) smoothing (Cleveland, 1979) or smoothing splines (which are discussed in detail later) to capture the relationship between two variables. These are popular techniques in what has become known as *scatterplot smoothing*. Scatterplot smooths start with a plot displaying data points placed on a plane defined by two variables. The scatterplot smooth seeks to illuminate an underlying trend between the two variables. These techniques are addressed in more detail in the section on smoothing. When relationships become more complex and involve more than two variables, trends identified by scatterplot smoothing are less useful. A type of model proposed by Hastie and Tibshirani (1990), called additive models or generalized additive models (GAMs), has become popular because it addresses these issues. In what follows, we provide a brief introduction to GAMs but leave a more technical explanation of GAM methods and their mixed model form generalized additive mixed models (GAMMs) to a later section.

GAMs have proved popular because they offer greater flexibility than scatterplot smoothing and they can be used in a manner very similar to linear regression. Over the past decade, increasing use has been made of GAMs in disciplines that often rely on temporal data, notably in ecological modeling (for an overview in relation to species distributions, see Guisan, Edwards, & Hastie, 2002); atmospheric modeling, for example, assessing trends in traffic-related emissions (Carslaw, Beevers, & Tate, 2007); and health applications, such as allowing inference in computer-aided diagnosis systems based on the interaction of nonlinear and continuous variables with categorical variables (Lado, Cadarso-Suárez, Roca-Pardiñas, & Tahoces, 2008), or adjustment for nonlinear confounding effects over time in risk of death from overdose (Bohnert, Prescott, Vlahov, Tardiff, & Galea, 2010).

The flexibility offered by GAMs allows the use of smooth nonlinear functions—often called nonparametric functions—in the same model as parametric terms. These parametric terms can describe linear relationships or can be dummy variables describing categorical and binary variables, and there are also ways of accommodating interactions. This ability to combine nonparametric and parametric terms has led to additive and generalized additive models being termed *semiparametric models* (Ruppert, Wand, & Carroll, 2003), and it makes them particularly useful in the analysis of continuous self-report measures of perceived emotion. The additive in the name of generalized additive models refers to the assumption of additivity; this is one of the features that makes them attractive. This means that the joint effect on the response variable of all the predictor variables is determined by the sum of all the individual effects of the predictor variables on the response variable; each predictor variable operates independently of the others, and so it does not matter at what level the other predictor variables are fixed. As well as allowing inferences to be made regarding changes over time, these models allow for the temporal information to be accounted for so that sound inferences can be made on other variables included in the model in an independent way. This is preferable to the practice of ignoring or collapsing temporal information in emotion perception studies, an unsatisfactory practice that can lead to unsound inferences. The assumption of additivity and independence of predictor terms that it implies mean that interactions between predictors need to be treated with care. It is prudent to check the assumption of additivity; if there is evidence for an interaction, then it is wise to treat the two interacting predictor terms within a single term in the additive model. More detail on this is provided in the interactions section.

Another common goal in emotion research is to find answers to questions about group differences or similarities in response to emotion stimuli without necessarily asking questions about the processes that led to the perceived emotion. In these circumstances, the single univariate parameter provided by static images means analysis of variance (ANOVA) or linear regression designs serve well. With more complex dynamic natural stimuli that play out over time, such designs would require the collapsing of time series data to a single parameter such as the mean. As has been pointed out, this is unsatisfactory; it wastes the temporal information in the data, has costs in terms of statistical power, and may lead to incorrect inferences. In such circumstances, semiparametric models offer the ability to account for the changes in the dynamic stimuli over time while also retaining the ability to ask about linear differences between groups.

This article aims to introduce GAMs and GAMMs as methods, with a focus on analyzing continuous self-report measures of perceived emotion. The data sets used in the article—introduced in the following section—are all drawn from perceived emotion studies but the techniques are broadly applicable to many research domains in psychology. For example, GAMs or GAMMs could be used in longitudinal and panel studies, signal detection (Knoblauch & Maloney, 2008), eye tracking and gaze tracking studies, modeling evoked response potentials, integrating psychological and geographical components (Wieling, Nerbonne, & Baayen, 2011), and in situations in which non-linear relationships between two variables are suspected. The goal of this article is to make researchers aware of these techniques and the new possibilities for analyses that they permit rather than to present an in depth mathematical treatment of the GAM method. However, we attempt to provide enough technical detail to allow an intuitive understanding of the issues; this starts with a section on smoothing that covers many of the basic concepts before addressing GAMs and GAMMs in more detail in the sections that follow. The perceived emotion studies are used to address the importance of dealing with correlation within the data, and issues of effect size and sample size. It highlights some techniques for making inferences in these models, concentrating on showing how smoothed nonparametric components can be used in combination with continuous and categorical linear covariates to make inferences about differences between groups and their interactions.

## Data

For purposes of illustration, we use three data sets. These are comprised of continuous self-report ratings of emotion perception. The data were collected from individuals watching short video clips of people exhibiting behavior associated with emotional states. The first uses estimated ratings of the emotional valence—the degree of positive or negative emotion displayed—from six clips from Set 1 of the Belfast Induced Natural Emotion Database (Sneddon, McRorie, McKeown, & Hannraty, 2012) in which 159 or 160 decoders (those perceiving the emotion and providing the ratings while observing the clip) provided continuous ratings for females encoding emotion expressions (encoders are the people recorded, the focus of the video clip). The behavior of encoders was recorded while they experienced an emotion induction procedure designed to elicit either amusement or disgust and the recordings used as stimulus clips in this study are all 30-s extracts. Three clips (2, 4, 6) show encoders exposed to the amusement inducing procedure, and three clips (1, 3, 5) show encoders exposed to the disgust inducing procedure. In the amusement inducing procedure, encoders watched an amusing film, an extract from an episode of *Father Ted*—a well-known television comedy series in the United Kingdom and Ireland (Baker, Shortt, Perkins, & Lowney, 1996). The 30-s extracts used as stimulus clips show encoders watching the same segment of the film, the start point selected on the basis of a cue on the soundtrack. In the disgust inducing procedure, encoders were presented with a black box with a 10-cm diameter hole cut in the top. Encoders were asked to reach into the box which contained a bowl of cold, cooked, cut spaghetti in sauce. The 30-s extracts used as stimulus clips show encoders from the point when their hand entered the box. Often in these emotional videos, sample rates of up to 50 frames per second are used for

capturing emotional expressions and even higher sample rates are used in audio analyses. In our clips, we have opted for a sample rate for the continuous self-report ratings of 100 ms—the videos were played back at 25 fps (for a useful discussion of the issues surrounding sample rate, see Schubert, 2010). This sampling rate gives 300 data points for each individual trajectory, and for most clips in Set 1, there are 160 participants giving 48,000 data points in the model. The details for the first data set are given in Table 1. This data set was mainly chosen to illustrate these procedures due to its large sample size.

The second data set (see Table 2) uses estimated ratings of the emotional valence from six different clips from Set 1 of the Belfast Induced Natural Emotion Database (Sneddon et al., 2012). For each of the six clips in this data set, 40 decoders provided continuous ratings of males and females encoding emotion expressions. The stimulus clips in this data set consisted of recordings of female (Clips 7 and 8) and male (Clip 9) encoders exposed to the amusement inducing procedure described above and recordings of female (Clips 10 and 12) and male (Clip 11) encoders exposed to the disgust inducing procedure described above. This data set was chosen for use as it contains a balanced sample of males and females and therefore allows the addition of a simple categorical variable to the models.

The third data set involved data from just one clip from Set 1 of the Belfast Induced Natural Emotion Database (Sneddon et al., 2012). In this data set, 86 decoders provided continuous ratings for a male stimulus clip in which a male encoder was exposed to a fear inducing procedure. In this procedure, the experimenter carefully placed a black box in front of the encoder. The experimenter moved slowly and spoke quietly while moving and touching the box which had a hinged metal grill on top and was covered with several warning icons and images of spiders. The decoders also completed the NEO-FFI Big Five personality scale (McCrae & Costa, 2004), and the agreeableness dimension was used in this data set. This data set was chosen as it contains a combination of continuous time series data and continuous variables from the Big Five personality scale, permitting a simple illustration of an interaction between continuous variables within the GAMM framework.

The stimulus clips were shown without sound to decoders on a laptop computer using a variant of a computer logging program called GTrace (Cowie, McKeown, & Douglas-Cowie, 2012). A 10 cm × 10 cm window containing the stimulus clip appeared on screen alongside an interactive horizontal scale (Sneddon et al., 2011). Decoders used the mouse to move a colored spot along the scale to trace their changing judgment of the intensity of the emotional expression of the target individual. The bi-directional

Table 2  
Data Characteristics for Data Set 2

Clip	Emotion	Set	<i>N</i>	Time (s)	Hz	Data points
7	Amusement	2	40	30	10	12,000
8	Amusement	2	40	30	10	12,000
9	Amusement	2	40	30	10	12,000
10	Disgust	2	40	30	10	12,000
11	Disgust	2	40	30	10	12,000
12	Disgust	2	40	30	10	12,000

scale was anchored at the left end by the text *very strongly negative* and at the right by *very strongly positive* with a central *neutral*. Decoders were instructed to use the computer mouse to move the dot along the scale to “indicate how strongly you think the person in the video clip is expressing either positive or negative emotion.”

Smoothing

A key concept in understanding GAMs and GAMMs is the idea of smoothing; one of the key features of these models is that they permit the incorporation of smooth components within a modeling framework very similar in nature to the linear model and linear mixed model frameworks. Smoothing seeks to illuminate the trend in one variable as a function of one or more other variables, typically in a nonparametric way. Following Hastie and Tibshirani (1990), we call the techniques that do this *smoothers* and the resultant trend line a *smooth*. The goal and usefulness of smoothers is often simply to provide a description of relationships that may not be linear in nature. This can be as simple as producing a scatterplot smooth or, alternatively, can seek to find a function to replace the regression slope. This function would describe the change required to accommodate the irregular nature of the relationship between a predictor variable and a response variable—a nonparametric mean function. Smooths can serve these descriptive purposes, or, as mentioned, smoothing functions can have an important role as components in models—such as GAMs and GAMMs—where the dependence of the mean of the response variable on one or more predictors is nonlinear in nature.

The choice of a smoothing technique requires consideration of the bias/variance tradeoff—a common issue in the statistical learning literature (Briscoe & Feldman, 2011; Hastie, Tibshirani, & Friedman, 2001). This is a complex issue that we can summarize here by the terms *underfitting*—representing high bias—which involves a model that is too simple and inflexible, and *overfitting*—representing high variance—which fits a model too closely to the sample data. A smooth seeks to minimize the variance in some optimal way without the introduction of bias. The extreme case of underfitting would take a straight line at the mean to represent the relationship on a scatterplot—a high bias model; extreme overfitting would involve interpolating every data point—a high variance model.

We illustrate this relationship with an explanatory example that can apply to many statistical situations in psychology including the time series data in emotion perception studies. Hastie and Tibshirani (1990) pointed out that one of the simplest smoothers is a categorical predictor variable; although not normally thought of as a smoothing technique, it does involve taking the average of

Table 1  
Data Characteristics for Data Set 1

Clip	Emotion	Set	<i>N</i>	Time (s)	Hz	Data points
1	Disgust	1	160	30	10	48,000
2	Amusement	1	160	30	10	48,000
3	Disgust	1	160	30	10	48,000
4	Amusement	1	160	30	10	48,000
5	Disgust	1	159	30	10	47,700
6	Amusement	1	160	30	10	48,000



predefined categories which creates a smooth of the data over two or more levels of the categorical variable as they relate to the response variable. A technique similar to the categorical predictor example is the ill-advised but still commonly used statistical practice of the dichotomization of continuous variables by median split (MacCallum, Zhang, Preacher, & Rucker, 2002); this process can be thought of as a poor smoothing technique. The problem with such a “smooth” is the loss of much of the information and statistical power in the continuous variable for reasons that are difficult to defend (DeCoster, Iselin, & Gallucci, 2009). Splitting a continuous variable into two big groups includes more information, and produces a slightly better model, than the simple mean—high bias—model outlined previously. This process can be viewed as the crudest form of one of the simplest types of smoother, the bin smoother. In more concrete terms, a valence response variable could be paired with a start phase and end phase as two levels of a categorical predictor variable in time series data. Alternatively, it could be a dichotomized personality variable like agreeableness giving low agreeableness and high agreeableness as two levels of the predictor variable. The variance of the variable has been reduced by averaging the values of the variable in two large “bins,” but a source of bias has been introduced by using this arbitrary categorization—the model underfits the data due to the choice of a bad model. Further splitting the variable into equal bins of four and then eight gives smooths with increased variance and less bias. Finally, if we further subdivide the bins all the way until they contain just one instance of the predictor variable we return to the original data; interpolating all this original data would result in the high variance and overfitted model mentioned previously.

In these examples, if we imagine that the underlying relationship between the two variables is linear and the residuals are normally distributed then the averages in the bins may approximate a linear mean function—the regression slope—with the distances between the bin means defined by a parametric constant—the  $\beta$  coefficient. We can get from one bin average to another by adding the parametric constant. However, if the relationship is not linear then we would expect this nonlinearity to be picked up by this technique and the distances between the bin means would not be similar but irregular and not easily definable by a single parametric constant; in this case, the bin averages when joined would produce a curvilinear or jagged line on a scatterplot.

There are many types of smoothers; the simplest cases—like the previous example—involve just one predictor variable, and as these can be illustrated easily on a scatterplot, the process is called *scatterplot smoothing*. To create these smooths, one must typically determine the size of the neighborhood or bandwidth that will influence the smooth—this would be the width of the bin in a bin smoother. The parameter that defines the neighborhood is typically known as the smoothing parameter. Additionally, one must determine how values of the response variable will be averaged within these neighborhoods. Decisions regarding these choices influence the fundamental trade-off between bias and variance. Informally a large neighborhood will reduce the variance but increase the bias while a small neighborhood will result in increased variance but reduced bias.

A bin smoother represents a crude approach. Cutoff points are chosen according to some criteria—equal divisions of the data in the previous example—and the response variable is averaged within these points; this leads to jumps at the boundaries between

bins and it is rarely a safe assumption that such discontinuities exist in the data. These abrupt changes can be overcome by allowing the cutoff points to overlap as in a moving average or running mean smoother; here, an average for a point is calculated from a set number of nearest neighbors above and below the current point—the neighborhood. This has the benefit of simplicity but an obvious bias is introduced by these methods in the way they deal with the endpoints, where the neighborhoods must be shortened due to a lack of data at the endpoints. The importance of this issue depends on the importance of the endpoints in an analysis and can be mitigated to some extent by the choice of neighborhood size. An improvement can be made by calculation of a least-squares line within the neighborhood instead of the mean, creating a running-line smooth; this improves the problem of endpoint bias but often produces jagged lines due to discrete jumps, as points outside the neighborhood do not contribute to the least-squares line and those inside contribute equally. This issue can be overcome by giving more weight to the points closest to the central point within a neighborhood and less weight to those at the edges; this is the basis of the popular locally-weighted running-line smoother more commonly known as loess.

Smoothers, such as the ones described so far, assist in recognizing and describing nonlinear relationships between two variables. When it comes to creating models that deal with nonlinear relationships between the predictor and response variable, two important parametric methods should be considered. The first is to transform either the response or the predictor in some way to make the relationship a linear one, typically using square root, log, or reciprocal transformations. In the first instance, it might prove simpler to adopt an approach based on the use of transformations. However, where complex nonlinear relationships exist, the utility of transformations is limited. Another important approach is polynomial regression. It creates a nonlinear model, but it is linear and parametric in the relationship of the  $\beta$  coefficients of the predictor variables with the response variable, as can be seen in the quartic polynomial regression shown in Equation 1:

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \beta_3 x_i^3 + \beta_4 x_i^4 + \varepsilon_i$$

$$\varepsilon_i \sim N(0, \sigma_\varepsilon^2), \quad (1)$$

where  $y_i$  is the response variable—valence in the emotion perception data;  $\beta_0$  is the intercept;  $x_i$  is a predictor variable—time coded at one point every 100 ms in the emotion perception data; and  $\varepsilon_i$  is the error term which is assumed to be independent and normally distributed ( $N(0, \sigma_\varepsilon^2)$ ). This is a linear model that requires finding suitable  $\beta$  coefficients for the intercept and each of the polynomial terms in Equation 1; so in this example, it involves an intercept and four separate  $\beta$  coefficients for up to the 4th order polynomial term. This will produce a curvilinear line that may represent certain data sets, and it has the very useful property of being linear in the parameters. However, it is of limited use as it is global in nature, meaning that it deals with all the data at once as opposed to the local neighborhood approach of a loess smoother. It requires the specification of the order polynomial—a specification that can be guided by a criterion (e.g., variance explained, or an information criterion such as the Akaike information criterion [AIC]). However, these factors reduce its flexibility and utility in situations where it is difficult to know the form a curve may take ahead of fitting a model.

A final important class of smoothing techniques are called splines which make special use of local polynomial regression. Regression splines offer a more flexible solution by using piecewise polynomial regressions over local neighborhoods joined together at points known as knots. The nature of these splines and the selection of neighborhood sizes using knots is dealt with in more detail in the following section.

### Generalized Additive Models

Here, we provide only a brief overview of the mathematical background and cover the basics necessary to understand GAMs. An accessible (for the mathematically proficient social or behavioral scientist) review is provided by Andersen (2009), and detailed mathematical accounts of GAMs and related techniques can be found in Hastie and Tibshirani (1990, 1993), Wood (2000, 2003, 2004, 2006a, 2006b, 2006c, 2008, 2011), and Ruppert et al. (2003). Throughout this article, the software used is the R package mgcv; a list of available R packages that provide GAM functionality is provided in Appendix A.

A GAM is a generalized linear model in which at least one of the linear predictors is made up of a sum of a smooth function of one—typically time in the emotion perception context—or more predictor variables. This is achieved by assuming that such a smooth function can be approximated by a linear combination of a given number of basis functions. In short, GAMs provide a means by which we can substitute a function for a parameter estimate with the effect of making the estimate not a single scalar number but a smooth nonparametric function.

If we write the typical regression function as

$$\begin{aligned} y_i &= \beta_0 + \beta_1 x_i + \varepsilon_i \\ \varepsilon_i &\sim N(0, \sigma_\varepsilon^2), \end{aligned} \quad (2)$$

then the simplest GAM takes the form:

$$\begin{aligned} y_i &= \beta_0 + f(x_i) + \varepsilon_i \\ \varepsilon_i &\sim N(0, \sigma_\varepsilon^2). \end{aligned} \quad (3)$$

In these two equations,  $y_i$  is the response variable—this could be valence;  $\beta_0$  is the intercept;  $x_i$  is a predictor variable—typically time in the emotion perception examples; and  $\varepsilon$  is the error term which is assumed to be independent and normally distributed ( $N(0, \sigma_\varepsilon^2)$ ). The important difference is that the parameter  $\beta$  in Equation 2 is replaced by a smoothing function  $f$  in Equation 3.

To estimate the smooth function  $f(x)$  it must be represented as a linear model, and this requires choosing a suitable basis. In the case of a simple linear regression model such as Equation 2 with observations 1 to  $i$ , the basis functions are 1 and  $x$  such that the design matrix  $\mathbf{X}$  will look like the following:

$$\mathbf{X} = \begin{bmatrix} 1 & x_1 \\ \vdots & \vdots \\ 1 & x_i \end{bmatrix}. \quad (4)$$

When we wish to estimate the smooth function, we similarly require a suitable set of basis functions that define the space of functions of which  $f$  is an element. One possibility here is to use a polynomial basis similar to that used in Equation 1 such that

$$f(x) = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3 + \beta_4 x^4. \quad (5)$$

This would then result in a model equivalent to Equation 1 and a correspondingly more complex model design matrix:

$$\mathbf{X} = \begin{bmatrix} 1 & x_1 & x_1^2 & x_1^3 & x_1^4 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & x_i & x_i^2 & x_i^3 & x_i^4 \end{bmatrix}. \quad (6)$$

However, as mentioned, such polynomial regressions suffer from inflexibility and, for this reason, using cubic splines as a basis is a preferable option; these similarly allow  $f(x)$  to be estimated in such a way that it can become part of a linear model, but the representation is much more complex than the example shown in Equation 5. Ruppert et al. (2003) worked through an example in some detail using linear spline basis functions that provides a good intuitive understanding of how spline smooths are constructed.

### Smoothing and Knot Selection

Smooths in GAMs are typically created using a cubic spline basis (although see Ruppert et al., 2003, for a range of other basis options); cubic spline curves are piecewise combinations of smaller cubic polynomials joined at points called knots. Cubic splines are particularly useful as they allow the generation of smooth curves by forcing the piecewise components to join as smoothly as possible at the knot. The data collected in perceived emotion studies seldom has abrupt shifts that cannot be handled by smooth curves based on cubic splines.

Once again we come across the fundamental tradeoff between reduction in variance and bias introduced by the technique. We want to ensure a good fit to the data but also to avoid overfitting, so choosing the correct smoothing parameter and selecting the placement of knots are fundamental issues in GAMs. In creating the larger curve each piecewise section is comprised of a basis function from a cubic regression spline basis multiplied by a  $\beta$  parameter. The design matrix for the model then consists of the sequence of cubic sections separated by the knots. The  $\beta$  parameters must be estimated and the model is now treated as an ordinary linear model; this is achieved using a design matrix of the type shown in Equation 7:

$$\mathbf{X} = \begin{bmatrix} 1 & x_1 & (x_1, x_1^*) & (x_1, x_2^*) & (x_1, x_3^*) \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & x_i & (x_i, x_1^*) & (x_i, x_2^*) & (x_i, x_3^*) \end{bmatrix}, \quad (7)$$

where we use  $(x_1, x_1^*)$  to represent the basis function that corresponds to the knot at point  $x_1^*$  using the  $*$  to denote that it is a knot and  $k$  is the number of the knot. The smooth term function can then be represented as a sum of the basis functions:

$$f(x) = \beta_0 + \beta_1 x + \sum_{k=1}^K b_k(x, x_k^*), \quad (8)$$

where  $\beta_0$  and  $\beta_1$  are equivalent to the linear basis functions in the simple linear model design matrix in Equation 4, and the  $(x, x_k^*)$  are the new basis functions required for the spline, and  $b_k$  is a vector of  $\beta$  coefficients that must be estimated for the basis functions. The smooth function  $f(x)$  is now linear in the parameters in the same way that a polynomial regression is linear in the parameters.

The number of basis functions used, and therefore parameters to be estimated, depends on the number of knots used. This means that knot placement decides the number of bends or “wiggleness” of the smooth and the points at which the bends should occur. One option for knot placement is to manually provide the optimal placement and fit by ordinary least squares methods; this is the “regression spline” approach. The “smoothing spline” approach starts with a very large number of knots and therefore parameters—often as many parameters as there are observations—and these are then reduced to an appropriate number by deletion, which is computationally inefficient. Another way of describing this deletion is that some of the coefficients associated with these knots are shrunk to zero while the others are left intact. Eilers and Marx (1996) proposed an alternative method that represented a sort of compromise between the regression and smoothing spline approaches. We start by slightly overfitting the function; this is done by making an estimation of slightly more knots than is thought necessary and imposing a penalty that shrinks all of the coefficients toward zero. This is in effect an ordinary least-squares fit with a “wiggleness” or roughness penalty associated with it, and the roughness penalty is controlled by a parameter  $\lambda$  that is the smoothing parameter. The large number of initial knots lessens the need for precise placement of the knots, but the computational efficiency is much greater than the smoothing spline method as we only need to start with slightly more knots than needed. These splines are known as penalized splines or *P*-splines.

A result of using penalized splines is that the issue of finding a good tradeoff between fit to the data and smoothness reduces to an estimation of the smoothing parameter  $\lambda$ , where if  $\lambda = 0$ , all data are interpolated, and  $\lambda \rightarrow \infty$  leads to a constant slope. One way of dealing with this issue is to use cross validation where each datum is omitted in turn and a new model is fitted to the remaining data; the position of the omitted datum is predicted from the new model, and the squared difference between the datum and its predicted value is calculated. The smoothing parameter is chosen to minimize the average squared difference between all the data and the predicted values when they are left out. This lessens the risk of overfitting as the difference is calculated from a predicted variable rather than an estimated one. In practice, the problem of finding a suitable  $\lambda$  parameter in *mgcv* defaults to estimation by a generalized cross-validation technique that represents an efficient compromise between performance and computationally prohibitive alternatives (Andersen, 2009; Wood, 2006a).

### GAMMs and Autocorrelation

Two related problems arise in using continuous perceived emotion data that need to be addressed if GAMs are to be useful; these concern the independence of the residuals. Simple GAMs assume that the data are independent yet the data are correlated within-participant. There is likely to be less variability within an individual participant than between the participants; therefore, there is structure at the level of the individual decoder that is not accounted for by a simple GAM. In a related issue, a further source of error arises as the residuals cannot be assumed to be independent because they are autocorrelated. It is important to note that the problem is not with autocorrelation in the time series itself but autocorrelation in the residuals that leads to violation of the independence assumption (Huitema & McKean, 1998). The model

could be improved by incorporating this structure within the model rather than assuming it can be accommodated in normally distributed random error. An important risk associated with both these issues is the likelihood of underestimating the standard errors, which equates to an overestimation in the level of precision we can claim for an estimate. Consequently, any inference based on these estimates is not safe. A better model can be achieved using a generalized additive mixed model with autocorrelated errors (Lin & Zhang, 1999; Wood, 2006a). GAMMs are an extension of generalized linear mixed models (GLMMs) to include a smooth function that allows smooth terms to be treated as random effects. The move to a mixed model has a number of desirable properties in the case of perceived emotion data. In addition to the treatment of smooths as random effects, it allows the decoders to be treated as a sample from a population—which is much more in keeping with our theoretical viewpoint—rather than a fixed population, and facilitates the handling of autocorrelation of the residuals. In matrix notation, the linear model in Equation 2 is typically written as

$$\begin{aligned} \mathbf{y} &= \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \\ \boldsymbol{\varepsilon} &\sim N(0, \mathbf{I}\sigma^2), \end{aligned} \quad (9)$$

where  $\mathbf{y}$  is the response vector,  $\mathbf{X}$  is the model design matrix (as in Equation 4),  $\boldsymbol{\beta}$  is the  $\beta$  coefficient vector, and  $\boldsymbol{\varepsilon}$  is the error term. In a mixed model this extends to

$$\begin{aligned} \mathbf{y} &= \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u} + \boldsymbol{\varepsilon} \\ \mathbf{u} &\sim N(0, \boldsymbol{\psi}_\theta) \\ \boldsymbol{\varepsilon} &\sim N(0, \boldsymbol{\Lambda}\sigma^2), \end{aligned} \quad (10)$$

where  $\mathbf{u}$  contains a random effects vector, and  $\mathbf{Z}$  is a model matrix for these random effects,  $\boldsymbol{\psi}$  is the covariance matrix, and  $\theta$  the unknown parameters within that covariance matrix.  $\boldsymbol{\Lambda}$  is a matrix that is part of the error term and can be used to model the residual autocorrelation but is often just the identity matrix  $\mathbf{I}$ , in which case it is no different from the error term in Equation 9.

Ruppert et al. (2003) emphasized the closely related nature of penalized splines with mixed models. In these cases, the smooth components of a GAMM become ordinary components of a generalized linear mixed model (GLMM) and are estimated using ordinary GLMM software (*lme* from the R *nlme* library in the case of *mgcv*). Equation 10 is simply the standard linear mixed effects model, but in the GAMMs case, the model design matrix  $\mathbf{Z}$  would take the more complex form of the design matrix incorporating knots shown in Equation 7. The vector  $\mathbf{u}$  corresponds to the random effects coefficients used to estimate the amount of smoothing. Importantly, the smoothing parameters,  $\lambda$ , are now treated as variance components  $\theta$  within the covariance matrix  $\boldsymbol{\psi}$  and are estimated as the ratio of the variance of the random effect  $\sigma_u^2$  and the variance of the errors  $\sigma_\varepsilon^2$ . These estimations use maximum likelihood (ML) or restricted maximum likelihood (REML) techniques for additive mixed models or use penalized quasi likelihood (PQL) techniques for the generalized additive mixed model case (estimated in *mgcv* using *glmmPQL* from the R *MASS* library). In essence, the fixed effect parts of a smooth, the unpenalized components, are added to the fixed effects model matrix  $\mathbf{X}$ , and the random effects parts, the penalized components, are added to the random effects model matrix  $\mathbf{Z}$ , and the model parameters are estimated as an ordinary GLMM.

These GAMMs can get complex very quickly and are unwieldy when written in their full linear form. Here, we present a simple illustrative model. This model presents a GAMM with a global smoothing function and a random intercept. The smoothing parameters and random intercepts are estimated variance components of a mixed model, and each participant possesses a separate function. The function or smooth is the same for each participant and differs only in the intercept; in the terminology of Gelman and Hill (2007), it would be a varying-intercept model. Later in the categorical interactions section we examine a type of model known as a varying-coefficient model that allow smooths to vary levels of a factor variable. There are also models that are equivalent to the varying-intercept varying-slope models in the terminology of Gelman and Hill, where each individual has a specific smooth; these are known as subject-specific curves and are addressed in detail by Durbán, Harezlak, Wand, and Carroll (2005). The illustrative model takes the following form:

$$\begin{aligned} \text{valence}_{ij} &= U_j + f(x_{ij}) + \epsilon_{ij} \\ U_j &\sim N(0, \sigma_U^2) \\ \epsilon_{ij} &= \rho\epsilon_{j,i-1} + \xi_{ij} \\ \xi_{ij} &\sim N(0, \sigma_\xi^2), \end{aligned} \quad (11)$$

where  $\text{valence}_{ij}$  is the response variable for individual  $j$  at time  $i$ .  $U_j$  represents the random effect intercept which is assumed to be normally distributed  $N(0, \sigma_U^2)$ .  $f(x_{ij})$  is the smooth term specifying the expected valence for individual  $j$  at time  $i$ .  $\epsilon_{ij}$  is the error term that includes an independent component  $\xi_{ij}$  and a first-order autoregressive or AR(1) component  $\rho\epsilon_{j,i-1}$  (where  $|\rho| < 1$ ) accounting for the serially correlated errors (Ruppert et al., 2003). An important aspect of this model is that we now have random effects that assess some level of individual differences in the random intercept, and that estimate the amount of smoothing. This is achieved by having a combined vector  $\mathbf{u}$  from Equation 10 that contains the random effect intercept and smoothing random effects, and an extended model design matrix  $\mathbf{Z}$  incorporating the knots shown in Equation 7 and columns for the random intercept. This allows the simultaneous estimation of between participant variation  $\sigma_U^2$  and amount of smoothing  $\sigma_u^2$  as variance components.

Figure 1 shows smooth plots for stimulus Clips 1–6 for both the simple GAM smooths and the GAMM smooths with autocorrelated error components and incorporating the within-participant correlation. The shaded areas represent pointwise confidence estimates or variability bands in the terminology used by Ruppert et al. (2003) for the smooth bounded at 2 standard errors above and below the smooth estimate. This provides approximate 95% pointwise confidence intervals for the smooth (this can be achieved simply by adding the `se=TRUE` and `shade=TRUE` commands to plots; see Appendix B). There is very little difference in the smooth terms, but the underestimation of the standard errors in the simple GAMs is easy to see by looking at the difference in thickness of the variability bands—the GAM smooths have very thin confidence bands which represent the underestimation of the standard errors in comparison with the thicker variability bands evident in the GAMM smooths.

Covariates can be added to these models and can be treated as any fixed effect covariate would be in a linear version of the model

with no smooth. These can be continuous (e.g., personality traits), categorical (e.g., country), or simple binary covariates (e.g., sex). A model including these examples would take the following form:

$$\begin{aligned} \text{valence}_{ij} &= U_j + f(x_{ij}) + \beta_1 \text{EDA}_i + \beta_2 a_j + \beta_3 c_j + \beta_4 s_j + \epsilon_{ij} \\ U_j &\sim N(0, \sigma_U^2) \\ \epsilon_{ij} &= \rho\epsilon_{j,i-1} + \xi_{ij} \\ \xi_{ij} &\sim N(0, \sigma_\xi^2). \end{aligned} \quad (12)$$

Here, the variables are the same as Equation 2 but include the one Level 1 variable  $\text{EDA}_i$  for electrodermal activity (also known as galvanic skin response), which is modeled here as a fixed effect parameter that would be a candidate to be modeled as a second smooth term. There would also be three Level 2 variables: (a) the continuous variable  $a_j$  for agreeableness; (b) a factor variable with three or more categories  $c_j$  for country—these would be recoded as binary (dummy) variables for each level of the factor; and (c) the binary variable  $s_j$  for sex. This model is much more complex than would be advisable and contains all of these extra variables for illustrative purposes. In general, more parsimonious models are favorable, but of course if a predictor warrants inclusion for either theoretical reasons or because it provides a useful explanatory covariate, then it should be included.

Full GAMMs with autocorrelated errors are computationally intensive (Yang, Qin, Zhao, Wang, & Song, 2012) and take time to compute. In practice, we have found it can be useful to conduct exploratory work using the simpler GAMs for exploratory data analysis (Tukey, 1977). Yang et al. (2012) proposed a GAM with autoregressive terms that is less computationally intensive, but it currently only works with count data and in Poisson regression situations. However, the underestimation of the standard error remains a problem with GAM models, so where necessary the more complete GAMMs should be used where inferences are being made.

Often the questions of interest in emotion perception research concern the details of what causes individual peaks and troughs within a smooth, that is, the signals of emotion that lead to an increase in the response variable at a certain moment in time. For example, does a smile, or more specifically does the combination of Facial Action Coding System Action Unit 12 and Action Unit 6 (Ekman & Friesen, 1978) cause a peak in a continuous self-report measure of perceived happiness, or does a particular acoustic signal cause a decoder to rate someone as feeling more negative? One approach to answering these sorts of question is suggested in the inference section, but a full exploration is beyond the scope of this article. The issues covered in the remainder of this article deal with ensuring that estimates of the optimal shared trajectory across time for a group of participants are valid. A valid smooth function can then inform us about how changes over time relate to other variables or allow us to isolate the changes over time to allow inference to be made concerning other variables. Time therefore is seen as a valid variable, but the reader should not forget that it is not time itself that causes the variation in the response variable, but it provides a useful encapsulation of the set of signals that lead to a shared response among the participants.



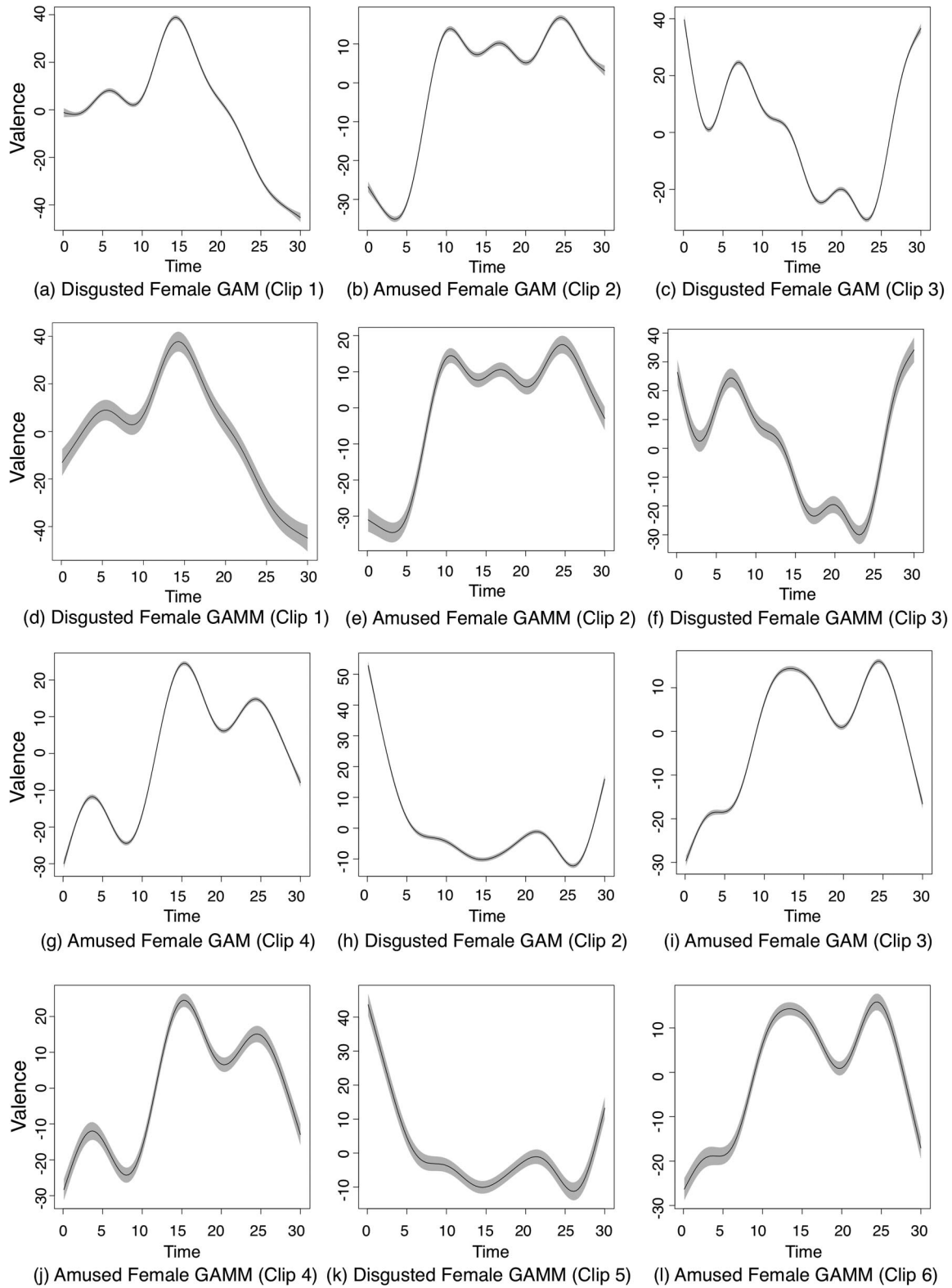


Figure 1. Plots of simple generalized additive model (GAM) smooths and generalized additive mixed model (GAMM) smooths ( $N = 160$ ; shading represents pointwise confidence estimates, or variability bands in the terminology used by Ruppert et al., 2003, for the smooth bounded at 2 SEs above and below the smooth estimate); a–c and g–i plot simple GAM smooths, d–f and j–l plot the same data using GAMMs.

### Assessing Effect Size

Using both GAM and GAMM techniques, we can find a number of useful statistics concerning dynamic emotional stimuli. The first of these is the overall effect size for a given clip as measured by  $R^2$  for the model. We take our definition of effect size from Kelley and Preacher (2012, p. 140): “a quantitative reflection of the magnitude of some phenomenon that is used for the purpose of addressing a question of interest.” The question of interest in this case is how much variability in the movement of a slider in response to emotional stimuli is shared between the individuals in a sample.

Here,  $R^2$  serves in its role as an effect size measure rather than assessing the goodness-of-fit of the model and we will consider the simple interpretation of  $R^2$  as the square of the correlation of the fitted smooth with the observed variables. We have mentioned that we are principally interested in isolating the emotional signals present within a clip so that we can make inferences about group differences and similarities. However, informing the researcher about the amount of variance explained by a model that just describes the shared change over time for a group of participants rating a particular emotional clip can be useful in practice.

Idiosyncratic changes from individuals should occur randomly and cancel each other out, while shared changes should remain and be reflected in the smooth. The correlation between the fitted smooth and the observed values of the response variable that is captured in the  $R^2$  statistic can serve as a quantitative reflection of this effect. Assessing the effect size of a minimal model containing only a smooth over time can tell us something about the level of emotionality present in a clip. This is a pragmatically useful piece of information when planning experiments or ranking video clips within a database of emotional material (e.g., Sneddon et al., 2012). If  $N$  participants have been asked to rate the same emotional clip along a particular dimension (e.g., valence), we can use the effect size of a simple model to provide us with an estimate of how much emotional material is present within the clip. That is, how much of the variance in a given clip can be explained due to the shared movements of the input device across the individual participants? We assume that any other variance is for reasons other than the movements required of the participant in the instructions. So in the language of Kelley and Preacher (2012),  $R^2$  can provide us with a quantitative reflection of the magnitude of emotionality present in any given clip.

In Table 3, we present the  $R^2$  statistics derived from simple GAMs containing only a single smooth term representing time for each of the clips in Data Sets 1 and 2. These can be compared with  $R^2$  statistics for more complex GAMMs incorporating individual participants as a random effect and including an autocorrelated error structure. We have found that effect size estimates do not differ much in practice, so given the computational expense of the GAMMs, simple GAMs may be more useful for the practical descriptive purpose of gauging the effect size, and  $R^2$  may prove a useful measure as a proxy for the level of emotional signals observed in each clip. However, where inferences are being made, a more complete GAMM is advised.

We can see that for the first six clips we have a fairly consistent medium effect size ranging from  $R^2 = .202$  to  $R^2 = .318$ , whereas the second set of six clips shows a greater variability, with an effect size ranging from  $R^2 = .057$  to  $R^2 = .642$ . Even with this greater range of effect sizes and the smaller sample sizes used in

Table 3

Data for GAMM's  $R^2$  Effect Sizes for Each of the Six Clips

Clip	Emotion	Set	$N$	GAMM	GAM
1	Disgust	1	160	0.281	0.283
2	Amusement	1	160	0.310	0.312
3	Disgust	1	160	0.299	0.303
4	Amusement	1	160	0.318	0.319
5	Disgust	1	159	0.202	0.206
6	Amusement	1	160	0.282	0.282
7	Amusement	2	40	0.280	0.280
8	Amusement	2	40	0.325	0.326
9	Amusement	2	40	0.642	0.642
10	Disgust	2	40	0.057	0.057
11	Disgust	2	40	0.210	0.211
12	Disgust	2	40	0.345	0.346

Note. GAMM = generalized additive mixed model; GAM = generalized additive model.

the second set of six clips, there is very little difference between effect size observed in the GAMMs and the simple GAMs. Figure 2 displays the plots of the extremes of this range of effect sizes where the more obvious curve in Clip 9 can be interpreted as greater level of shared perceived emotion as it relates to changes in valence, and the flatter line in Clip 10 can be interpreted as less agreement in perceived emotion as it relates to changes in valence.

### Sample Size

Given this measure of effect size, a further practical question of use to many researchers is how many participants are required to provide a reasonable estimated smooth, one that can give a researcher reasonable confidence that it represents the smooth in the population?

Here, we share the results of an exercise to assess model behavior under differing sample sizes using  $R^2$  as an effect size measure. We adopted an approach similar to the subsampling bootstrap technique of Politis and Romano (1994); Politis, Romano, and Wolf (1999); and earlier subsampling approaches (Forsythe & Hartigan, 1970; Hartigan, 1969). It is also similar to an approach by Fritz and MacKinnon (2007), who used generated data at different sample sizes to assess the sample size required to detect mediated effects in mediation models. As the GAMMs proposed in this article are computationally very intensive especially when they account for autocorrelation in the residuals, this exercise was completed with GAMs.

In our approach, we make use of the large sample size of 160 participants in Clips 1–6. We assume that the 160 participants represent the population and take a random sample of 100 subsamples starting with just 2 participants and increasing to subsamples of 159 participants. On these subsamples, we perform simple GAMs of the relationship between variance and time and compare the mean  $R^2$  for the model to the “population”  $R^2$ . Figure 3 displays these relationships for each of the six data sets.

We judge that for the data presented in this article somewhere between 20 and 30 participants represents the point where the required number of participants allow a confidence that the emotional signal is present. Beyond this point, there are diminishing returns from an increase in sample size for the data we have examined. We make no claim that these results will generalize to

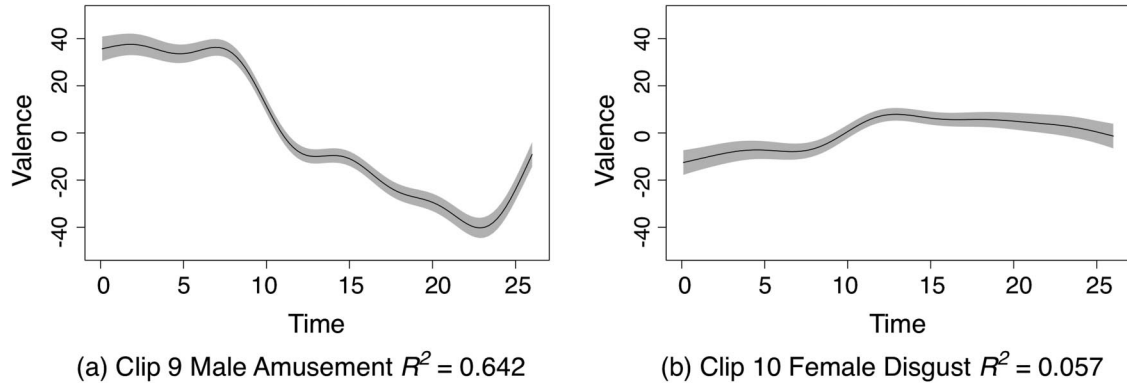


Figure 2. Plots of generalized additive mixed model smooths for Clip 9 (large effect size) and Clip 10 (small effect size).

other situations. The appropriate sample size for such analyses remains an empirical question and is likely to vary with context and participant group; we provide the results of this exercise to encourage other researchers to share similar evidence.

### Inference

Drawing inferences from GAMs and GAMMs presents interesting problems. One of the primary issues is the fact that the smooth terms

are nonparametric and do not lend themselves readily to straightforward numerical comparison. Standard linear model and linear mixed model techniques such as ANOVA tests can be used to compare models that drop components or test against some minimal null model. However, an interesting inferential question arises in the situation of GAMs and GAMMs concerning the nature of certain areas of curves: Do the features of the curve correspond to something real or are they simply a spurious undulation? This is especially

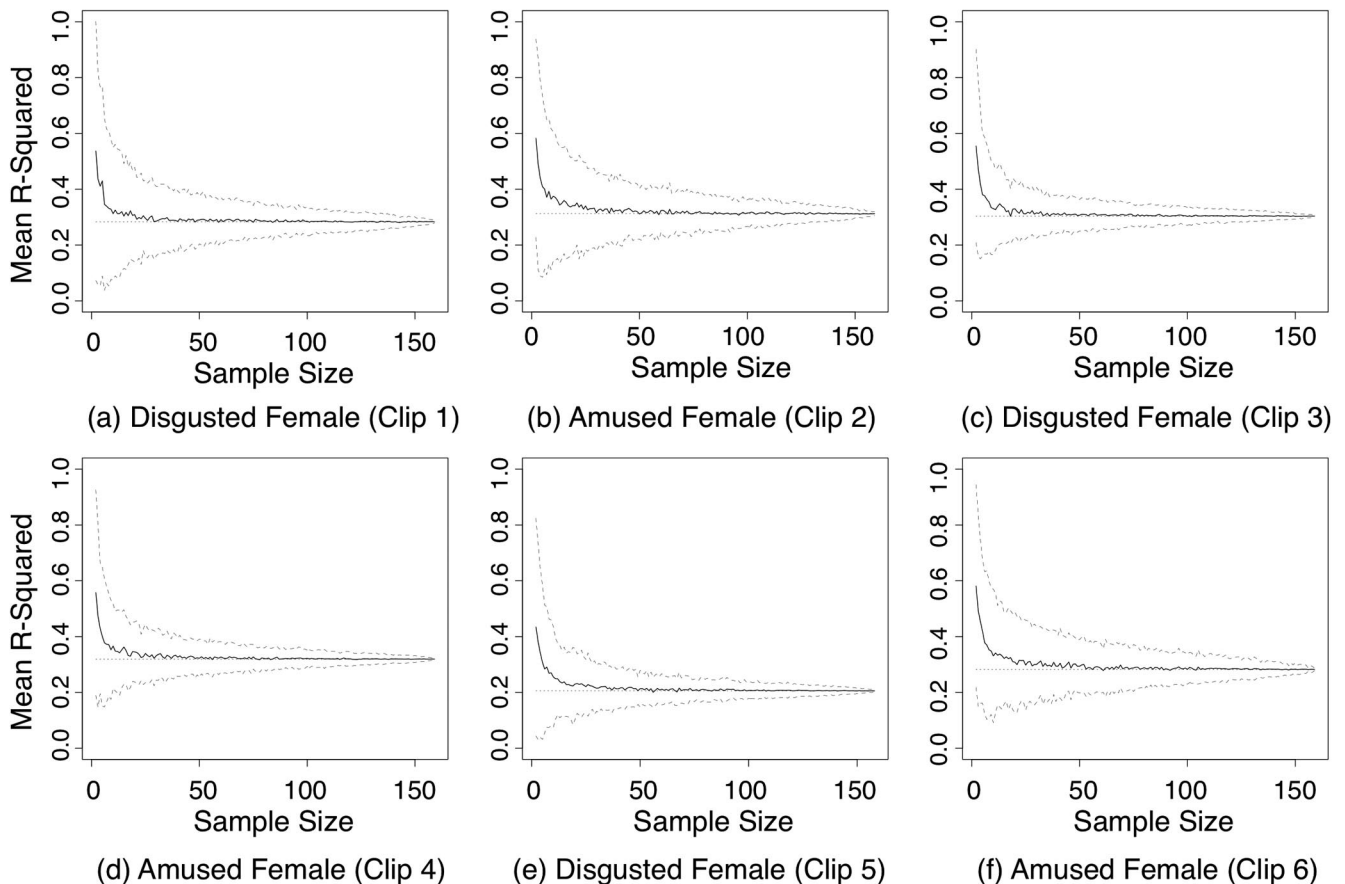


Figure 3. Plots of mean  $R^2$  as a function of sample size for each of the data sets.

interesting in the case of emotion perception where we may want to know whether a given curve represents the occurrence of an emotional signal or set of emotional signals or if the curve feature is just there by chance. One approach to answering this question is provided by Chaudhuri and Marron (1999) with their significant zero crossings of derivatives (SiZer) technique. This approach involves taking the first derivative of the regression function to create a new curve with corresponding variability bands. Any area of the new curve in which the variability band is greater than zero corresponds to an area in the original regression function that is increasing significantly; where the variability band crosses below zero the original regression function is deemed to be decreasing significantly. A full exploration of these techniques would require a separate article; here, we restrict further discussion to inference in linear variables and interactions.

**Inference in Linear Variables**

How do we use these models to make statistical inferences about covariate variables once the smooths have been accounted for? As GAMs allow us to combine nonparametric and parametric terms for categorical predictors, we can look at the importance of the effect size of one level of a factor in place of another using the  $\beta$  coefficients of binary or categorical variables, much as they would be used in an ordinary linear model. Similarly, the  $\beta$  coefficients for continuous predictors can be used in the same way as for a linear model. This means making inferences concerning the extra terms shown in the model specified by Equation 12. In effect, we can explain the shared change in the response variable over time and partial it out to assess if there are differences due to levels of a factor or binary variable. In the following examples, a binary variable (between females and males) is the focus, examining differences in the overall level of valence they provided in response to observing the emotional clip.

Table 4 displays the relevant statistics for a GAMM, a GAM, and an ordinary linear model for Clip 9 and Clip 12. We can see that in the GAMM for Clip 9 we see no difference due to sex, a non-significant estimate of a valence score that is 3.1 higher for males than females, with a standard error of 4.97. The GAMM for Clip 12 provides evidence for a difference due to sex with an estimate of 19.37 higher in males than females, and we can say this with reasonable confidence when we look at the standard error. The GAMs provide similar  $R^2$  values to the GAMMs as they take account of the change in the response variable over time; the estimates for the difference due to sex are also not dissimilar.

However, we see the problem of the underestimated standard error in both the models for Clip 9 and Clip 12. This leads to inflated  $t$  values and the conclusion of significant estimates for both Clip 9 and Clip 12. This would lead to the incorrect inference that in both Clips 9 and 12 males and females are different when the GAMM only provides evidence for differences in Clip 12.

When we compare both these models with the linear versions of the model that take no account of change over time we find that the  $R^2$  for Clip 9 is very small, which is not surprising given the lack of any effect, and the  $R^2$  for Clip 12 is 0.08. The estimates for the difference due to sex are the same as those for the GAM. However, again we get the problem of underestimated standard errors in both models leading to a greatly inflated  $t$  value in the linear model for Clip 12 and a  $t$  value that approaches the conventional significance level of 0.05 in Clip 9. This would lead to the erroneous conclusion of a difference between males and females in Clip 9 and an over estimation of the precision with which we can state the relevant statistic in Clip 12.

An additional check that we would like to conduct on these models to be confident that they are valid is to rule out the possibility that there are interactions with the smooth terms.

**Interactions**

So far the explanations of the models have assumed that the GAM specified smooth has no impact on other variables of interest; in other words, we have accepted the assumption of additivity and expect no interaction between the smooth over time and other variables of interest. There may, however, be cases where a researcher may seek to see how smooths interact with other variables. One reason may be to rule out the possibility of an interaction with the goal of ensuring the assumption of additivity is sound, or it may be the case that an interaction has been predicted. Models that include interactions can take two main forms; smooths over time can interact with categorical factor or binary variables or they may interact with continuous variables.

**Categorical interactions.** Interactions with categorical variables are achieved by adopting a version of the GAM paradigm known as varying-coefficient models (Hastie & Tibshirani, 1993; for alternative approaches, see Durbán et al., 2005; Ruppert et al., 2003). In such models, smooths are multiplied by a covariate which has the effect of creating a separate smooth function for each level of a categorical variable (this is achieved

Table 4  
*Comparison of Linear Parametric Differences in Sex Using GAMMs, GAMs, and Linear Models for Clips 9 and 12*

Measure	Clip 9			Clip 12		
	GAMM	GAM	Linear	GAMM	GAM	Linear
$R^2$	0.64	0.64	0.0003	0.42	0.42	0.08
Intercept	28.18	30.45	30.45	−8.35	−9.85	−9.85
$\beta$ male	3.1	−1.32	−1.32	19.37	22.32	22.32
SE	4.97	0.41	0.69	5.49	0.59	0.75
$t$	0.63	−3.19	−1.9	3.53	37.63	29.74
$p$	.53	<.01	.056	<.001	<.001	<.001

Note. GAMM = generalized additive mixed model; GAM = generalized additive model.



in the R package *mgcv* using “by” variables in the formulation of the smooth term). Once again using Clip 9 and Clip 12, we can illustrate this by creating separate male and female smooths over time. Table 5 shows the statistics from varying-coefficient GAMMs and GAMs fitted to the Clip 9 and Clip 12 data. We can see in the GAMM for Clip 9 that again there is no difference due to sex, a non-significant estimate of a valence level of  $-1.39$ , this time a little lower for males than females and closer to the equivalent parameter estimate for the GAM and linear models in Table 4. However with a standard error of 5.25, this is of little consequence. There is a slight increase in the  $R^2$  associated with this model, but it comes at a price of increased degrees of freedom to accommodate the two smooths, and a log-likelihood ratio test between the two models favors the more parsimonious explanation of leaving the interaction out in this case. As there is no significant effect due to sex, it should not really be included in the model, and it would depend on the goals of the experimenter if there was value in creating this model. Similarly, in the GAMM for Clip 12, again we see a strong difference due to sex. The GAM versions of these models provide similar coefficient estimates but once again suffer from an underestimation of the standard error. Figure 4 shows the separate smooths for male and females for the Clip 9 and Clip 12. It is obvious that there is almost no difference between the male and female smooths for Clip 9; the lines fall within the variability bands of each smooth indicating there is no significant difference. In Clip 12, however, we can see a difference in the smooths between males and females, and there is no overlap in the variability bands at the peaks and troughs of the lines. As the overall pattern is largely the same we may conclude that the males do not interpret the emotionality they are observing to be as extreme as the females decoders.

**Continuous interactions.** Although in depth exploration of interactions with continuous variables is beyond the scope of the current article, the R package *mgcv* does permit this using tensor product smooths, or using tensor product spline bases (for further details on this method, see Wood, 2006a, pp. 162–167). This, in essence, permits two smooths to be combined to produce a surface capturing the interaction between the two continuous variables, one of which is the change in the response variable over time. One problem here is the use of different units of measurement in two covariates. However, the use of tensor product bases allows the

smooths to be invariant to linear rescaling of covariates making it possible to display the smooth interaction relationship between two variables where the measured quantities are in different units (Wood, 2006a, 2006b). Figure 5 displays two smooths for data taken from decoders of emotion looking at a video clip with a male exhibiting fear. Figure 5a shows a straightforward GAMM smooth for all the participants, whereas Figure 5b shows a plot of a smooth derived from a model that includes the continuous variable agreeableness in a bivariate interaction with the continuous variable time. This has the effect of displaying the changes in the time smooth as a result of its interaction with the level of the personality characteristic agreeableness in the decoders. We can see considerable agreement in the overall pattern with a sharp rise followed by a trough followed by a further one or two gentle peaks. However, the interaction highlights differences: For high levels of agreeableness, movements after the initial peak are somewhat muted; for medium levels of agreeableness, movements after the initial peak exhibit a more obvious “two gentle peaks” structure as in the simple smooth Figure 5a; finally, for low levels of agreeableness, movement after the initial peak is more intense, and the gentle two peak structure has given way to a stronger single peak structure. This interaction smooth surface then shows considerable agreement for the first peak and more ambiguity dependent on agreeableness in the structures that come after.

## Ambiguity

These techniques permit an assessment of the level of ambiguity present in emotion and non-verbal signals. There are certain areas within each of the clips that contain greater or lesser degrees of shared response, and if we assume that decoders act upon the information that is present in the clips we can draw the conclusion that where there is less agreement there is less useful information. To quantify this, we can take advantage of the variability bands derived from the standard error estimates of the GAMMs. Where there is a greater level of agreement between decoders, the signals in that section of clip are less ambiguous and conversely; where there is a lesser level of agreement between decoders, the signals are more ambiguous. This information may be used to help isolate areas within a stimulus clip that may be of particular interest, areas that show strong agreement and that therefore warrant more detailed study and areas in which there are enduring periods of ambiguity. Figure 6 shows three plots derived from GAMMs, and the curves reveal that, in periods of strong movement, the variability bands are often close, suggesting more agreement between the decoders. This is not surprising, as such movement is usually due to the presence of strong emotional signals in the data such as a smile or a strong look of disgust. In areas where the traces plateau, again we have less agreement and greater ambiguity due to low level or ambiguous signals. However, we do occasionally see ambiguous rises such as the first upward slope observed in Figure 6a, where the rise is pronounced, but there is a greater ambiguity than in the second sharper rise that immediately follows.

## Some Issues

As GAMs are generalizations of linear regression, most of the standard linear regression diagnostic techniques have an

Table 5  
*Linear Differences in Sex Using Categorical Interaction GAMMs and GAMs With Separate Male and Female Smooths for Clips 9 and 12*

Measure	Clip 9		Clip 12	
	GAMM	GAM	GAMM	GAM
$R^2$	0.65	0.65	0.48	0.48
Intercept	28.18	30.45	-8.35	-9.85
$\beta$ male	-1.39	-1.32	22.33	22.32
SE	5.25	0.41	5.94	0.57
$t$	-0.27	-3.21	3.76	39.48
$p$	.79	<.01	<.001	<.001

Note. GAMM = generalized additive mixed model; GAM = generalized additive model.

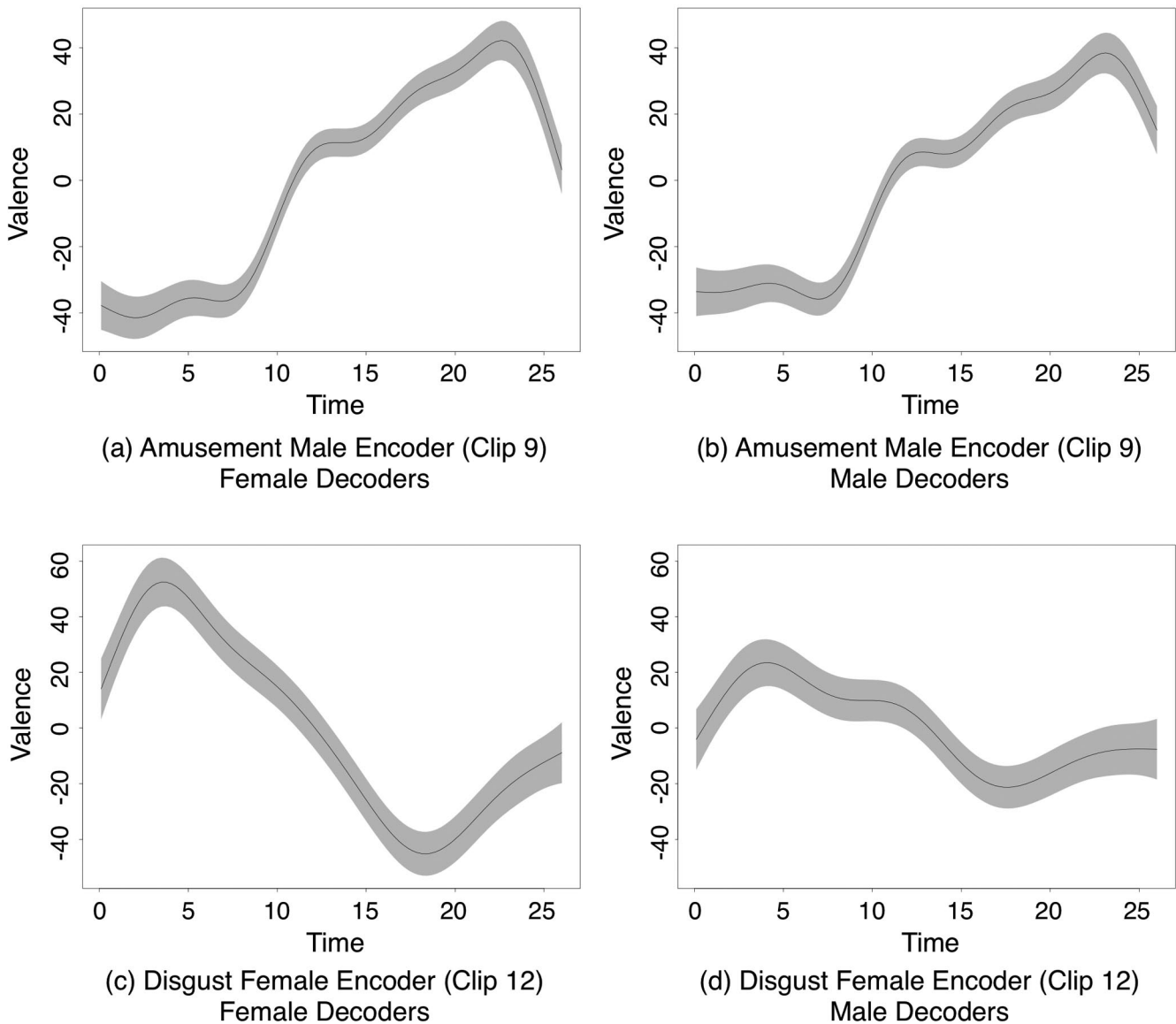


Figure 4. Categorical interactions with separate smooths for male and female decoders.

equivalent that is suitable for GAMs (some can be obtained using the `gam.check` command in the R package `mgcv`). The presence of concavity in the data can be an issue. This is the nonparametric equivalent of multicollinearity, in which the function of a covariate can be approximated by a linear combination of the functions of other covariates. Where concavity is present in the data, variance estimates can be biased and standard errors underestimated. This is less of a problem in the penalized splines approach (used in the R package `mgcv`) compared to backfitting algorithms used in other software (Figueiras, Roca-Pardiñas, & Cadarso-Suárez, 2005; additionally, the `concurvity` command can be used in the R package `mgcv` to assess the extent of the problem).

Throughout this article, we have largely concentrated on additive models or additive mixed models but have referred to them in their generalized sense. Generalized additive models

are to additive models as generalized linear models are to linear models. In the generalized form, assumptions of the normal distribution of the response variable are relaxed and models can be fitted to response variables with any distribution from the exponential family (e.g., normal, Poisson, binomial, gamma). This means that more types of data can be accounted for using these techniques, such as binary and multinomial categorical data. We have not addressed these types of data but wished to highlight that these techniques may open the possibility of creating models that are useful in the analysis of categorical time series data that are also prevalent in the emotion perception literature.

Finally, there is a known source of error that is not addressed in these models. This is due to the response lag in the decoder's manipulation of the mouse—the time that it takes an encoder to respond to emotional signals on an encoder's face or body. This

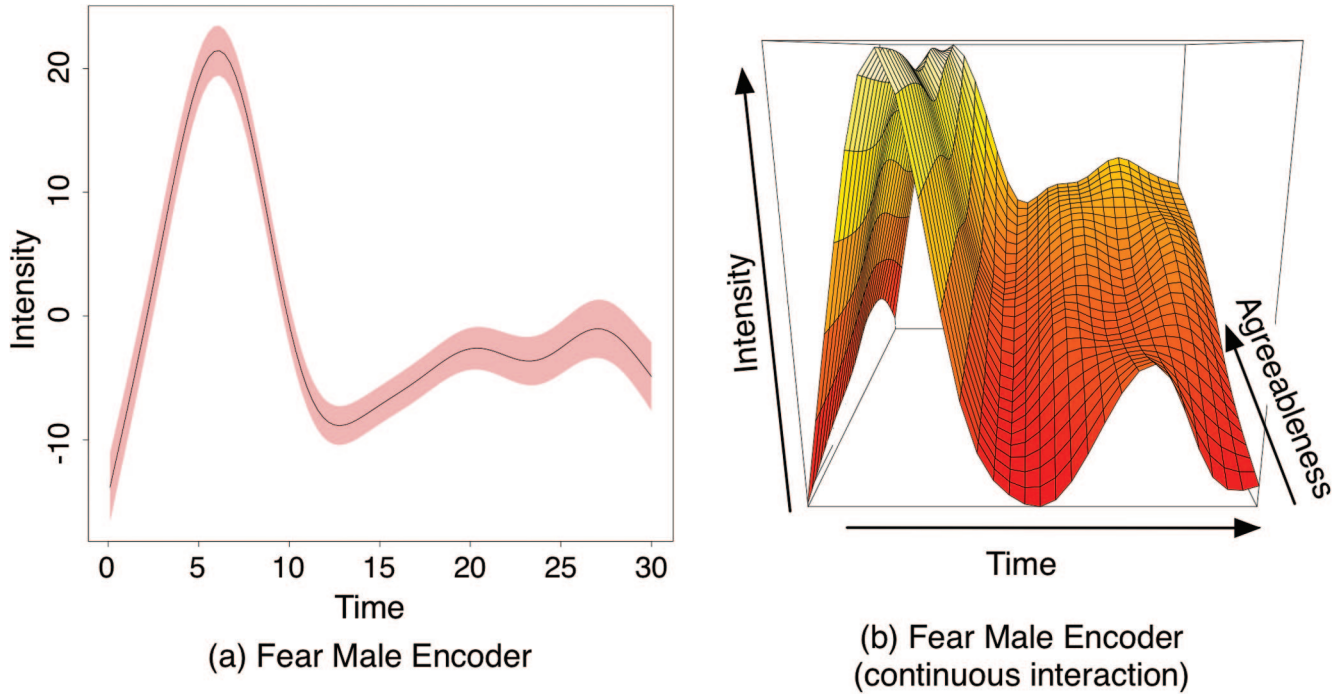


Figure 5. Two smooths derived from models of data taken from decoders of emotion looking at a video clip with a male exhibiting fear. Panel (a) displays a generalized additive mixed model smooth with no continuous interaction. Panel (b) displays a smooth surface derived from a model that includes the continuous variable agreeableness in a bivariate interaction with the continuous variable time.

will differ across individuals and lead to some individuals having peaks and troughs that are unaligned. Resolution of this issue is beyond the scope of the current article, but there are possibilities using dynamic time-warping (Giorgino, 2009)—a technique used in word recognition studies or similar techniques that may be used to bring such response lags into alignment. Additionally, there are promising techniques for peak and trough alignment using dy-

namic probabilistic canonical correlation analysis (Nicolaou, Pavlovic, & Pantic, 2012).

### Conclusions

The overwhelming majority of previous research on the perception of the facial expression of emotion has relied on a recognition

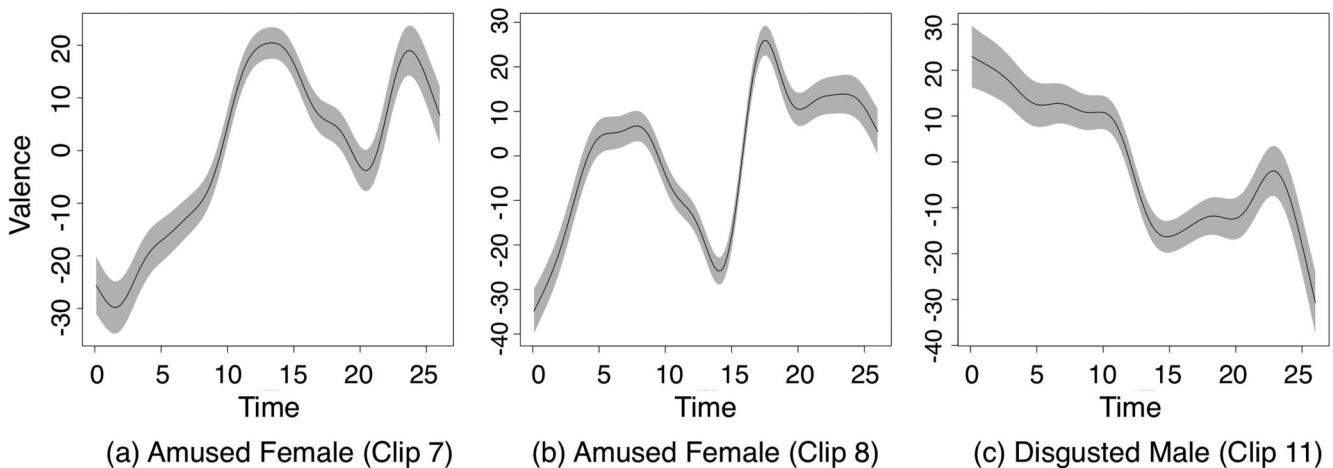


Figure 6. Plots of generalized additive mixed model smooths where the smooths vary as random effects. Broader confidence bands can be interpreted as reduced agreement between the decoders.

paradigm and on stimuli consisting of posed static photographs of faces (Tcherkassof et al., 2007). It has been suggested (Wehrle et al., 2000) that implicit acceptance of this limitation may be based on a “basic emotions” (Ekman, 1992) theoretical position that views emotional expressions as innate prototypical patterns that unfold in a set sequence and for which subtle dynamic changes are of little importance. However, recently there have been an increasing number of pleas for research on perception of the emotion expressed in faces to be based on dynamic and spontaneous stimuli (e.g., Ambadar, Schooler, & Cohn, 2005; Biele & Grabowska, 2006; Edwards, 1998; Kamachi et al., 2001; Naab & Russell, 2007; Sneddon et al., 2011; Tcherkassof et al., 2007).

In this article, we suggest that generalized additive models and generalized additive mixed models provide a solution to this particular problem in emotion perception research. They provide a route that allows the emotion perception field to assess more than just the standard static image stimuli, and enable a move to examining stimuli containing dynamic changes in emotion expression over time. Using SiZer methods, inferences can be drawn concerning features of dynamic perception ratings and also standard modeling methods permit inferences concerning differences in other variables included in the models. While these models have been explored with a focus on time related data within an emotion perception domain, they are broadly applicable to any research situation in which time is a factor or indeed in which it is likely that the relationships between variables are not linear, a possibility that is too often overlooked in psychology. Time related data are already commonly used in areas of psychology that collect functional data, however, and such data are likely to become increasingly common if, as Miller (2012) suggests, smartphones represent the next important research technology in psychology. Although not dealt with in this article, generalized additive models are also very useful in accounting for spatial data and combinations of spatio-temporal data, a type of data that has not been extensively used in psychology but that is also likely to become increasingly important as the use of smartphones as research tools increases. Generalized additive mixed models provide a solution to the specific problem of analyzing emotion perception and may result in the ability to address unanswered problems in a broader range of psychological domains. They offer statistical analysis solutions within a framework that is very similar to the widely used linear regression techniques and which should provide a transition for most researchers in practical terms.

With regard to emotion perception research, this technique opens up the possibility of comparing dynamic patterns of behavior, and this, for the first time, should allow emotion researchers to compare perceptions of spontaneous sequences of facial expressions and expressions of emotion. This in turn may allow us to look beyond the recognition paradigm and to break our reliance on posed static photographs when seeking to understand real-life expression of emotion.

## References

Ambadar, Z., Schooler, J. W., & Cohn, J. F. (2005). Deciphering the enigmatic face: The importance of facial dynamics in interpreting subtle

facial expressions. *Psychological Science*, 16, 403–410. doi:10.1111/j.0956-7976.2005.01548.x

Andersen, R. (2009). Nonparametric methods for modeling nonlinearity in regression analysis. *Annual Review of Sociology*, 35, 67–85. doi:10.1146/annurev.soc.34.040507.134631

Baker, H., Shortt, P., Perkins, G. (Writers), & Lowney, D. (Director). (1996). Hell [Television series episode]. In M. Bell (Executive producer), *Father Ted*. London, United Kingdom: Hat Trick Productions.

Bates, D., Maechler, M., & Bolker, B. (2012). lme4: Linear mixed-effects models using s4 classes (R package Version 0.999999-0) [Computer software]. Retrieved from <http://CRAN.R-project.org/package=lme4>

Biele, C., & Grabowska, A. (2006). Sex differences in perception of emotion intensity in dynamic and static facial expressions. *Experimental Brain Research*, 171, 1–6. doi:10.1007/s00221-005-0254-0

Bohnert, A. S. B., Prescott, M. R., Vlahov, D., Tardiff, K. J., & Galea, S. (2010). Ambient temperature and risk of death from accidental drug overdose in New York City, 1990–2006. *Addiction*, 105, 1049–1054. doi:10.1111/j.1360-0443.2009.02887.x

Briscoe, E., & Feldman, J. (2011). Conceptual complexity and the bias/variance tradeoff. *Cognition*, 118, 2–16. doi:10.1016/j.cognition.2010.10.004

Carslaw, D. C., Beevers, S. D., & Tate, J. E. (2007). Modelling and assessing trends in traffic-related emissions using a generalised additive modelling approach. *Atmospheric Environment*, 41, 5289–5299. doi:10.1016/j.atmosenv.2007.02.032

Chaudhuri, P., & Marron, J. S. (1999). SiZer for exploration of structures in curves. *Journal of the American Statistical Association*, 94, 807–823. doi:10.1080/01621459.1999.10474186

Cleveland, W. S. (1979). Robust locally weighted regression and smoothing scatterplots. *Journal of the American Statistical Association*, 74, 829–836. doi:10.1080/01621459.1979.10481038

Cowie, R., McKeown, G., & Douglas-Cowie, E. (2012). Tracing emotion: An overview. *International Journal of Synthetic Emotions*, 3, 1–17. doi:10.4018/jse.2012010101

Crainiceanu, C., Reiss, P., Goldsmith, J., Huang, L., Huo, L., & Scheipl, F. (2012). Refund: Regression with functional data (R package Version 0.1-6) [Computer software]. Retrieved from <http://CRAN.R-project.org/package=refund>

DeCoster, J., Iselin, A.-M. R., & Gallucci, M. (2009). A conceptual and empirical examination of justifications for dichotomization. *Psychological Methods*, 14, 349–366. doi:10.1037/a0016956

Durbán, M., Harezlak, J., Wand, M. P., & Carroll, R. J. (2005). Simple fitting of subject-specific curves for longitudinal data. *Statistics in Medicine*, 24, 1153–1167. doi:10.1002/sim.1991

Edwards, K. (1998). The face of time: Temporal cues in facial expressions of emotion. *Psychological Science*, 9, 270–276. doi:10.1111/1467-9280.00054

Eilers, P., & Marx, B. (1996). Flexible smoothing with B-splines and penalties. *Statistical Science*, 11, 89–121. doi:10.1214/ss/1038425655

Ekman, P. (1992). Are there basic emotions? *Psychological Review*, 99, 550–553. doi:10.1037/0033-295X.99.3.550

Ekman, P., & Friesen, W. (1978). *Facial Action Coding System: A technique for the measurement of facial movement*. Palo Alto, CA: Consulting Psychologists Press.

Elfenbein, H. A., & Ambady, N. (2002). On the universality and cultural specificity of emotion recognition: A meta-analysis. *Psychological Bulletin*, 128, 203–235. doi:10.1037/0033-2909.128.2.203

Figueiras, A., Roca-Pardiñas, J., & Cadarso-Suárez, C. (2005). A bootstrap method to avoid the effect of concavity in generalised additive models in time series studies of air pollution. *Journal of Epidemiology & Community Health*, 59, 881–884. doi:10.1136/jech.2004.026740

Finch, W. H., Chang, M., & Davis, A. S. (2011). Application of CART, neural networks, and generalized additive models: A case study. *Multiple Linear Regression Viewpoints*, 37, 8–21.

Finch, W. H., Chang, M., Davis, A. S., Holden, J. E., Rothlisberg, B. A., & McIntosh, D. E. (2011). The prediction of intelligence in preschool



- children using alternative models to regression. *Behavior Research Methods*, 43, 942–952. doi:10.3758/s13428-011-0102-z
- Forsythe, A., & Hartigan, J. A. (1970). Efficiency of confidence intervals generated by repeated subsample calculations. *Biometrika*, 57, 629–639. doi:10.1093/biomet/57.3.629
- Fritz, M. S., & MacKinnon, D. P. (2007). Required sample size to detect the mediated effect. *Psychological Science*, 18, 233–239.
- Gelman, A., & Hill, J. (2007). *Data analysis using regression and multilevel/hierarchical models*. Cambridge, England: Cambridge University Press.
- Giorgino, T. (2009). Computing and visualizing dynamic time warping alignments in R: The dtw package. *Journal of Statistical Software*, 31, 1–24.
- Guisan, A., Edwards, T., & Hastie, T. (2002). Generalized linear and generalized additive models in studies of species distributions: Setting the scene. *Ecological Modelling*, 157(2–3), 89–100. doi:10.1016/S0304-3800(02)00204-1
- Gunes, H., & Schuller, B. (2013). Categorical and dimensional affect analysis in continuous input: Current trends and future directions. *Image and Vision Computing*, 31, 120–136. doi:10.1016/j.imavis.2012.06.016
- Hartigan, J. (1969). Using subsample values as typical values. *Journal of the American Statistical Association*, 64, 1303–1317. doi:10.1080/01621459.1969.10501057
- Hastie, T. (2011). Gam: Generalized additive models (R package Version 1.06.2) [Computer software]. Retrieved from <http://CRAN.R-project.org/package=gam>
- Hastie, T., & Tibshirani, R. (1990). *Generalized additive models*. Boca Raton, FL: Chapman and Hall/CRC.
- Hastie, T., & Tibshirani, R. (1993). Varying-coefficient models. *Journal of the Royal Statistical Society Series B*, 55, 757–796.
- Hastie, T., Tibshirani, R., & Friedman, J. H. (2001). *The elements of statistical learning: Data mining, inference, and prediction*. doi:10.1007/978-0-387-21606-5
- Huitema, B., & McKean, J. (1998). Irrelevant autocorrelation in least-squares intervention models. *Psychological Methods*, 3, 104–116. doi:10.1037/1082-989X.3.1.104
- Kamachi, M., Bruce, V., Mukaida, S., Gyoba, J., Yoshikawa, S., & Akamatsu, S. (2001). Dynamic properties influence the perception of facial expressions. *Perception*, 30, 875–887. doi:10.1068/p3131
- Kelley, K., & Preacher, K. J. (2012). On effect size. *Psychological Methods*, 17, 137–152. doi:10.1037/a0028086
- Knoblauch, K., & Maloney, L. T. (2008). Estimating classification images with generalized linear and additive models. *Journal of Vision*, 8, 10.1–19. doi:10.1167/8.16.10
- Lado, M. J., Cadarso-Suárez, C., Roca-Pardiñas, J., & Tahoces, P. G. (2008). Categorical variables, interactions and generalized additive models. Applications in computer-aided diagnosis systems. *Computers in Biology and Medicine*, 38, 475–483. doi:10.1016/j.combiomed.2008.01.011
- Levitin, D. J., Nuzzo, R. L., Vines, B. W., & Ramsay, J. O. (2007). Introduction to functional data analysis. *Canadian Psychology/Psychologie canadienne*, 48, 135–155.
- Lin, X., & Zhang, D. (1999). Inference in generalized additive mixed models by using smoothing splines. *Journal of the Royal Statistical Society Series B*, 61, 381–400. doi:10.1111/1467-9868.00183
- MacCallum, R. C., Zhang, S., Preacher, K. J., & Rucker, D. D. (2002). On the practice of dichotomization of quantitative variables. *Psychological Methods*, 7, 19–40. doi:10.1037/1082-989X.7.1.19
- Marshall, D. B., & English, D. J. (2000). Neural network modeling of risk assessment in child protective services. *Psychological Methods*, 5, 102–124. doi:10.1037/1082-989X.5.1.102
- McCrae, R. R., & Costa, P. T., Jr. (2004). A contemplated revision of the NEO Five-Factor Inventory. *Personality and Individual Differences*, 36, 587–596. doi:10.1016/S0191-8869(03)00118-1
- Miller, G. (2012). The smartphone psychology manifesto. *Perspectives on Psychological Science*, 7, 221–237. doi:10.1177/1745691612441215
- Naab, P. J., & Russell, J. A. (2007). Judgments of emotion from spontaneous facial expressions of New Guineans. *Emotion*, 7, 736–744. doi:10.1037/1528-3542.7.4.736
- Nicolaou, M. A., Pavlovic, V., & Pantic, M. (2012). Dynamic probabilistic CCA for analysis of affective behaviour. In A. Fitzgibbon, S. Lazebnik, P. Perona, Y. Sato, & C. Schmid (Eds.), *Computer Vision—ECCV 2012: Lecture notes in computer science* (Vol. 7578, pp. 98–111). Berlin, Germany: Springer.
- Oravec, Z., Tuerlinckx, F., & Vandekerckhove, J. (2011). A hierarchical latent stochastic differential equation model for affective dynamics. *Psychological Methods*, 16, 468–490. doi:10.1037/a0024375
- Pham, T., & Wand, M. (2012). Gammslice: Generalized additive mixed model analysis via slice sampling (R package Version 1.2-6) [Computer software]. Retrieved from <http://CRAN.R-project.org/package=gammSlice>
- Politis, D. N., & Romano, J. P. (1994). Large sample confidence regions based on subsamples under minimal assumptions. *The Annals of Statistics*, 22, 2031–2050. doi:10.1214/aos/1176325770
- Politis, D. N., Romano, J. P., & Wolf, M. (1999). *Subsampling*. doi:10.1007/978-1-4612-1554-7
- Ramsay, J. O., & Silverman, B. W. (2005). *Functional data analysis* (Springer Series in Statistics). New York, NY: Springer.
- R Development Core Team. (2010). R: A language and environment for statistical computing [Computer software]. Retrieved from <http://www.R-project.org>
- Rozin, A., Rozin, P., & Goldberg, E. (2004). The feeling of music past: How listeners remember musical affect. *Music Perception*, 22, 15–39. doi:10.1525/mp.2004.22.1.15
- Ruppert, D., Wand, M. P., & Carroll, R. J. (2003). *Semiparametric regression*. doi:10.1017/CBO9780511755453
- Russell, J. A. (1994). Is there universal recognition of emotion from facial expression? A review of the cross-cultural studies. *Psychological Bulletin*, 115, 102–141. doi:10.1037/0033-2909.115.1.102
- Scheipl, F., Greven, S., & Kuechenhoff, H. (2008). Size and power of tests for a zero random effect variance or polynomial regression in additive and linear mixed models. *Computational Statistics and Data Analysis*, 52, 3283–3299. doi:10.1016/j.csda.2007.10.022
- Schubert, E. (2010). Continuous self-report methods. In P. N. Juslin & J. Sloboda (Eds.), *Handbook of music and emotion* (pp. 223–253). Oxford, England: Oxford University Press.
- Sneddon, I., McKeown, G., McRorie, M., & Vukicevic, T. (2011). Cross-cultural patterns in dynamic ratings of positive and negative natural emotional behaviour. *PLoS ONE*, 6, e14679. doi:10.1371/journal.pone.0014679
- Sneddon, I., McRorie, M., McKeown, G., & Hannraty, J. (2012). The Belfast Induced Natural Emotion Database. *IEEE Transactions on Affective Computing*, 3, 32–41. doi:10.1109/T-AFFC.2011.26
- Strobl, C., Malley, J., & Tutz, G. (2009). An introduction to recursive partitioning: Rationale, application, and characteristics of classification and regression trees, bagging, and random forests. *Psychological Methods*, 14, 323–348. doi:10.1037/a0016973
- Tan, X., Shiyko, M. P., Li, R., Li, Y., & Dierker, L. (2012). A time-varying effect model for intensive longitudinal data. *Psychological Methods*, 17, 61–77. doi:10.1037/a0025814
- Tcherkassoff, A., Bollon, T., Dubois, M., Pansu, P., & Adam, J. (2007). Facial expressions of emotions: A methodological contribution to the study of spontaneous and dynamic emotional faces. *European Journal of Social Psychology*, 37, 1325–1345. doi:10.1002/ejsp.427
- Tukey, J. (1977). *Exploratory data analysis* (Addison-Wesley Series in Behavioral Science). Boston, MA: Addison-Wesley.
- Wagner, H., MacDonald, C., & Manstead, A. (1986). Communication of individual emotions by spontaneous facial expressions. *Journal of Personality and Social Psychology*, 50, 737–743. doi:10.1037/0022-3514.50.4.737
- Walls, T. A., & Schafer, J. L. (2006). *Models for intensive longitudinal data*. doi:10.1093/acprof:oso/9780195173444.001.0001

- Wehrle, T., Kaiser, S., Schmidt, S., & Scherer, K. (2000). Studying the dynamics of emotional expression using synthesized facial muscle movements. *Journal of Personality and Social Psychology*, 78, 105–119. doi:10.1037/0022-3514.78.1.105
- Wieling, M., Nerbonne, J., & Baayen, R. H. (2011). Quantitative social dialectology: Explaining linguistic variation geographically and socially. *PLoS ONE*, 6, e23613. doi:10.1371/journal.pone.0023613
- Wood, S. N. (2000). Modelling and smoothing parameter estimation with multiple quadratic penalties. *Journal of the Royal Statistical Society Series B*, 62, 413–428. doi:10.1111/1467-9868.00240
- Wood, S. N. (2003). Thin plate regression splines. *Journal of the Royal Statistical Society Series B*, 65, 95–114. doi:10.1111/1467-9868.00374
- Wood, S. N. (2004). Stable and efficient multiple smoothing parameter estimation for generalized additive models. *Journal of the American Statistical Association*, 99, 673–686. doi:10.1198/016214504000000980
- Wood, S. N. (2006a). *Generalized additive models: An introduction with R*. Boca Raton, FL: Chapman and Hall/CRC.
- Wood, S. N. (2006b). Low-rank scale-invariant tensor product smooths for generalized additive mixed models. *Biometrics*, 62, 1025–1036. doi:10.1111/j.1541-0420.2006.00574.x
- Wood, S. N. (2006c). On confidence intervals for generalized additive models based on penalized regression splines. *Australian & New Zealand Journal of Statistics*, 48, 445–464. doi:10.1111/j.1467-842X.2006.00450.x
- Wood, S. N. (2008). Fast stable direct fitting and smoothness selection for generalized additive models. *Journal of the Royal Statistical Society Series B*, 70, 495–518. doi:10.1111/j.1467-9868.2007.00646.x
- Wood, S. N. (2011). Fast stable restricted maximum likelihood and marginal likelihood estimation of semiparametric generalized linear models. *Journal of the Royal Statistical Society Series B*, 73, 3–36. doi:10.1111/j.1467-9868.2010.00749.x
- Wood, S. N. (2012). Gamm4: Generalized additive mixed models using mgcv and lme4 (R package Version 0.1-6) [Computer software]. Retrieved from <http://CRAN.R-project.org/package=gamm4>
- Wood, S. N., & Augustin, N. H. (2002). GAMs with integrated model selection using penalized regression splines and applications to environmental modelling. *Ecological Modelling*, 157(2–3), 157–177. doi:10.1016/S0304-3800(02)00193-X
- Yang, L., Qin, G., Zhao, N., Wang, C., & Song, G. (2012). Using a generalized additive model with autoregressive terms to study the effects of daily temperature on mortality. *BMC Medical Research Methodology*, 12, 165. doi:10.1186/1471-2288-12-165

## Appendix A

### Generalized Additive Model (GAM) Software

There are currently many software options that support GAMs. Here, we provide a list of options available to fit GAMs, generalized additive mixed models (GAMMs), and semiparametric regression models in the R environment—these are available as R packages on the Comprehensive R Archive Network (CRAN; R Development Core Team, 2010). The earliest R package to offer GAMs was *gam* (Hastie, 2011); however, it is limited to generalized additive models and does not support generalized additive mixed models. One of the oldest most mature packages is *mgcv* (Wood, 2006a; Wood & Augustin, 2002); it supports generalized additive mixed models and is the software used in this article. An extension of

*mgcv* is *gamm4* (Wood, 2012), which uses the mixed modeling package *lme4* (Bates, Maechler, & Bolker, 2012) as the basis for fitting GAMMs rather than the *nmle* package used in *mgcv*. The *SemiPar* package (Ruppert, Wand, & Carroll, 2003) is useful but is less well maintained than the *mgcv* package. The *refund* package offers GAMM functionality (Crainiceanu et al., 2012). The package *RLRsim* provides relevant restricted likelihood ratio tests for a restricted set of cases (Scheipl, Greven, & Kuechenhoff, 2008). Finally, *gammSlice* is the most recent addition, and it supports Markov chain Monte Carlo (MCMC)-based inference for GAMM analyses (Pham & Wand, 2012).

(Appendices continue)

## Appendix B

### R Code for Implementation of the Generalized Additive Model (GAM) and the Generalized Additive Mixed Model (GAMM)

The data should be arranged in long form; for example, the output of a `head (Clip1)` R command should produce output similar to [Table B1](#), and the output of a `tail (Clip1)` R command should produce output similar to [Table B2](#).

Indicative R code used to create models such as those illustrated in [Figure 1](#) is provided here. The most basic command using defaults for the GAMs would be as follows:

```
gamMod1 <- gam(valence~s(time), data=Clip1)
gamMod2 <- gam(valence~s(time), data=Clip2)
gamMod3 <- gam(valence~s(time), data=Clip3)
gamMod4 <- gam(valence~s(time), data=Clip4)
gamMod5 <- gam(valence~s(time), data=Clip5)
gamMod6 <- gam(valence~s(time), data=Clip6)
```

Summary information for GAMs can be printed using the `summary` command, for example, `summary(gamMod1)`.

The most basic command using defaults for the GAMMs would be as follows:

```
gammMod1 <- gamm(valence ~ s(time), data=Clip1, random=list (PartNo=~1),
  correlation=corAR1())
gammMod2 <- gamm(valence ~ s(time), data=Clip2, random=list (PartNo=~1),
  correlation=corAR1())
gammMod3 <- gamm(valence ~ s(time), data=Clip3, random=list (PartNo=~1),
  correlation=corAR1())
gammMod4 <- gamm(valence ~ s(time), data=Clip4, random=list (PartNo=~1),
  correlation=corAR1())
gammMod5 <- gamm(valence ~ s(time), data=Clip5, random=list (PartNo=~1),
  correlation=corAR1())
gammMod6 <- gamm(valence ~ s(time), data=Clip6, random=list (PartNo=~1),
  correlation=corAR1())
```

**Table B1**

*The First Six Lines of the R Data Frame for Data Set 1 Clip 1*

Row	Part no.	Clip	Sex	Time	Valence
1	1	F03b	Female	0.1	0.256
2	2	F03b	Female	0.1	0.767
3	3	F03b	Female	0.1	1.278
4	4	F03b	Female	0.1	0.767
5	5	F03b	Female	0.1	0.000
6	6	F03b	Female	0.1	0.000

**Table B2**

*The Last Six Lines of the R Data Frame for Data Set 1 Clip 1*

Row	Part no.	Clip	Sex	Time	Valence
47995	155	F03b	Male	30	-58.745
47996	156	F03b	Male	30	-28.885
47997	157	F03b	Male	30	-87.167
47998	158	F03b	Male	30	-100.000
47999	159	F03b	Male	30	-61.854
48000	160	F03b	Male	30	-56.492

(table continues)

(Appendices continue)

Retrieving summary information for GAMMs can be printed using the `summary` command; however, `$gam` or `$lme` must be appended to specify which part of the model you require.

```
summary(gamMod1$gam)
summary(gamMod1$lme)
```

The smooth formula is the part with the `s(time)`; this can be altered to place an upper limit on the number of knots by changing the parameter `k` (the basis dimension for the smooth):

```
gamMod7 <- gam(valence~s(time, k=15), data=Clip7)
gammMod7 <- gamm(valence ~ s(time, k=15), data=Clip7, random=list(PartNo=~1),
  correlation = corAR1())
```

Similarly, you can alter the type of smoother using the `bs` command in the smooth formula:

```
"cr"—A penalized cubic regression spline;
"ps"—Eilers and Marx style P-splines;
"ad"—Adaptive smoothers based on "ps";
"tp"—Optimal low rank approximation to thin plate spline.
```

For example,

```
gamMod8 <- gam(valence~s (time, bs="ps"), data=Clip8)
gammMod8 <- gamm(valence ~ s (time, bs="cr"), data=Clip8, random=
  list (PartNo=~1), correlation = corAR1())
```

Varying coefficient models are achieved by including a `by` command within the smooth term:

```
gamMod12 <- gam(valence~s (time, by=Sex), data=Clip12), random=list (PartNo=~1),
gammMod12 <- gamm(valence ~ s (time, by=Sex), data=Clip12,
  correlation = corAR1())
```

Plots like the ones in [Figures 1 and 4](#) can be obtained using

```
plot (gamMod1, shade=TRUE, shade.col="rosybrown2", rug=FALSE, se=T, xlab="Time",
  ylab="Valence", main="Disgust Female Encoder")
plot (gammMod1$gam, shade=TRUE, shade.col="rosybrown2", rug=FALSE, se=T, xlab="Time",
  ylab="Valence", main="Amusement Female Encoder")
```

Received June 12, 2012  
 Revision received May 19, 2013  
 Accepted May 25, 2013 ■