Appl. Statist. (2005)
54, *Part* 4, *pp.* 781–793

Probabilistic feature analysis of facial perception of emotions

Michel Meulders, Paul De Boeck and Iven Van Mechelen

Katholieke Universiteit Leuven, Belgium

and Andrew Gelman

Columbia University, New York, USA

[Received October 2000. Final revision November 2004]

Summary. According to the hypothesis of configural encoding, the spatial relationships between the parts of the face function as an additional source of information in the facial perception of emotions. The paper analyses experimental data on the perception of emotion 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 modelling approach. First, the probabilistic feature model allows us to extract empirically 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 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: Bayesian analysis; Facial expression; Perception of emotion; Probabilistic feature model

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 the perception of emotion 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 an FE.

According to the hypothesis of configural encoding, the spatial relationships between the parts of the face function as an additional source of information when processing an FE. To test this hypothesis, we 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 relationships 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 relationships between the 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) and Morris *et al.* (2002)) do not preserve such spatial relationships. The effect of preserving or distorting spatial relationships on the processing and the rating of the face can be

Address for correspondence: Michel Meulders, Department of Psychology, Katholieke Universiteit Leuven, Tiensestraat 102, B-3000 Leuven, Belgium.
E-mail: Michel.Meulders@psy.kuleuven.ac.be

 0035–9254/05/54781

evaluated in a straightforward way by using a standard analysis technique such as generalized linear modelling. More specifically, the hypothesis of configural encoding (i.e. the spatial relationships between the parts of the face function as an additional source of information when processing an FE) can be evaluated by testing the interaction between upper and lower facial halves.

However, important drawbacks of such a generalized linear model analysis include that

(a) it does not yield a direct insight into the relevance of different facial feature configurations for the processing of the FE and
(b) it fails to give insight into the way that configurations of features are combined when processing the FE.

As an alternative, this paper proposes a probabilistic feature model to analyse the experimental data. In contrast with the generalized linear modelling 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 an FE depends on two types of event that can be represented as the realization of latent Bernoulli variables:

(a) it is assumed that certain features representing properties of the face are activated when a person judges an FE and
(b) 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 an 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 analysed 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 by using posterior predictive checks. We conclude in Section 6.

The data that are analysed in the paper can be obtained from

`http://www.blackwellpublishing.com/rss`

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 FEs. The raters were shown 10 photographed faces, corresponding to emotions of happiness and fear for each of five different stimulus people ($s = 1, \ldots, 5$), from a standard set of pictures of emotional FEs (Ekman and Friesen, 1976). In addition, 10 computer-generated chimerical faces were constructed by 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). Fig. 1 shows the prototypical and chimerical faces for one of the stimulus people. 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,

Fig. 1. Happy (HH) and fearful (FF) prototypical FEs and chimerical FEs with a happy upper half and fearful lower half (HF) and a fearful upper half and happy lower half (FH) for one stimulus person

relaxed and warm), nine negative emotions (angry, confused, contempt, distant, embarrassed, fearful, pained, repulsed and sad) and a single neutral emotion (surprise).

Fig. 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 that is proportional to the number of ties are added to the plot.

Inspection of Fig. 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 people. Second, for the happy FE (compared with other types of FE), 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 quite a high probability of being elicited by the chimerical expression HF and that quite different emotional states such as confusion, embarrassment, repulsion and surprise all have quite a high probability of being 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 and low frequencies indicate respectively strong and weak associations between the corresponding elements. In this

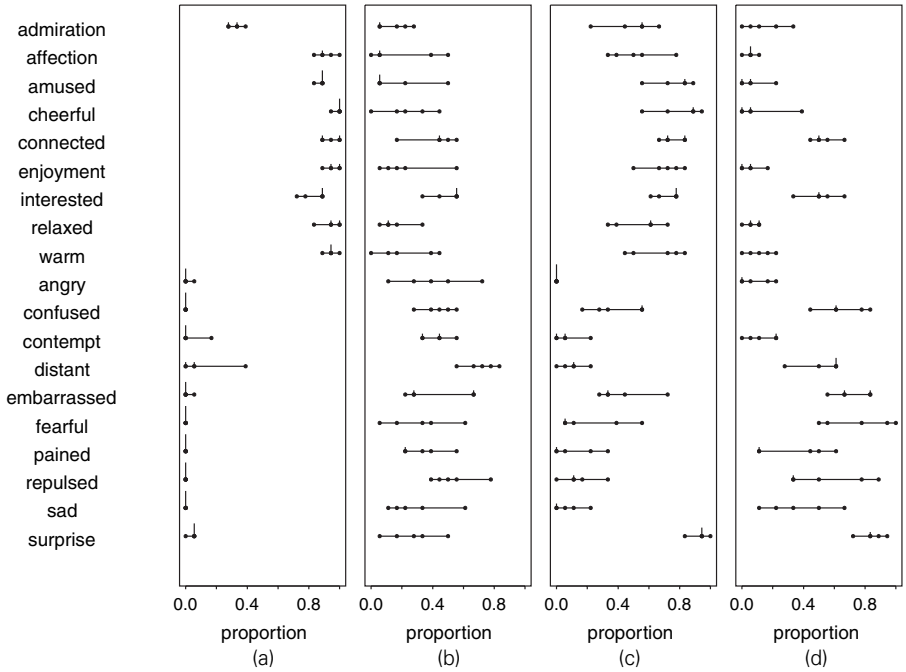

Fig. 2. Graphical representation of the proportion of raters who perceive a particular emotion in (a) happy (HH), (b) fearful (FF), (c) chimerical HF and (d) chimerical FH facial expressions from five stimulus people: •, observed proportion; |, ties, with length proportional to the number of ties

paper the model of Maris *et al.* (1996) will be referred to as the probabilistic feature model (PFM).

PFMs have been applied to analyse a wide variety of phenomena in various substantive contexts such as psychiatric diagnosis (Maris *et al.*, 1996; Gelman *et al.*, 2004), marketing research (Candel and Maris, 1997), cross-cultural research (Meulders, De Boeck, Van Mechelen, Gelman and Maris, 2001) and personality assessment (Meulders *et al.*, 2002, 2003).

To explain the perception of emotion in FEs, 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 to 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 the perception of emotion. In contrast, a generalized linear model 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 generalized linear model 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 equal 0 otherwise, and let d_{itse} denote a specific observation. The index t indicates the four types of FE (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 $D_{+tse} = \Sigma_i D_{itse}$. PFMs assume that each observed response D_{itse} is obtained as a mapping of latent variables X_{ei}^{tsf} and Y_{tsi}^{ef} ($f = 1, \ldots, F$) which have the following interpretation:

$$X_{ei}^{tsf} = \begin{cases} 1 & \text{if feature } f \text{ representing properties of a face is activated in FE } t \\ & \text{of stimulus person } s \text{ when rater } i \text{ judges whether emotion } e \text{ is} \\ & \text{perceived in this FE,} \\ 0 & \text{otherwise;} \end{cases}$$

$$Y_{tsi}^{ef} = \begin{cases} 1 & \text{if the activation of feature } f \text{ is required for emotion } e \text{ to be} \\ & \text{perceived when rater } i \text{ is judging the association between FE } t \\ & \text{of stimulus person } s \text{ and emotion } e, \\ 0 & \text{otherwise.} \end{cases}$$

The model further assumes that

$$X_{ei}^{tsf} \sim \text{Bern}(\sigma_{tsf}), \tag{1}$$

$$Y_{tsi}^{ef} \sim \text{Bern}(\rho_{ef}), \tag{2}$$

with all latent variables being independent. The parameters σ_{tsf} and ρ_{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 that 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, for example, Russell (2003)).The postulate that emotion links are renewed at each new judgment is only one possible meaningful way of specifying the psychological process of activation of a fuzzy emotion definition of a rater who makes a judgment. 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 $\mathbf{X}_{ei}^{ts} = (X_{ei}^{ts1}, \ldots, X_{ei}^{tsF})$ and $\mathbf{Y}_{tsi}^{e} = (Y_{tsi}^{e1}, \ldots, Y_{tsi}^{eF})$ have been 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, i.e.

$$D_{itse} = 1 \Leftrightarrow \forall f : X_{ei}^{tsf} \geqslant Y_{tsi}^{ef}. \tag{3}$$

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

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

with $\mathbf{x}_{ei}^{ts}$ and $\mathbf{y}_{tsi}^{e}$ denoting vectors of latent realizations underlying observation d_{itse}. From expressions (1)–(3) we may derive that

$$
\begin{aligned}
P(D_{itse}=1|\boldsymbol{\sigma},\boldsymbol{\rho}) &= \prod_f P(X_{ei}^{tsf} \geqslant Y_{tsi}^{ef}|\boldsymbol{\sigma},\boldsymbol{\rho}) \\
&= \prod_f \{1 - P(X_{ei}^{tsf} < Y_{tsi}^{ef}|\boldsymbol{\sigma},\boldsymbol{\rho})\} \\
&= \prod_f \{1 - P(X_{ei}^{tsf} = 0, Y_{tsi}^{ef} = 1|\boldsymbol{\sigma},\boldsymbol{\rho})\} \\
&= \prod_f \{1 - (1-\sigma_{tsf})\rho_{ef}\},
\end{aligned}
$$

with $\boldsymbol{\sigma}$ and $\boldsymbol{\rho}$ being vectors that comprise all parameters σ_{tsf} and ρ_{ef} respectively.

3.2. Estimation

We shall follow a Bayesian approach to obtain statistical inferences for the PFM, using a hierarchical model to capture the variation of parameters from different stimulus people. Let $\mathbf{d}$ be a vector that comprises all observations and let $\boldsymbol{\phi}$ be a vector of hyperparameters. For the hierarchical PFM, inferences are based on the posterior distribution $p(\boldsymbol{\sigma},\boldsymbol{\rho},\boldsymbol{\phi}|\mathbf{d})$, which is proportional to the product of the likelihood $p(\mathbf{d}|\boldsymbol{\sigma},\boldsymbol{\rho})$, the prior $p(\boldsymbol{\sigma},\boldsymbol{\rho}|\boldsymbol{\phi})$ and the hyperprior $p(\boldsymbol{\phi})$. 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.

3.2.1. Likelihood

As each observation d_{itse} is based on independent realizations of Bernoulli variables $\mathbf{X}_{ei}^{ts}$ and $\mathbf{Y}_{tsi}^{e}$ it follows that $D_{itse} \sim \text{Bern}(\pi_{tse})$, with $\pi_{tse} = \Pi_f \{1-(1-\sigma_{tsf})\rho_{ef}\}$. Consequently, the *likelihood* of the data $\mathbf{d}$ can be expressed as

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

This also implies that $D_{+tse} \sim \text{Bin}(18, \pi_{tse})$, so that the PFM actually models frequencies d_{+tse} as a (non-linear) function of the parameters $\boldsymbol{\sigma}$ and $\boldsymbol{\rho}$.

3.2.2. Augmented likelihood

The augmented likelihood is defined as the joint distribution of all observed and latent variables, i.e. $p(\mathbf{d},\mathbf{x},\mathbf{y}|\boldsymbol{\sigma},\boldsymbol{\rho})$. 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. $p(\mathbf{x}|\boldsymbol{\sigma})$), the realization of links between emotions and features (i.e. $p(\mathbf{y}|\boldsymbol{\rho})$) and the conjunctive mapping rule (i.e. $p(\mathbf{d}|\mathbf{x},\mathbf{y})$). As a result, the augmented likelihood has a simple structure as it is proportional to a product of Bernoulli likelihoods:

$$
\begin{aligned}
p(\mathbf{d},\mathbf{x},\mathbf{y}|\boldsymbol{\sigma},\boldsymbol{\rho}) &= p(\mathbf{d}|\mathbf{x},\mathbf{y})\ p(\mathbf{x}|\boldsymbol{\sigma})\ p(\mathbf{y}|\boldsymbol{\rho}) \\
&= p(\mathbf{d}|\mathbf{x},\mathbf{y}) \prod_i \prod_t \prod_s \prod_e \prod_f \sigma_{tsf}^{x_{ei}^{tsf}} (1-\sigma_{tsf})^{1-x_{ei}^{tsf}} \rho_{ef}^{y_{tsi}^{ef}} (1-\rho_{ef})^{1-y_{tsi}^{ef}}.
\end{aligned}
$$

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.

3.2.3. Prior distribution

The prior distribution of the PFM can be specified by considering certain groups of parameters in $(\boldsymbol{\sigma},\boldsymbol{\rho})$ to be a sample of independent beta distributions. In the present context, we consider

separate population distributions for feature activation probabilities that are associated with a specific feature and a specific type of face. In this way, feature activation probabilities for FEs of different stimulus people 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 that are associated with FEs of different stimulus people. Furthermore, the feature emotion probabilities are considered as a sample of a single population distribution. The resulting prior distribution is

$$p(\boldsymbol{\sigma}, \boldsymbol{\rho}|\boldsymbol{\phi}) = \prod_t \prod_s \prod_f \text{beta}(\sigma_{tsf}|\alpha_\sigma^{tf}, \beta_\sigma^{tf}) \prod_e \prod_f \text{beta}(\rho_{ef}|\alpha_\rho, \beta_\rho), \tag{5}$$

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

$$\begin{aligned} p(\boldsymbol{\sigma}, \boldsymbol{\rho}|\mathbf{d}, \mathbf{x}, \mathbf{y}, \boldsymbol{\phi}) = &\prod_t \prod_s \prod_f \text{beta}\left\{\sigma_{tsf}|\alpha_\sigma^{tf} + \sum_e \sum_i x_{ei}^{tsf}, \beta_\sigma^{tf} + \sum_e \sum_i (1 - x_{ei}^{tsf})\right\} \\ &\times \prod_e \prod_f \text{beta}\left\{\rho_{ef}|\alpha_\rho + \sum_t \sum_s \sum_i y_{tsi}^{ef}, \beta_\rho + \sum_t \sum_s \sum_i (1 - y_{tsi}^{ef})\right\}. \end{aligned}$$

3.2.4. Hyperprior distribution

For each of the beta priors in equation (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 $u = \alpha/(\alpha + \beta)$ and the ratio $v = 1/(\alpha + \beta)$. For the feature activation and feature emotion probabilities, we denote the transformed hyperparameters as $(u_\sigma^{tf}, v_\sigma^{tf})$ and (u_ρ, v_ρ) respectively. As we have only a sample of five parameters to estimate the hyperparameters $(u_\sigma^{tf}, v_\sigma^{tf})$ that are associated with FE t and feature f, we have only little information to obtain a reliable estimate of the parameter v_{tf}, which may be regarded as a measure of the variability of the distribution. Therefore we impose the restriction that $v_\sigma^{tf} = v_\sigma$ $(t = 1, \ldots, 4; f = 1, \ldots, F)$.

The specification of a uniform prior for u and v is standard practice (for a similar example, see Gelman *et al.* (2003), page 128) with the understanding that, if the posterior distribution is insufficiently informative, we 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 $\boldsymbol{\phi}^*$ to denote the entire collection of transformed hyperparameters, the joint distribution $p(\boldsymbol{\sigma}, \boldsymbol{\rho}, \boldsymbol{\phi}^*)$ can be expressed as

$$\begin{aligned} p(\boldsymbol{\sigma}, \boldsymbol{\rho}, \boldsymbol{\phi}^*) = &\prod_t \prod_s \prod_f \text{beta}\left(\sigma_{tsf}\middle|\frac{u_\sigma^{tf}}{v_\sigma}, \frac{1 - u_\sigma^{tf}}{v_\sigma}\right) \prod_e \prod_f \text{beta}\left(\rho_{ef}\middle|\frac{u_\rho}{v_\rho}, \frac{1 - u_\rho}{v_\rho}\right) \\ &\times \prod_t \prod_f U(u_\sigma^{tf}|0, 1)\, U(v_\sigma|0, 1)\, U(u_\rho|0, 1)\, U(v_\rho|0, 1). \end{aligned}$$

Samples from the posterior distribution $p(\boldsymbol{\sigma}, \boldsymbol{\rho}, \boldsymbol{\phi}^*|\mathbf{d})$ can be drawn by using the Gibbs sampler, drawing directly from the conjugate full conditional posterior distributions for the latent parameters $\mathbf{x}_{ei}^{ts}, \mathbf{y}_{tsi}^{e}, \sigma_{tsf}$ and ρ_{ef} and using the Metropolis algorithm to update the hyperparameters $u_\sigma^{tf}, v_\sigma, u_\rho$ and v_ρ in turn. Convergence can be monitored by using the multiple-sequence

diagnostic of Gelman and Rubin (1992), and then the posterior sample can be used to derive point estimates and $100(1-\alpha)\%$ 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 to 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 concern only the relative fit of models it is recommended to evaluate whether the model selected captures important aspects of the data and whether it fits the data in a global way. The sample of the posterior may be a basis for model evaluation via the use of posterior predictive checks (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 one, two, three or four features. For these models the DIC values equal 8452, 6222, 5842 and 5779 respectively, and Bayesian p-values of the Pearson χ^2 discrepancy measure (see Appendix A) equal 0, 0.004, 0.094 and 0.513 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.

(a) Are there important differences between feature activation probabilities associated with FEs of the same type from different stimulus people?
(b) Which features are relevant when processing a particular type of FE?
(c) Do the data provide evidence for configural encoding?
(d) Does the model respect the multivariate nature of the data, i.e. 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 σ (which are 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 people (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 people.

An inspection of the estimated hyperparameters u_σ^{tf} in Fig. 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 labelled 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

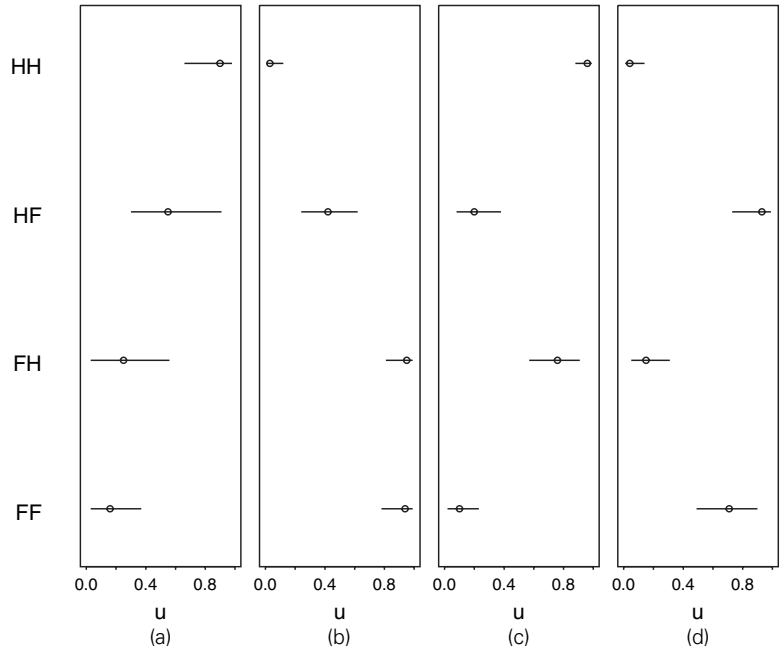

Fig. 3. Posterior median (∘) and 95% posterior interval (———) of hyperparameters u_{σ}^{tf} of feature activation probabilities for a four-feature model: (a) feature 1 (HU); (b) feature 2 (FU); (c) feature 3 (HL); (d) feature 4 (FL)

and lower parts. As an exception, the FU feature has a moderate probability of being activated in the happy–fearful FE. However, this result is mainly caused by moderate activation probabilities for the FEs of two stimulus people (average $\sigma_{\mathrm{HF,FU}}$ of 0.55) whereas the activation probabilities for the other three stimulus people are lower (average $\sigma_{\mathrm{HF,FU}}$ of 0.30).

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 PFMs with fewer 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 labelled FU. Finally, the HU feature contributes least to the model fit.

The PFM is especially suited to investigating the hypothesis of configural encoding as for modelling 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 an 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 Fig. 4, most positive emotions, however, require only the activation of the happy lower feature to be perceived. Exceptions are the emotions ‘affection’ and ‘relaxed’ that also show a moderately strong link with the happy upper feature. In contrast, most negative emotions require the activation of two features 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

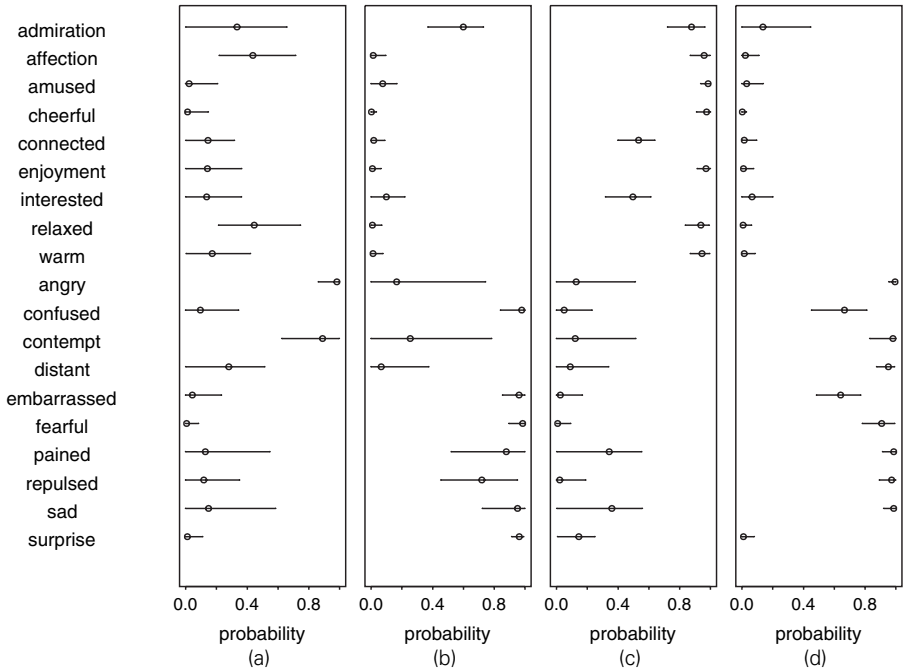

Fig. 4. Posterior median (∘) and 95% posterior interval (———) of fear emotion probabilities for a four-feature model: (a) feature 1 (HU); (b) feature 2 (FU); (c) feature 3 (HL); (d) feature 4 (FL)

such as confused, embarrassed and repulsed show a similar 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 of being perceived in the chimerical happy–fearful FE. Finally, the neutral emotion surprise is linked only 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 analysis-of-variance model for a completely randomized factorial design (Kirk (1995), page 441) with four factors as test statistics. More specifically, this analysis-of-variance model partitions the total sum of squares

$$\mathrm{SS}_{\mathrm{TOT}} = \sum_u \sum_l \sum_s \sum_e (d_{+ulse} - \bar{d}_{+\ldots})^2$$

into main effects, second-order interactions and so on (Table 1). As the total sum of squares can vary across replicated data sets, we use the proportion of the variation that is 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.

Table 1. Observed proportion of the variation accounted for by components of the analysis-of-variance model and posterior predictive *p*-value for a four-feature model

Component	*Observed proportion*	*p-value*
SS_U	0.0008	0.36
SS_L	0.0172	0.57
SS_S	0.0040	0.53
SS_E	0.2047	0.67
SS_{UL}	0.0001	0.75
SS_{US}	0.0010	0.81
SS_{UE}	0.1446	0.29
SS_{LS}	0.0043	0.68
SS_{LE}	0.4782	0.41
SS_{SE}	0.0336	0.06
SS_{ULS}	0.0002	0.98
SS_{ULE}	0.0238	0.64
SS_{USE}	0.0415	0.43
SS_{LSE}	0.0357	0.24
SS_{ULSE}	0.0103	1.00

6. Discussion

In this paper we analysed experimental data on the perception of emotion from de Bonis *et al.* (1999). Our analysis goes beyond the analysis of de Bonis *et al.* (1999) 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.* (1999) 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 (which is 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 and 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.* (1999). More specifically, it was argued in this paper that a PFM has several advantages for testing the configurality hypothesis compared with a generalized linear modelling approach. First, the PFM allows us to identify empirically the relevant features of emotion perception from the data whereas a generalized linear model would take the features that are manipulated in the experiment for granted. Note that the performance of PFMs in extracting 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), page 71). In the present application, the PFM identified the four manipulated facial halves to be the relevant features for the perception of emotion and, as such, it provided a check on the manipulations that are involved in the experiment. Second, the PFM is especially suited to investigating the configural encoding hypothesis as it includes a (conjunctive) mechanism for modelling 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 half 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 de Bonis *et al.* (1999), the PFM shows that basic emotions other than happiness or fear

can be elicited by chimerical FEs; the analysis provides a clear picture of this 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 generalized linear models is not always necessary since PFMs can in some cases be expressed as generalized linear models (Meulders, De Boeck and Van Mechelen, 2001). For instance, a PFM in which either feature activation probabilities or feature emotion probabilities are considered to be fixed would be equivalent to a generalized linear model with a binomial random component, a log-link and a linear predictor that depends on the type of face and the emotion.

Acknowledgements

We thank the Associate Editor and several reviewers for helpful comments. The research that is 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 the Katholieke Universiteit Leuven (F/96/6 Fellowship to Andrew Gelman, OT/96/10 project awarded to Iven Van Mechelen, grant GOA/2000/02 awarded to Paul De Boeck and Iven Van Mechelen, and grant PDM/02/066 awarded to Paul De Boeck).

Appendix A: Computation of the deviance information criterion

Using $\boldsymbol{\theta}$ as notation for parameters that are involved in the likelihood of the model, the DIC is defined as

$$\mathrm{DIC} = \overline{D(\boldsymbol{\theta})} + p_D,$$

with $\overline{D(\boldsymbol{\theta})}$ 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 $p_D = \overline{D(\boldsymbol{\theta})} - D(\bar{\boldsymbol{\theta}})$, with $\bar{\boldsymbol{\theta}}$ being the mean of the posterior sample. For the PFM, the deviance function is specified as $-2\log\{p(\mathbf{d}|\boldsymbol{\sigma}, \boldsymbol{\rho})\}$.

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 (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 $T(\mathbf{d})$ that depend only on the data $\mathbf{d}$ and discrepancy measures $T(\mathbf{d}, \boldsymbol{\theta})$ that depend on both the parameters $\boldsymbol{\theta}$ and the data. An example of the latter type of measure is the Pearson χ^2-measure for evaluating global goodness of fit. In the context of the present application, the Pearson χ^2-measure is defined as

$$\chi^2(\mathbf{d}, \boldsymbol{\theta}) = \sum_{t,s,e} \frac{\{d_{+tse} - E(D_{+tse}|\boldsymbol{\theta})\}^2}{\mathrm{var}(D_{+tse}|\boldsymbol{\theta})},$$

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

References

de Bonis, M., De Boeck, P., Pérez-Diaz, F. and Nahas, M. (1999) A two-process theory of facial perception of emotions. *C. R. Acad. Sci.* III, **322**, 669–675.

Candel, M. J. J. M. and Maris, E. (1997) Perceptual analysis of two-way two-mode frequency data: probability matrix decomposition and two alternatives. *Int. J. Res. Marktng*, **14**, 321–339.

Dempster, A. P., Laird, N. M. and Rubin, D. B. (1977) Maximum likelihood from incomplete data via the EM algorithm (with discussion). *J. R. Statist. Soc.* B, **39**, 1–38.

Ekman, P. and Friesen, W. V. (1976) *Pictures of Facial Affect*. Palo Alto: Consulting Psychologists Press.
Gelman, A., Carlin, J. B., Stern, H. S. and Rubin, D. B. (2003) *Bayesian Data Analysis*, 2nd edn. London: Chapman and Hall.
Gelman, A., Meng, X. M. and Stern, H. (1996) Posterior predictive assessment of model fitness via realized discrepancies. *Statist. Sin.*, **4**, 733–807.
Gelman, A. and Rubin, D. B. (1992) Inference from iterative simulation using multiple sequences. *Statist. Sci.*, **7**, 457–472.
Gelman, A., Van Mechelen, I., Verbeke, G., Heitjan, D. and Meulders, M. (2004) Multiple imputation for model checking: completed-data plots with missing and latent data. *Biometrics*, to be published.
Kirk, R. E. (1995) *Experimental Design: Procedures for the Behavioral Sciences*, 3rd edn. New York: Brooks–Cole.
Maris, E., De Boeck, P. and Van Mechelen, I. (1996) Probability matrix decomposition models. *Psychometrika*, **61**, 7–29.
Massaro, D. W. (1998) *Perceiving Talking Faces: from Speech Perception to a Behavioral Principle*. London: MIT Press.
Meng, X. L. (1994) Posterior predictive *p*-values. *Ann. Statist.*, **22**, 1142–1160.
Meulders, M., De Boeck, P., Kuppens, P. and Van Mechelen, I. (2002) Constrained latent class analysis of three-way three-mode data. *J. Class.*, **19**, 277–302.
Meulders, M., De Boeck, P. and Van Mechelen, I. (2001) Probability matrix decomposition models and main-effects generalized linear models for the analysis of replicated binary associations. *Comput. Statist. Data Anal.*, **38**, 217–233.
Meulders, M., De Boeck, P. and Van Mechelen, I. (2003) A taxonomy of latent structure assumptions for probability matrix decomposition models. *Psychometrika*, **68**, 61–77.
Meulders, M., De Boeck, P., Van Mechelen, I., Gelman, A. and Maris, E. (2001) Bayesian inference with probability matrix decomposition models. *J. Educ. Behav. Statist.*, **26**, 153–179.
Morris, J. S., de Bonis, M. and Dolan, R. J. (2002) Human amygdala responses to fearful eyes. *NeuroImage*, **17**, 214–222.
Rosenberg, S. and De Boeck, P. (1997) Emotion depicted and experienced: Picasso's portraiture. In *Emotion, Creativity and Arts* (eds L. Dorfman, C. Martindale, D. G. Leontiev, V. Cupchick, P. Petrov and N. Matchotka), pp. 371–386. Perm: Perm Institute of Art and Culture.
Rubin, D. B. (1984) Bayesianly justifiable and relevant frequency calculations for the applied statistician. *Ann. Statist.*, **12**, 1151–1172.
Russell, J. A. (2003) Core affect and the psychological construction of emotion. *Psychol. Rev.*, **110**, 145–172.
Spiegelhalter, D. J., Best, N. G., Carlin, B. P. and van der Linde, A. (2002) Bayesian measures of model complexity and fit (with discussion). *J. R. Statist. Soc.* B, **64**, 583–639.
Tanner, M. A. and Wong, W. H. (1987) The calculation of posterior distributions by data augmentation. *J. Am. Statist. Ass.*, **82**, 528–540.