Probabilistic feature analysis of facial perception of emotions

Michel Meudlers, Paul De Boeck, Iven Van Mechelen

Katholieke Universiteit Leuven

Andrew Gelman

Columbia University New York

We thank the Associate Editors and several reviewers for helpful comments. The research reported in this paper was partially supported by the Fund for Scientific Research-Flanders (Belgium) (project G.0207.97 awarded to Paul De Boeck and Iven Van Mechelen), and the Research Fund of K.U. Leuven (F/96/6 fellowship to Andrew Gelman, OT/96/10 project awarded to Iven Van Mechelen, GOA/2000/02 awarded to Paul De Boeck and Iven Van Mechelen, and PDM/02/066 awarded to Paul De Boeck). Correspondence concerning this paper should be addressed to Michel Meudlers, Department of Psychology, Tiensestraat 102, B-3000 Leuven, Belgium. Email: Michel.Meudlers@psy.kuleuven.ac.be
Abstract

According to the hypothesis of configural encoding, the spatial relations between the parts of the face function as an additional source of information in the facial perception of emotions. This paper analyzes experimental data on emotion perception (de Bonis et al., 1999) in order to investigate whether there is evidence for configural encoding in the processing of facial expressions. It is argued that analysis with a probabilistic feature model has several advantages that are not implied by, for example, a generalized linear modeling approach. First, the probabilistic feature model allows us to empirically extract the facial features that are relevant in processing the face, rather than focusing on the features that were manipulated in the experiment. Second, the probabilistic feature model allows for a direct test of the hypothesis of configural encoding as it explicitly formalizes a mechanism for the way in which information about separate facial features is combined in processing the face. Third, the model allows us to account for a complex data structure while still yielding parameters that have a straightforward interpretation.

Keywords: facial expression, emotion perception, probabilistic feature model, Bayesian analysis
1 Introduction

The recognition of facial expressions (FEs) is an important aspect of most face-to-face communications. However, little is known about the mechanisms that lie at the basis of emotion perception, and several theoretical questions are still actively debated in the literature (for an overview, see Massaro, 1998). An important question, for instance, is whether a single holistic cue or multiple cues are used in processing a FE.

According to the hypothesis of configural encoding, the spatial relations between the parts of the face function as an additional source of information when processing a FE. In order to test this hypothesis, one could conduct a typical experiment in which subjects are asked whether they perceive certain emotions in FEs that are manipulated to distort or preserve the spatial relations between the upper and lower halves of the face. Prototypical FEs of basic emotions (Ekman and Friesen, 1976) such as happiness or fear naturally preserve the spatial relations between upper and lower halves, whereas chimerically constructed FEs that consist of upper and lower halves of two distinct FEs expressing basic emotions (see de Bonis et al., 1999) do not preserve such spatial relations. The effect of preserving or distorting spatial relations on the processing and the rating of the face can be evaluated in a straightforward way using a standard analysis technique such as generalized linear modeling (GLM). More specifically, the hypothesis of configurality can be evaluated by testing the interaction between upper and lower facial halves.

However, important drawbacks of such a GLM analysis include that: (1) it does
not yield a direct insight into the relevance of different facial feature configurations for the processing of the FE, and (2) it fails to give insight into the way configurations of features are combined when processing the FE. As an alternative, this paper proposes a probabilistic feature model to analyze the experimental data. In contrast to the GLM approach, this model extracts the relevant features from the data and it formalizes a mechanism for combining information about distinct features in processing the FE. In particular, the probabilistic feature model assumes that the perception of an emotion in a FE depends on two types of events that can be represented as the realization of latent Bernoulli variables: (1) it is assumed that certain features representing properties of the face are activated when a person judges a FE and (2) it is assumed that the activation of features may or may not be a necessary condition for a particular emotion to be perceived in a certain FE. The model further assumes that an emotion will be perceived in a FE if all the required features are activated.

The outline of the paper is as follows: In section 2 we discuss the data that will be analyzed in this paper. We explain the probabilistic feature model and discuss estimation in a Bayesian framework for this model in Section 3. In section 4 we discuss the issue of model selection and model checking. In Section 5 we fit the model to our data and assess the model fit using posterior predictive checks. We conclude in Section 6.
2 Data

de Bonis et al. (1999) conducted an experiment in which they asked raters \( i = 1, \ldots, 18 \) to indicate whether or not they perceived each of a set of emotions \( e = 1, \ldots, 19 \) in different types of facial expressions. The raters were shown 10 photographed faces, corresponding to emotions of happiness and fear for each of 5 different stimulus persons \( s = 1, \ldots, 5 \), from a standard set of pictures of emotional facial expressions (Ekman and Friesen, 1976). In addition, 10 computer-generated chimerical faces were constructed using common morphing techniques, by combining, for each stimulus person, the happy upper half and the fearful lower half, or the fearful upper half and the happy lower half. Happy and fearful faces are denoted as HH (happy upper and happy lower part) and FF (fearful upper and fearful lower part). The chimerical faces are denoted as HF (happy upper and fearful lower part) and FH (fearful upper and happy lower part). Figure 1 shows the prototypical and chimerical faces for one of the stimulus persons. The set of emotions consisted of 19 emotion words taken from a study by Rosenberg and De Boeck (1997). The set included nine positive emotions: admiration, affection, amused, cheerful, connected, enjoyment, interested, relaxed, warm; nine negative emotions: angry, confused, contempt, distant, embarrassed, fearful, pained, repulsed, sad; and a single neutral emotion: surprise.
Figure 2 displays the proportion of raters who perceive each emotion for each of the five pictures within each type of face. As observations may be tied (different pictures within a type of face may elicit a certain emotion from the same proportion of raters), vertical bars with a length proportional to the number of ties are added to the plot.

Inspection of Figure 2 shows that the observations for the happy FE differ in at least two ways from the observations for fearful and chimerical FEs: First, the observed proportions for the happy FE tend to be more extreme and less different across different stimulus persons. Second, for the happy FE (as compared to other types of FEs), positive and negative emotions constitute clear clusters in that most positive emotions are elicited and most negative emotions are not. Furthermore, it is remarkable that so-called basic emotions such as anger and contempt have a rather high probability to be elicited by the chimerical expression HF and that quite different emotional states such as confusion, embarrassed, repulsion and surprise all have a rather high probability to be perceived in prototypical fearful FEs.

3 Probabilistic feature models

Our approach to probabilistic feature analysis is based on the probability matrix decomposition model (Maris et al., 1996) which is a method of data analysis for
two-way frequency tables. In most applications, the entries of such tables reflect the numbers of raters according to whom elements in the rows and columns of the table are related; in such cases, high (low) frequencies indicate strong (weak) associations between the corresponding elements. In the present paper the model of Maris et al. will be referred to as the probabilistic feature model (PFM).

PFMs have been applied to analyze a wide variety of phenomena in different substantive contexts such as psychiatric diagnosis (Maris et al., 1996; Gelman et al., accepted), marketing research (Candel and Maris, 1997), cross-cultural research (Meulders et al., 2001b), and personality assessment (Meulders et al., 2002; Meulders et al., 2003).

To explain emotion perception in FE, PFMs assume a twofold process: First, it is assumed that, during the perception of some FE, certain facial features may (or may not) be activated, and that these features may (or may not) be linked to the emotion that is judged. Second, it is assumed that the rating of the face follows from applying the conjunctive rule that the emotion will be elicited by the FE under study if all features that are linked to the emotion are also activated in the FE. PFMs are especially suited for modelling the data of the current experiment because they extract the relevant features from the data through a process of feature activation and because they include a (conjunctive) mechanism for explaining the combination of relevant facial features in emotion perception. In contrast, a GLM which models the sum of the binary responses of the 18 raters for each combination of FE, stimulus person, and emotion as binomial with a linear link and a linear
predictor that depends on the upper and lower facial halves and their interaction, would take the experimentally manipulated features for granted. Moreover, this GLM would not provide an explicit model for the way in which information from distinct facial features is combined in processing the face.

### 3.1 The model

Let binary variables $D_{itse}$ equal 1 if rater $i$ perceives emotion $e$ in FE $t$ of stimulus person $s$, and 0 otherwise, and let $d_{itse}$ denote a specific observation. The index $t$ indicates the four types of FEs (i.e., HH, HF, FH and FF). The number of raters who perceive a certain emotion in a particular facial expression is denoted by variables $D_{+tse} = \sum_i D_{itse}$. PFM$s$ assume that each observed response $D_{itse}$ is obtained as a mapping of latent variables $X^{tsf}_{ei}$ and $Y^{ref}_{tsi}$ ($f = 1, \ldots, F$) which have the following interpretation:

$$X^{tsf}_{ei} = \begin{cases} 1 & \text{if feature } f \text{ representing properties of a face is activated in FE } t \\
0 & \text{otherwise} \end{cases}$$

with $f$ representing properties of a face is activated in FE $t$ of stimulus person $s$ when rater $i$ judges whether emotion $e$ is perceived in this FE.

$$Y^{ref}_{tsi} = \begin{cases} 1 & \text{if the activation of feature } f \text{ is required for emotion } e \text{ to be perceived when rater } i \text{ is judging the association between FE } t \\
0 & \text{otherwise} \end{cases}$$

of stimulus person $s$ and emotion $e$. 


The model further assumes that:

\[ X_{ts}^{ei} \sim \text{Bern}(\sigma_{tsf}) \]  
\[ Y_{tsi}^{ef} \sim \text{Bern}(\rho_{ef}), \]  

with all latent variables being independent. The parameters \( \sigma_{tsf} \) and \( \rho_{ef} \) are further denoted as feature-activation and feature-emotion probabilities, respectively.

From a psychological point of view, the independence assumptions regarding feature-activation in FEs and the realization of feature-emotion links are motivated as follows: First, the assumption that feature-activation is renewed at each encounter \((i, t, s, e)\) implies an independent processing of a specific FE by rater \(i\) for each emotion \(e\). This assumption may be meaningful as FEs are complex stimuli that do not necessarily activate the same features each time they are being perceived. Second, the assumption that feature-emotion links are renewed at each encounter \((i, t, s, e)\) links up with the concept of fuzzy emotion definitions (see, e.g., Russell, 2003). Note that the postulate of the emotion links being renewed at each new judgement is only one possible meaningful way to specify the psychological process of activation of a fuzzy emotion-definition of a rater who makes a judgement. Other possible specifications are possible but these have not been further developed here as the present paper focuses on the hypothesis of configural encoding.

Once the \(2 \times F\) latent variables \(X_{ei}^{ts} = (X_{ei}^{ts1}, \ldots, X_{ei}^{tsF})\) and \(Y_{tsi}^{e} = (Y_{tsi}^{e1}, \ldots, Y_{tsi}^{eF})\) are realized for a particular combination \((i, t, s, e)\), the observed response \(d_{itse}\) is obtained by applying the conjunctive rule that the emotion will be perceived in the FE if all the features that are linked to the emotion are also activated in the FE,
that is
\[ D_{itse} = 1 \Leftrightarrow \forall f : X_{ei}^{tsf} \geq Y_{tsi}^{ef}. \] (3)

This mapping rule can be formally expressed by specifying the conditional distribution of the observation given the underlying latent variables as follows:

\[ p(d_{itse}|x_{ei}^{ts}, y_{tsi}^{ef}) = \left[ \prod_f (1 - (1 - x_{ei}^{tsf})y_{tsi}^{ef}) \right]^{d_{itse}} \left[ 1 - \prod_f (1 - (1 - x_{ei}^{tsf})y_{tsi}^{ef}) \right]^{(1-d_{itse})}, \]

with \( x_{ei}^{ts} \) and \( y_{tsi}^{ef} \) denoting vectors of latent realizations underlying observation \( d_{itse} \).

From (1), (2), and (3) we may derive that

\[ P(D_{itse} = 1|\sigma, \rho) = \prod_f P(X_{ei}^{tsf} \geq Y_{tsi}^{ef}|\sigma, \rho) \]
\[ = \prod_f [1 - P(X_{ei}^{tsf} < Y_{tsi}^{ef}|\sigma, \rho)] \]
\[ = \prod_f [1 - P(X_{ei}^{tsf} = 0, Y_{tsi}^{ef} = 1|\sigma, \rho)] \]
\[ = \prod_f [1 - (1 - \sigma_{tsf})\rho_{ef}]. \]

with \( \sigma \) and \( \rho \) being vectors that comprise all parameters \( \sigma_{tsf} \) and \( \rho_{ef} \), respectively.

### 3.2 Estimation

We will follow a Bayesian approach to obtain statistical inferences for the PFM, using a hierarchical model to capture the variation of parameters from different stimulus persons. Let \( d \) be a vector that comprises all observations and let \( \phi \) be a vector of hyperparameters. For the hierarchical PFM, inferences are based on the posterior distribution \( p(\sigma, \rho, \phi|d) \), which is proportional to the product of the likelihood \( p(d|\sigma, \rho) \), the prior \( p(\sigma, \rho|\phi) \) and the hyperprior \( p(\phi) \). The specific
form of the likelihood and the families of densities to be used for the prior and the hyperprior distributions will be discussed next.

1. **Likelihood.** As each observation $d_{itse}$ is based on independent realizations of Bernoulli variables $X_{eti}^t$ and $Y_{etsi}^e$, it follows that $D_{itse} \sim \text{Bern}(\pi_{tse})$, with $\pi_{tse} = \prod_f [1 - (1 - \sigma_{tsf}) \rho_{ef}]$. Consequently, the likelihood of the data $d$ can be expressed as

$$p(d|\sigma, \rho) = \prod_i \prod_t \prod_s \prod_e \prod_f (\pi_{tse})^{d_{itse}} (1 - \pi_{tse})^{1 - d_{itse}}.$$ (4)

The above also implies that $D_{itse} \sim \text{Bin}(18, \pi_{tse})$, so that the PFM actually models frequencies $d_{+tse}$ as a (nonlinear) function of the parameters $\sigma$ and $\rho$.

2. **Augmented likelihood.** The augmented likelihood is defined as the joint distribution of all observed and latent variables, that is, $p(d, x, y|\sigma, \rho)$. An elegant property of PFMs is that the augmented likelihood can be factorized in three parts that reflect the assumptions of the two-process model, namely the activation of features in the FE (i.e. $p(x|\sigma)$), the realization of links between emotions and features (i.e. $p(y|\rho)$) and the conjunctive mapping rule (i.e. $p(d|x, y)$). As a result, the augmented likelihood has a simple structure as it is proportional to a product of Bernoulli likelihoods:

$$p(d, x, y|\sigma, \rho) = p(d|x, y)p(x|\sigma)p(y|\rho) \quad = \quad p(d|x, y)$$

$$\times \prod_i \prod_t \prod_s \prod_e \prod_f (\sigma_{tsf})^{x_{esi}^f} (1 - \sigma_{tsf})^{1 - x_{esi}^f} (\rho_{ef})^{y_{etsi}^e} (1 - \rho_{ef})^{1 - y_{etsi}^e}.$$
The simple structure of the augmented likelihood may be exploited by using an EM (Dempster et al., 1977) or data augmentation algorithm (Tanner and Wong, 1987) for parameter estimation.

3. Prior distribution. The prior distribution of the PFM can be specified by considering certain groups of parameters in $(\sigma, \rho)$ to be a sample of independent beta distributions. In the present context, we consider separate population distributions for feature-activation probabilities associated to a specific feature and a specific type of face. In this way, feature-activation probabilities for FEs of different stimulus persons are made structurally dependent which makes sense from a substantive point of view. The mean of the population distribution reflects the average level of clustered probabilities whereas the variance of the population distribution is a measure of the heterogeneity of clustered probabilities associated to FEs of different stimulus persons. Furthermore, the feature-emotion probabilities are considered as a sample of a single population distribution. The resulting prior distribution is

$$p(\sigma, \rho|\phi) = \prod_t \prod_s \prod_f \text{Beta}(\sigma_{tsf}|a_{\sigma}, b_{\sigma}) \prod_e \prod_f \text{Beta}(\rho_{ef}|a_{\rho}, b_{\rho}),$$

with $\phi$ a vector of hyperparameters $(\alpha, \beta)$. The augmented posterior distribution $p(\sigma, \rho|d, x, y, \phi)$ has the same functional form as the prior, and so it can be expressed as a product of beta distributions:

$$p(\sigma, \rho|d, x, y, \phi) = \prod_t \prod_s \prod_f \text{Beta}(\sigma_{tsf}|a_{\sigma} + \sum_e x_{ei} t_{sf}, b_{\sigma} + \sum_e (1 - x_{ei} t_{sf}))$$

$$\times \prod_e \prod_f \text{Beta}(\rho_{ef}|a_{\rho} + \sum_t \sum_s y_{tsi} e_{sf}, b_{\rho} + \sum_t \sum_s (1 - y_{tsi} e_{sf})).$$
4. **Hyperprior distribution.** For each of the beta priors in (5), the hyperparameters are assigned a distribution \( p(\alpha, \beta) \). As we have no specific hypotheses about the mean or the variance of the population distributions we specify uniform distributions in the interval \([0, 1]\) for the mean \( u = \frac{\alpha}{\alpha + \beta} \) and the ratio \( v = \frac{1}{\alpha + \beta} \). For the feature-activation and feature-emotion probabilities, we denote the transformed hyperparameters as \((u_{t\sigma}^{tf}, v_{t\sigma}^{tf})\) and \((u_\rho, v_\rho)\), respectively. As we have only a sample of five parameters to estimate the hyperparameters \((u_{t\sigma}^{tf}, v_{t\sigma}^{tf})\) associated to FE \(t\) and feature \(f\), we have only little information to get a reliable estimate of the parameter \(v_{t\sigma}^{tf}\), which may be regarded as a measure of the variability of the distribution. Therefore we impose the restriction that \(v_{t\sigma}^{tf} = v_{t\sigma}(t = 1, \ldots, 4; f = 1, \ldots, F)\).

The specification of a uniform prior for \(u\) and \(v\) is standard practice (for a similar example, see Gelman et al., 2003, p. 128) with the understanding that if the posterior distribution is insufficiently informative, one could go back and assign a more informative prior distribution based on the scientific literature. In the present application it turns out that \(u\) and \(v\) are estimated relatively precisely (and are not close to the boundaries), which indicates that the data are sufficiently informative.

Using \(\phi^*\) to denote the entire collection of transformed hyperparameters, the joint distribution \(p(\sigma, \rho, \phi^*)\) can be expressed as

\[
p(\sigma, \rho, \phi^*) = \prod_t \prod_s \prod_f \text{Beta} \left( \sigma_{tsf} \left| \frac{u_{t\sigma}^{tf}}{v_\sigma}, \frac{1 - u_{t\sigma}^{tf}}{v_\sigma} \right. \right) \prod_e \prod_f \text{Beta} \left( \rho_{ef} \left| \frac{u_\rho}{v_\rho}, \frac{1 - u_\rho}{v_\rho} \right. \right)
\]
\begin{align*}
\times \prod_{t, f} U(u_\sigma^f|0, 1)U(v_\sigma|0, 1)U(u_\rho|0, 1)U(v_\rho|0, 1).
\end{align*}

Samples from the posterior distribution \( p(\sigma, \rho, \phi^*|d) \) can be drawn using the Gibbs sampler, drawing directly from the conjugate full conditional posterior distributions for the latent parameters \( x_{tsi}^{ts}, y_{tsi}^{tsi}, \sigma_{tsf}, \rho_{ef} \) and using the Metropolis algorithm to update the hyperparameters \( u_\sigma^f, v_\sigma, u_\rho, v_\rho \) in turn. Convergence can be monitored using the multiple-sequence diagnostic of Gelman and Rubin (1992), and then the posterior sample can be used to derive point estimates and \( 100 \times (1 - \alpha) \% \) posterior intervals of the parameters.

4 Model selection and model checking

An important topic in fitting PFMs is to choose the number of features so that the model has an optimal balance between complexity and goodness of fit. In this paper we use the Deviance Information Criterion (DIC, Spiegelhalter et al., 2002) for choosing between models with different numbers of features. The DIC is especially suited for comparing complex hierarchical models in which the number of parameters is not clearly defined and is easily computed on the basis of the posterior sample. The model with the lowest DIC value should be selected. In Appendix A, we describe the computation of the DIC for PFMs.

As model selection criteria only concern the relative fit of models it is recommended to evaluate whether the selected model captures important aspects of the data and whether it fits the data in a global way. The sample of the posterior may
serve as a basis for model evaluation via the use of posterior predictive checks (PPCs; Gelman et al., 2003). In Appendix B, we define a global goodness-of-fit test for the PFM and we describe computational procedures for computing Bayesian $p$-values.

5 Analysis

We performed inference for hierarchical PFMs with 1, 2, 3, or 4 features. For these models DIC values equal 8452, 6222, 5842 and 5779, respectively, and Bayesian $p$-values of the Pearson-$\chi^2$ discrepancy measure (see Appendix A) equal 0, .004, .094 and .513, respectively. Hence, the four-feature model which is selected on the basis of the DIC also fits the data in a global way.

In the following paragraphs we give a detailed presentation of the results for the four-feature model. We focus on the following substantive questions: (1) Are there important differences between feature-activation probabilities associated to FEs of the same type from different stimulus persons? (2) Which features are relevant when processing a particular type of FE? (3) Do the data provide evidence for configural encoding? (4) Does the model respect the multivariate nature of the data, that is, does it capture higher-order interactions between the variables that are manipulated in the experiment, namely, the upper-half (U), the lower-half (L), the stimulus person (S), and the emotion (E)?

The posterior distributions of the feature-activation probabilities $\sigma$ (not reported in this paper) indicate that the activation of features in prototypical FEs is usually reliable (small posterior intervals) and consistent across FEs of different stimulus
persons (posterior intervals strongly overlap). For chimerical FEs, however, feature activation is sometimes unreliable (especially for the happy-upper feature) and more inconsistent across FEs of different stimulus persons.

Inspection of the estimated hyperparameters $u^{lf}_i$ in Figure 3 indicates that the four features that are extracted by the model are the facial feature configurations that are manipulated in the experiment; they can be labeled Happy Upper (HU), Happy Lower (HL), Fear Upper (FU), and Fear Lower (FL). This interpretation follows because FEs tend to have high feature-activation probabilities for features that correspond to their upper and lower parts and because they tend to have low feature-activation probabilities for features that do not correspond to their upper and lower parts. As an exception, the FU feature has a moderate probability to be activated in the Happy-Fearful FE. However, this result is mainly caused by moderate activation probabilities for the FEs of two stimulus persons (average $\sigma_{HF,FU}$ of .55) whereas the activation probabilities for the other three stimulus persons are lower (average $\sigma_{HF,FU}$ of .30).

As indicated in the section on model checking four features are needed to obtain a sufficient fit to the data. However, further analysis also indicates that all features do not equally contribute in fitting the data. In particular, excluding the features FL, HL, FU and HU from the four-feature model decreases the variance that is accounted for in the observed frequencies by the model from 94% to 52%, 42%, 67% and 85%,
respectively. Hence, the lower parts of the face provide relatively more information for processing emotions in FEs. This finding is also supported by the results of PFM with less than four features: The two-feature model extracts features that can be interpreted as HL and FL and the three-feature model additionally extracts a feature that can be labeled FU. Finally, the HU feature contributes least to model fit.

The PFM is especially suited to investigate the hypothesis of configural encoding for modeling the perception of emotions in FEs. It assumes a conjunctive rule for combining the information of separate facial features. More specifically, the model assumes that an emotion will be perceived in a FE if all the features that are linked to the FE are also activated in the FE. For a particular emotion, configural encoding then shows up if that emotion has high feature-emotion probabilities for two features pertaining to the two halves of the face.

As shown in Figure 4, most positive emotions, however, only require the activation of the happy lower feature in order to be perceived. Exceptions are the emotions “affection” and “relaxed” that also show a moderately strong link with the happy-upper feature. On the other hand, most negative emotions require the activation of two features in order to be perceived: The emotion “fear” shows strong links with the features FU and FL and not with the other features (HU and HL). Other emotional states such as confused, embarrassed, and repulsed show a simi-
lar pattern of feature-emotion probabilities, but posterior intervals are often larger than for “fear”. Furthermore, the basic emotions “anger” and “contempt” require the activation of HU and FL which means that they have a high probability to be perceived in the chimerical Happy-Fearful FE. Finally, the neutral emotion surprise is only linked to the Fear-Upper feature.

To evaluate whether the PFM can capture interactions between the design variables U, L, S, and E, we applied posterior predictive checks (see Appendix A) using the linear components of the ANOVA model for a completely randomized factorial design (Kirk, 1995, p. 441) with four factors as test statistics. More specifically, this ANOVA model partitions the total sum of squares $SS_{TOT} = \sum_u \sum_l \sum_s \sum_e (d_{ulse} - \bar{d}_{+,...})$ into main effects, second order interactions and so on (see Table 1). As the total sum of squares can vary across replicated data sets, we use the proportion of the variation accounted for by interaction components as test statistics. Table 1 shows the observed values of these test statistics and the corresponding posterior predictive $p$-values that were computed for the four-feature model. The results of the analysis indicate that most components are well represented by the model in the sense that posterior predictive $p$-values are not close to 0 or 1. As an exception, the ULS and ULSE interactions are slightly overestimated by the model and the SE interaction tends to be underestimated. The third-order interactions ULE, USE, and LSE, which account together for about 10% of the variation in the observed
frequencies, are all captured by the model.

6 Discussion

In this paper we analyzed experimental data on emotion perception from de Bonis et al. (1999). Our analysis goes beyond the analysis of de Bonis et al. in several ways: We presented a fully Bayesian analysis of a hierarchical variant of the PFM (including Bayesian model selection and model checking) whereas de Bonis et al. used an EM algorithm to obtain inferences for a non-hierarchical PFM and presented no goodness-of-fit tests. A comparison of the two models (not included in this paper) indicates that the hierarchical PFM yields a better fit to the data than the non-hierarchical model both in a global way as with respect to specific aspects of the data. Furthermore, the present paper focused on evaluating the hypothesis of configural encoding whereas this topic was not discussed by de Bonis et al. More specifically, it was argued in this paper that a PFM has several advantages for testing the configurality hypothesis compared to a GLM approach. First, the PFM allows us to empirically identify the relevant features of emotion perception from the data whereas a GLM would take the features that are manipulated in the experiment for granted. Note that the performance of PFMs to extract relevant features from the data was found to be very good in a simulation study when the DIC was used as the selection criterion (see Meulders et al., 2003, p. 71). In the present application, the PFM identified the four manipulated facial halves to be the relevant features for emotion perception and, as such, it provided a check on the manipulations in-
volved in the experiment. Second, the PFM is especially suited to investigate the configurality hypothesis as it includes a (conjunctive) mechanism for modeling the combination of facial features in processing the FE. The probabilistic feature analysis showed that there is only weak evidence for configural encoding in perceiving positive emotions because the smile on the happy lower halve is often sufficient for perceiving such emotions. Furthermore, the analysis indicated that there is strong evidence for configural encoding in perceiving fear. Third, as already indicated by the Bonis et al., the PFM shows that basic emotions other than happiness or fear can be elicted by chimerical FEs; the analysis provides a clear picture of the latter phenomenon as it indicates which features should be activated for an emotion to be perceived. Finally, we may note that the distinction between PFMs and GLMs is not always necessary since PFMs can in some cases be expressed as GLMs (Meulders et al., 2001a). For instance, a PFM in which either feature-activation probabilities or feature-emotion probabilities are considered to be fixed would be equivalent to a GLM with a binomial random component, a log link and a linear predictor that depends on the type of face and the emotion.

7 Appendix A: Computation of DIC

Using $\theta$ as notation for parameters involved in the likelihood of the model, the DIC is defined as

$$DIC = \overline{D(\theta)} + p_D,$$
with $D(\bar{\theta})$ being the posterior mean of the deviance of the model and with $p_D$ being an estimate of the effective number of parameters in the model, namely, $p_D = D(\bar{\theta}) - D(\bar{\theta})$, with $\bar{\theta}$ being the mean of the posterior sample. For the PFM, the deviance function is specified as $-2\log p(d|\sigma, \rho)$.

8 Appendix B: Computation of Bayesian $p$-values

The posterior sample can easily be used to evaluate fit measures with the technique of posterior predictive checks (PPCs) (Rubin, 1984; Meng, 1994; Gelman et al., 1996). In describing the computational procedures for Bayesian $p$-values we may distinguish between two types of fit measures: statistics $T(d)$ that only depend on the data $d$ and discrepancy measures $T(d, \theta)$ that depend on both the parameters $\theta$ and the data. An example of the latter type of measure is the Pearson-$\chi^2$ measure for evaluating global goodness of fit. In the context of the present application, the Pearson-$\chi^2$ measure is defined as:

$$\chi^2(d, \theta) = \sum_{t,s,e} \frac{[d_{tse} + E(D_{tse}|\theta)]^2}{\text{Var}(D_{tse}|\theta)},$$

with $E(D_{tse}|\theta) = 18\pi_{tse}$ and $\text{Var}(D_{tse}|\theta) = 18\pi_{tse}(1 - \pi_{tse})$. For statistics, the PPC $p$-value can be computed by generating new data sets $d^{rep}$ (using the draws from the observed posterior) and by computing the proportion of replicated data sets in which $T(d^{rep}) \geq T(d)$. For discrepancy measures, the $p$-value is computed as the proportion of replicated data sets in which realized discrepancies $T(d^{rep}, \theta)$ exceed or equal observed discrepancies $T(d, \theta)$. 
References

mode frequency data: probability matrix decomposition and two alternatives.

theory of facial perception of emotions. Comptes Rendus de l’Académie des
Sciences, Série III, 322, 669-675.

mation from incomplete data via the EM algorithm (with discussion). Journal

Consulting Psychologists Press.


Multiple imputation for model checking: completed-data plots with missing and latent data. *Biometrics.*


Table 1: Observed proportion of the variation accounted for by components of ANOVA model and posterior predictive $p$-value for a four-feature model.

<table>
<thead>
<tr>
<th>component</th>
<th>observed proportion</th>
<th>$p$-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>$SS_U$</td>
<td>.0008</td>
<td>.36</td>
</tr>
<tr>
<td>$SS_L$</td>
<td>.0172</td>
<td>.57</td>
</tr>
<tr>
<td>$SS_S$</td>
<td>.0040</td>
<td>.53</td>
</tr>
<tr>
<td>$SS_E$</td>
<td>.2047</td>
<td>.67</td>
</tr>
<tr>
<td>$SS_{UL}$</td>
<td>.0001</td>
<td>.75</td>
</tr>
<tr>
<td>$SS_{US}$</td>
<td>.0010</td>
<td>.81</td>
</tr>
<tr>
<td>$SS_{UE}$</td>
<td>.1446</td>
<td>.29</td>
</tr>
<tr>
<td>$SS_{LS}$</td>
<td>.0043</td>
<td>.68</td>
</tr>
<tr>
<td>$SS_{LE}$</td>
<td>.4782</td>
<td>.41</td>
</tr>
<tr>
<td>$SS_{SE}$</td>
<td>.0336</td>
<td>.06</td>
</tr>
<tr>
<td>$SS_{ULS}$</td>
<td>.0002</td>
<td>.98</td>
</tr>
<tr>
<td>$SS_{ULE}$</td>
<td>.0238</td>
<td>.64</td>
</tr>
<tr>
<td>$SS_{USE}$</td>
<td>.0415</td>
<td>.43</td>
</tr>
<tr>
<td>$SS_{LSE}$</td>
<td>.0357</td>
<td>.24</td>
</tr>
<tr>
<td>$SS_{ULSE}$</td>
<td>.0103</td>
<td>1.00</td>
</tr>
</tbody>
</table>
Figure 1: Happy (HH) and fearful (FF) prototypical facial expressions and chimerical facial expressions with happy upper half and fearful lower half (HF) and with fearful upper half and happy lower half (FH) for one stimulus person.
Figure 2: Graphical representation of the proportion of raters who perceive a particular emotion in happy (HH), fearful (FF) and chimerical (HF and FH) facial expressions from 5 stimulus persons. Each distinct observed proportion is represented as a dot; ties are indicated by vertical bars with length proportional to the number of ties.
Figure 3: Posterior median (o) and 95% Posterior Interval (—) of hyperparameters $u_\sigma^f$ of feature activation probabilities for a four-feature model.
Figure 4: Posterior median (o) and 95% Posterior Interval (—) of feature-emotion probabilities of four-feature model.