FADA algorithm

The core of the FADA algorithm results in an EM-like optimization of the parameters of the following general anechoic mixing model:

\[ W^*_i(t) = \sum_{j=1}^{P} a_{ij} s_j(t - \tau_{ij}) \]  

(1)

where each temporal signal \( W^*_i(t) \) is an AU temporal profile. Note that for the special case that \( \tau_{ij} = 0 \) for all pairs \((i,j)\), this eqn.(1) coincides with the classical instantaneous mixing model underlying NMF, except for the positivity constraints. As the anechoic algorithm used by Omlor and Giese (Omlor & Giese, 2007a, 2007b), FADA is based on eqn.(1) but it includes additional smoothness priors for the source functions. The introduction of such priors is justified by the observations that biological data usually have limited band-width and by the fact that such priors improve substantially the robustness of the estimation method. Band limited signals in (1) can approximated by a truncated Fourier series of the form:

\[ W^*_i(t) \approx \sum_{k=-Q}^{Q} c_{ik} \exp^{ikt} \]  

(2)

and

\[ s_j(t - \tau_{ij}) \approx \sum_{k=-Q}^{Q} \nu_{jk} \exp^{-ikt} \exp^{i\phi_{jk}} \]  

(3)

\( Q \) is a positive integer which is determined by Nyquist’s theorem according to the limit frequency of the signals. The symbol \( \exp \) signifies the imaginary unit and \( c_{ik} \) and \( \nu_{jk} \) complex numbers \((c_{ik} = |c_{ik}| e^{i\phi_{cik}} \) and \( \nu_{jk} = |\nu_{jk}| e^{i\phi_{jk}} \)). Substituting (2) and (3) in (1), and assuming uncorrelatedness of the sources \( s_j(t) \), the following iterative algorithm for the identification of the unknown parameters in eqn.(1) was applied:

After a random initialization of the estimated parameters, the following steps are carried out until convergence:

1. Compute the absolute values of the coefficients \( c_{ik} \) and solve the following positive demixing problem using positive ICA or non-negative matrix factorization:

\[ |c_{ik}|^2 = \sum_{j=1}^{P} |a_{ij}|^2 |\nu_{jk}|^2 \]  

(4)

with \( i = 0, 1, \ldots N_s \) and \( k = 0, 1, \ldots Q \). \( P \) indicates the total number of sources. Since the signals are real the Fourier coefficients in equations eqn.(2) and eqn.(3) for positive and negative indices \( k \) are complex conjugates of each other. For this reason it is sufficient to solve the demixing problem by considering only the coefficients with indices \( k \geq 0 \). For our implementation here we used non-negative independent component analysis (Hørjen-Sørensen, Winther, & Hansen, 2002) for the solution of the underlying demixing problem with non-negative components.

2. Initialize \( \phi_{\nu_{jk}} = 0 \) for all pairs \((j,k)\) and iterate the following steps:

   (a) Update the phases of the Fourier coefficients of the sources, which are defined by the identity \( \phi_{\nu_{jk}} = \text{angle}(\nu_{jk}) = \arctan(\text{Im}(\nu_{jk})/(\text{Re}(\nu_{jk})) \) by solving the following non-linear least square problem

\[ \min_{\Phi} \| C - V(\Phi) \|^2_F \]  

(5)

where \( (C)_{ik} = c_{ik} \), \( (V)_{ik} = \sum_{j=1}^{N} a_{ij} e^{-ikt} \nu_{jk} e^{i\phi_{\nu_{jk}}} \) and \( \Phi_{jk} = \phi_{\nu_{jk}} \). \( \| . \|^2_F \) indicates the Frobenius norm.
(b) Assuming that the source functions $s_j(t)$, defined by the parameters $\nu_{jk}$ are known, the mixing weights $a_{ij}$ and the delays $\tau_{ij}$ can be optimized for each signal $x_i$ by the following minimization:

$$[\hat{a}, \hat{t}] = \arg \min_{a, t} \| W_i^* (t) - s(t, t)' a \|^2_F$$  \hspace{1cm} (6)

Optimization with respect to $a$ and $t$ is feasible, assuming uncorrelatedness of the sources and independence of the time delays (Swindlehurst, 1998). The column vector $a$ concatenates all weights associated with dof $i$, i.e. $a = [a_{i1}, ..., a_{iP}]'$. The vector function $s(t, t) = [s_1(t - \tau_{i1}), ..., s_N(t - \tau_{iP})]'$ concatenates source functions associated with dof $i$, shifted by the associated time delays.

By construction, the values of the AU activation profiles are non-negative. The FADA algorithm does not include a non-negativity constraint for the identified source functions. To overcome this discrepancy, the FADA algorithm was not applied directly to the matrix $X$ but to the matrix $dX/dt$, the rows of which are the derivatives of the AU profiles in the matrix $X$. Once the component were identified, the original data were approximated by integrating the identified components over time. The constant values to add to the components after their integration were identified using an optimization procedure minimizing the error between actual and reconstructed data. An additional constraint was also imposed on the constant pentameters in order to assure the non-negativity after summation of the corresponding integrated source functions.

**Model type and order selection criteria (LAP, AIC and BIC)**

To determine whether human facial emotion perception is (approximately) Bayes-optimal, we compared human perception results from psychophysical experiments to Bayesian model selection. In the following, we briefly describe the applied approximate Bayesian model type and order (i.e. number of sources) selection criterion. This criterion is based on a Laplace approximation to the model evidence (Laplace, 1774; Bishop, 2007). Details of the derivation can be found in (Endres, Chiovetto, & Giese, 2013).

Let $X$ be the observable data such that $X_{jt}$ is signal $j \in \{1, ..., J\}$ at (discrete) time $t \in \{1, ..., T\}$. In our setting, the rows of the matrix $X$ are the weights $W^* (t)$ sampled at $T$ points in time. $\Theta_M$ is a vector of model parameters for a model indexed by $M$. In our setting $M$ is a tuple (model type, model order), where the model type is either a smooth anechoic mixture determined with the FADA algorithm (Chiovetto & Giese, 2013) or a synchronous (i.e. undelayed) mixture computed with NMF. Using standard terminology, we denote

$$\text{likelihood} : p(X | \Theta_M, M)$$
$$\text{prior} : p(\Theta_M | M).$$  \hspace{1cm} (7) \hspace{1cm} (8)

The likelihood is the probability density of the data given the model parameters and $M$, we omit references to hyperparameters for simplicity. The parameter prior is the probability density of the model parameters. Thus the marginal likelihood of $M$, or model evidence for $M$ is given by

$$p(X | M) = \int p(X, \Theta_M | \Phi, M) d\Theta_M = \int p(X | \Theta_M, M) p(\Theta_M | M) d\Theta_M$$  \hspace{1cm} (9)

where the second equality follows from the product rule for probability distributions. $\Phi$ indicates a tuple of hyperparameters. Once we have evaluated eqn.(9) for all $M$, we can select that $M$ which maximizes the model evidence, since we have no a-priori preference for any $M$. The hard part, as in most Bayesian models, is the computation of the integral over the model parameters $\Theta_M$ in eqn.(9). Instead of an exact solution, we therefore resort to a Laplace approximation, which is obtained by a quadratic expansion of the logarithm of the
integrand in eqn.(9) around its maximum located at $\Theta^*_M$ and dropping all terms of higher order. Thus, let

$$\Theta^*_M = \argmax_{\Theta_M} [\log(p(X|\Theta_M, M)) + \log(p(\Theta_M|M))]$$

$$H_{uv} = -\frac{\partial^2 \log(p(X|\Theta_M, M))}{\partial \Theta_{M,u}\partial \Theta_{M,v}} |_{\Theta^*_M}$$

$$-\frac{\partial^2 \log(p(\Theta_M|M))}{\partial \Theta_{M,u}\partial \Theta_{M,v}} |_{\Theta^*_M}$$

where $H$ is the Hessian matrix of second derivatives at $\Theta^*_M$. The integral in eqn.(9) then becomes multivariate Gaussian, which is tractable in closed form (Endres et al., 2013):

$$p(X|M) \approx p(X|\Theta^*_M, M) p(\Theta^*_M|M) \frac{(2\pi)^{F/2}}{\sqrt{|H|}}$$

where $F = \text{dim}(\Theta)$. This approximation typically yields good results when the integrand has one maximum that can be well fitted by a Gaussian. Our LAP criterion is the logarithm of this expression:

$$\text{LAP} = \log(p(X|\Theta^*_M, M)) + \log(p(\Theta^*_M|M)) + \frac{\text{dim}(\Theta)}{2} \log(2\pi) - \frac{1}{2} \log(|H|)$$

The three parts indicated by the braces have the following interpretation: goodness-of-fit is measured by the log-likelihood, the log-prior is a regularization term to insure well-defined behavior of the parameters when the observable data provide insufficient constraints on $\Theta_M$. The log-posterior volume of the parameter posterior measures how well the data constrain the $\Theta_M$: if this contribution is large, then the posterior uncertainty is high and consequently the LAP score increases. In other words, complex models will only be punished by LAP if there is sufficient data to justify that.

We compared the LAP criterion to two standard model complexity estimators below (see the Results section): the Bayesian information criterion (BIC) (Schwarz, 1978) and the Akaike information criterion (AIC) (Akaike, 1974). The BIC is given by

$$\text{BIC} = -2 \left( \log(p(X|\Theta^*, \Phi, M)) - \frac{1}{2} \text{dim}(\Theta) \log(N) \right)$$

where $N$ is the number of data-points. The best model is found by minimizing the BIC w.r.t. $M$. The BIC can be obtained from the LAP criterion in the limit $N \to \infty$, by dropping all terms from the LAP which do not grow with $N$ and multiplying by $-2$ (Bishop, 2007; Endres et al., 2013).

Information-theoretic reasoning lead to the AIC (Akaike, 1974): a good model should lose only a small amount of information when approximating (unknown) reality. Measuring information by Kullback-Leibler divergence (Cover & Thomas, 1991), the AIC is obtained:

$$\text{AIC} = -2 \left( \log(p(X|\Theta^*, \Phi, M)) - \text{dim}(\Theta) \right).$$

Like for the BIC, a good model has a low AIC score.

To apply any of these model selection criteria, we use the following likelihood model, which approximates the generative models of the FADA as well as the NMF algorithm:
Figure 1: Graphical model of the ideal observer which we employ to test cue fusion between movement primitive percepts. Like in (Ernst & Banks, 2002), we assume that a latent cause $R$, here the strength of an emotion, causes the observable cues $C_1, C_2$, here scaled FADA movement primitives. Standard cue fusion models assume that the observable effects become independent given the cause. We allow for an additional dependency between $C_1$ and $C_2$ given $R$, indicated by the horizontal connection. For details, see text.

\[
X_{jt} = \sum_i W_{ji} S_i(t - \tau_{ji}) + \eta_{jt}
\]

\[
\eta_{jt} \sim \mathcal{N}(0, \sigma_n)
\]

where $X_{jt}$ is signal $j$ at time $t$, $W_{ji}$ is the weight of source $i \in \{1, \ldots, I\}$ for signal $j$, $\tau_{ji}$ is the delay of source $i$ in signal $j$. $\eta_{jt}$ is a Gaussian noise term with standard deviation $\sigma_n$. For the LAP, we additionally need to choose priors on the weights, sources and delays:

\[
W_{ji} \sim \Gamma(\nu, S)
\]

\[
\tau_{ji} \sim \text{Exponential}()\]

\[
S_i(t) \sim \mathcal{GP}(\mu(t), k(t, t'))
\]

i.e. the weights $W_{ji}$ are positive and drawn from a gamma distribution (Bishop, 2007) with shape $\nu$ and scale $S$, the delays $\tau_{ij}$ are drawn from an exponential distribution with mean $\lambda$ and the source functions have a Gaussian process prior (Rasmussen & Williams, 2006) with mean function $\mu(t)$ and kernel $k(t, t')$. We estimate the mean function from the sources obtained with the FADA algorithm, and use a wave kernel (Genton, 2001)

\[
k(t, t') \propto \text{sinc}(2f_0|t - t'|) = \frac{\sin(2\pi f_0|t - t'|)}{2\pi f_0|t - t'|}
\]

because it imposes a soft temporal regularization onto the sources that resembles the ideal low-pass filter employed in the FADA algorithm. We are not imposing a strict positivity prior on the sources. However, the mean function of the Gaussian process prior on the sources (eqn.(19)) is positive due to FADA’s positivity constraints. This ensures that most of the probability mass of the prior is located where $S_i(t) > 0$, thereby promoting positive sources in the posterior. We showed earlier (Chiovetto, d’Avella, & Giese, 2016) that this soft positivity prior leads to negligible approximation errors.

**Cue fusion models**

The FADA algorithm can discover modular movement primitives, i.e. simple component movements that are linearly combined to form complex natural movements, see eqn.(1). Since we are interested in the perception of facial emotion expressions, we wondered in which form the individual primitives contribute to the perception of emotions, and how the contributions of different primitives interact. To this end, we investigated the hypothesis that the primitives can be viewed as “cues” that are fused by the observer to generate a unified percept. In the domain of multimodal perception, it has been demonstrated that human performance can come close to a Bayesian ideal observer in haptic-visual or audio-visual cue integration tasks (Ernst & Banks, 2002; Körding et al., 2007), or proprioceptive-visual integration informed by an internal forward model (Beck, Endres, Lindner, & Giese, 2014). The graphical generative model for such an
ideal observer in our setting is shown in fig. (1): an (unobservable) generating process gives rise to an emotion with strength \( R \) which causes two (or more) cues \( C_1 \) and \( C_2 \). The cues are amplitude-scaled versions of the movement primitives, which can be achieved by multiplying the emotion-specific weights \( W_{j,i} \) for primitive \( i \) with a constant (positive) factor. Causal dependencies are depicted by arrows, the undirected connection between the cues models any dependency between the primitives that is not due to the common emotion. For simplicity, we model the joint distribution of \( R, C_1 \) and \( C_2 \) as a multivariate Gaussian \( \mathcal{N}(\mu, \Sigma) \), ignoring that they are range-limited. This approximation is inspired by the earlier successes of similar models (Ernst & Banks, 2002). Let \( c = (C_1, C_2) \), \( z = (R, c)^T = (R, C_1, C_2)^T \), then

\[
\begin{align*}
\mu &= (\mu_R, \mu_C)^T = (\mu_R, \mu_{C_1}, \mu_{C_2})^T \\
\Sigma &= \begin{pmatrix}
\Sigma_R \\
\Sigma_{rc} \\
\Sigma_c
\end{pmatrix}
\end{align*}
\]

Under our experimental conditions, the observer was not asked to report a posterior distribution, but rather a single value that summarizes e.g. (Petersen & Pedersen, 2012), chapter 8, formula 353. In the following we write this posterior’s mean as a function of the cues

\[
\text{The expressions for the posterior mean and variance of } R \text{ can be deduced from formulas for conditioning multivariate Gaussians, see e.g. (Petersen & Pedersen, 2012), chapter 8, formula 353. In the following we write this posterior’s mean as a function of the cues}
\]

\[
R(C_1, C_2) := E \left[ R | C_1, C_2 \right]
\]

Under our experimental conditions, the observer was not asked to report a posterior distribution, but rather a single value that summarizes it. In the context of Bayesian estimation theory, choosing this value is usually done by minimizing a risk function given the posterior. If the risk is quadratic, then the best guess is the posterior expectation of \( R \). Denoting the entries of the inverse covariance matrix by

\[
\Sigma_c^{-1} = \begin{pmatrix}
p_{C_1} & p_{C_1, C_2} \\
p_{C_1, C_2} & p_{C_2}
\end{pmatrix}
\]

we find, by rearranging terms in the posterior mean in eqn.(25)

\[
\begin{align*}
R(C_1, C_2) &= \mu_R + \sigma^2_{R,C_1} (p_{C_1}(C_1 - \mu_{C_1}) + p_{C_1, C_2}(C_2 - \mu_{C_2})) \\
&\quad + \sigma^2_{R,C_2} (p_{C_1, C_2}(C_1 - \mu_{C_1}) + p_{C_2}(C_2 - \mu_{C_2})) \\
&= \mu_R - (\sigma^2_{R,C_1} p_{C_1} + \sigma^2_{R,C_2} p_{C_1, C_2}) \mu_{C_1} - (\sigma^2_{R,C_2} p_{C_2} + \sigma^2_{R,C_1} p_{C_1, C_2}) \mu_{C_2} \\
&\quad + (\sigma^2_{R,C_1} p_{C_1} + \sigma^2_{R,C_2} p_{C_1, C_2}) C_1 + (\sigma^2_{R,C_2} p_{C_1, C_2} + \sigma^2_{R,C_2} p_{C_2}) C_2 \\
&=: B_0 + B_1 C_1 + B_2 C_2
\end{align*}
\]
In other words, the expected emotional strength is a linear function of the cues, where the coefficients $B_i$ follow from the parameters of the multivariate Gaussian distribution. Note that this result holds even if there is a conditional dependence between the cues (graphically expressed by the undirected connection between $C_1$ and $C_2$ in fig. (1)).

In a typical cue fusion experiment, one would now proceed to measure the parameters of the distribution in eqn.(22) by occluding or masking one cue and measuring the marginal distribution of the other. Afterwards, one would test if the cues as fused as predicted, i.e. if eqn.(28) holds when both cues are presented. However, we can not do that, because the FADA movement primitives are spatially holistic: one primitive generates movement in the whole face. Thus, we instead test the prediction that $\tilde{R}$ with both cue strengths different from zero can be predicted from the participant-reported $R$s when one cue strength is zero. Eqn.(28) implies that

$$R(C_1, 0) = B_0 + B_1 C_1 \Rightarrow C_1 = \frac{R(C_1, 0) - B_0}{B_1} \quad (31)$$
$$R(0, C_2) = B_0 + B_2 C_2 \Rightarrow C_2 = \frac{R(0, C_2) - B_0}{B_2} \quad (32)$$

and therefore

$$R(C_1, C_2) = B_0 + B_1 \frac{R(C_1, 0) - B_0}{B_1} + B_2 \frac{R(0, C_2) - B_0}{B_2} = R(C_1, 0) + R(0, C_2) - B_0 \quad (33)$$

whence the expected emotional strength rating for $C_1, C_2 > 0$ can be computed by summing up the strength ratings measured when one of the cues is zero, minus a bias term. Note that of all models that predict $R(C_1, C_2)$ linearly from $R(C_1, 0)$ and $R(0, C_2)$:

$$R(C_1, C_2) = \beta_1 R(C_1, 0) + \beta_2 R(0, C_2) + \beta_0 \quad (34)$$

eqn.(33) is the only self-consistent one in the sense that

$$R(C_1 \neq 0, C_2 = 0) = R(C_1, 0) + R(0, 0) - B_0 = R(C_1, 0) + B_0 - B_0 = R(C_1, 0) \quad (35)$$

and likewise for $C_1 = 0, C_2 \neq 0$. In other words, the coefficients $\beta_1$ and $\beta_2$ in eqn.(34) must have coefficients equal to one, and $\beta_0 = B_0$.

Even a cursory inspection of the figure associated with the rating data in the article reveals non-linear saturation effects for high morphing levels ($(\gamma_1, \gamma_2)$ in the figure): while there is an almost linear increase of the rating with $\gamma_2$ if $\gamma_1 = 0$, the rating is almost constant w.r.t. $\gamma_2$ if $\gamma_1 = 1$. We therefore also tested a cue fusion model with an output non-linearity of the form

$$\tilde{R}(C_1, C_2) = f \left( R(C_1, C_2) \right) := z_0 \left( 1 - \exp \left( -z_1 R(C_1, C_2) \right) \right) \quad (36)$$

where $\tilde{R}(C_1, C_2)$ is the rating reported by the participants, and $R(C_1, C_2)$ is rating predicted by the linear Bayesian cue fusion model, eqn.(33). This non-linearity has the property of being almost linear for small $z_1$ and $z_0 = \frac{1}{z_1}$. Furthermore, it saturates at $z_0$ for large $x$. Combining eqn.(36) with eqn.(33), we obtain the predicted rating with output non-linearity:

$$\tilde{R}(C_1, C_2) = f \left( f^{-1} \left( R(C_1, 0) \right) + f^{-1} \left( R(0, C_2) \right) - B_0 \right) \quad (37)$$

Note that all three cue fusion predictions (eqn. 33, 34, and 37) are relationships between means, not individual trial outcomes. For the nonlinear model, this implies that averages are computed after $f^{-1}()$ has been applied to the individual ratings.
References


