Copula Ordinal Regression
for Joint Estimation of Facial Action Unit Intensity

Robert W alecki*, Ognjen Rudovic*, Vladimir Pavlovic† and Maja Pantic‡
*Department of Computing, Imperial College London, UK
† Department of Computer Science, Rutgers University, USA
‡ EEMCS, University of Twente, The Netherlands
{r.walecki14,o.rudovic,m.pantic}@imperial.ac.uk, vladimir@cs.rutgers.edu

Abstract

Joint modeling of the intensity of facial action units (AUs) from face images is challenging due to the large number of AUs (30+) and their intensity levels (6). This is in part due to the lack of suitable models that can efficiently handle such a large number of outputs/classes simultaneously, but also due to the lack of labelled target data. For this reason, majority of the methods proposed so far resort to independent classifiers for the AU intensity. This is sub-optimal for at least two reasons: the facial appearance of some AUs changes depending on the intensity of other AUs, and some AUs co-occur more often than others. Encoding this is expected to improve the estimation of target AU intensities, especially in the case of noisy image features, head-pose variations and imbalanced training data. To this end, we introduce a novel modeling framework, Copula Ordinal Regression (COR), that leverages the power of copula functions and CRFs, to detangle the probabilistic modeling of AU dependencies from the marginal modeling of the AU intensity. Consequently, the COR model achieves the joint learning and inference of intensities of multiple AUs, while being computationally tractable. We show on two challenging datasets of naturalistic facial expressions that the proposed approach consistently outperforms (i) independent modeling of AU intensities, and (ii) the state-of-the-art approach for the target task.

1. Introduction

Human facial expressions are typically described in terms of variation in configuration and intensity of facial muscle actions defined using the Facial Action Coding System (FACS) [6]. Specifically, the FACS defines a unique set of 30+ atomic non-overlapping facial muscle actions named Action Units (AUs) [19]. It also provides rules for scoring the intensity of each AU in the range from absent to maximal intensity on a six-point ordinal scale, denoted as neutral<A<B<C<D<E. Thus, using FACS, human coders can manually code nearly any anatomically possible facial expression, decomposing it into specific AUs and their intensities. However, this process is tedious and error-prone due to the large number of AUs and the difficulty in discerning their intensities [20]. On the other hand, automated estimation of the AU intensity is challenging for many reasons such as the subject-specific facial morphology and expressiveness level [24], as well as the changes in lighting and the head-pose variation. Co-occurrences of the intensity levels of different AUs are another important factor that affects their coding/automated estimation. For instance, the criteria for intensity scoring of AU7 (lid tightener) are changed significantly if AU7 appears with a maximal intensity of AU43 (eye closure), since this combination changes the appearance as well as timing of these AUs [6].
Furthermore, co-occurring AUs can be non-additive, in the case of which one AU masks another, or a new and distinct set of appearances is created [6]. As an example of the non-additive effect, AU4 (brow lowerer) appears differently depending on whether it occurs alone or in combination with AU1 (inner brow raise). When AU4 occurs alone, the brows are drawn together and lowered, while in AU1+4, the brows are drawn together but are raised due to the activation of AU1. This, in turn, significantly affects their appearance. Moreover, some AUs are often activated together, e.g. AU12 and AU6 in the case of smiles, but with different intensities depending on the type of smile (e.g., genuine vs. posed). Therefore, modeling dependencies among (the intensities of) multiple AUs is expected to result in models that are more robust to noisy features and imbalanced training data, leading to a more accurate estimation of the target AU intensities [26, 16].

Contributions. To address the primary challenge of computationally modeling the variable and complex dependencies that exist among intensities of multiple AUs, then leveraging the models for more accurate AU intensity prediction, we propose the Copula Ordinal Regression model for joint AU intensity estimation. Specifically, we propose to use the powerful framework of copula functions [29] to efficiently model dependencies of intensities among AUs. Copula functions generalize the notion of linear correlation to more flexible dependency structures specified using simple parametric functional families (copula families). The key advantage of copula models is that they retain representational and computational efficiency by decoupling the modeling of dependencies from the modeling of marginal densities, as detailed in Sec.2.2. The basic idea is that one starts with state-of-the-art independent (marginal probability) AU models and then captures the intrinsic AU dependence (joint probability) through copula functions, while guaranteeing that the marginals remain unaltered. This presents a distinct advantage over all previously surveyed models that tightly couple the marginal and joint model specification/estimation, resulting in often intractably complex models.

Even though copulas model dependencies using compact parametric functions, it is still necessary to estimate their parameters from data. To this end, we propose a new Conditional Random Field (CRF) model in Sec.2.2 and the accompanying learning and inference strategies in Sec.2.4. The CRF-based model considers sparse, graph-induced, cliques of AUs (inferred from data and illustrated in Fig.1), where dependencies in each clique are modeled using an independent copula model. The joint CRF model is then estimated using a new, efficient block descent algorithm that intuitively combines optimization of dependencies (copula association parameters) with learning of independent marginal model parameters (the intensity levels of each AU from the corresponding covariates, i.e., the locations of a set of fiducial facial points). To avoid the typically challenging evaluation of the CRF partition function, we propose to use a composite marginal likelihood objective with guaranteed optimality properties [30, 5]. The joint inference in this model is then accomplished using a fast loopy belief approximation method on the learned CRF model. We demonstrate the utility of COR on two benchmark datasets of spontaneous AUs, DISFA [22] and Shoulder Pain [18].

2. Methodology

Let us denote the training set as \( D = \{ Y, X \} \). \( Y = [y_1, \ldots, y_i, \ldots, y_N] \) is comprised of \( N \) instances of multivariate outputs stored in \( y_i = \{ y_{i1}, \ldots, y_{iq}, \ldots, y_{iQ} \} \), where \( Q \) is the number of AUs, and \( y_{iq} \) takes one of \( \{1, \ldots, L_q\} \) discrete intensity levels of the \( q \)-th AU. Furthermore, \( X = \)}
[\mathbf{x}_1, \ldots, \mathbf{x}_i, \ldots, \mathbf{x}_N]^T$ are input features (e.g., facial points) that correspond to the combinations of labels in $Y$. Thus, our goal is to simultaneously estimate the combination of the intensity levels $y^q$ of $Q$ AUs, given the facial features $x$. In what follows, we first introduce the ordinal regression framework for modeling single output ($Q = 1$). We then introduce the copula framework for modeling joint distributions, and formulate our model for joint learning and inference of intensity levels of multiple AUs.

2.1. Ordinal Regression

Let $l \in \{1, \ldots, L\}$ be the ordinal label for the intensity level of the $q$-th AU. In the ordinal regression framework notation \[1\], we define the latent projection $y^q \in \mathbb{R}$ as a function of covariates $x$, and then relate this latent projection to the ordinal level $(y^q)$ through the threshold bounds:

$$y^q_l = \beta^q x^T + \epsilon^q, \quad y^q_l = l \text{ iff } \psi^q_{l-1} < y^q_l \leq \psi^q_l,$$

where $x \in \mathbb{R}^D$, $\beta^q$ is the ordinal projection vector, $\psi^q_l$ is the lower bound threshold for count level $l$ ($\psi^q_0 = -\infty < \psi^q_1 < \psi^q_2 < \ldots < \psi^q_{L-1} < \psi^q_L = +\infty$). The error (noise) terms $\epsilon^q$ capture the idiosyncratic effects of all omitted variables for the $q$-th AU. They are assumed to be identically distributed across the intensity levels, each with a univariate continuous marginal distribution function $F(z^q) = \Pr(z^q < z^q)$. In the case of the normal distribution with zero mean and variance $(\sigma^q)^2$, the marginal distribution function is defined as the normal cumulative density function (cdf) $F(z^q) = \Phi(z^q) = \int_{-\infty}^{z^q} N(\xi; 0, 1) d\xi$. Then, classification in ordinal regression models is performed using the following ordinal likelihood \[1\]:

$$l^* = \arg\max_{l=1, \ldots, L} \Pr(y^q = l | y^q_l) = \arg\max_{l=1, \ldots, L} F(z^q_l) - F(z^q_{l-1}),$$

where $z^q_l = (\psi^q_{l-1} - \beta^q x^T)/\sigma^q$ are the cumulative probits. The model parameters are then stored in $\varphi^q = \{\psi^q_1, \psi^q_2, \ldots, \psi^q_{L-1}, \beta^q, \sigma^q\}$.

2.2. Copula Model

A copula is a method for generating a stochastic dependence relationship in the form of a multivariate distribution of random variables with pre-specified marginals \[28\]. Formally, a copula $C(u^1, u^2, \ldots, u^Q): [0, 1]^Q \rightarrow [0, 1]$ is a multivariate distribution function on the unit cube with uniform marginals \[31\]. The main idea of copulas closely related to that of histogram equalization: for a random variable $y^q$ with (continuous) cdf $F$, the random variable $u^q := F(y^q)$ is uniformly distributed on the interval $[0, 1]$. Using this property, the marginals can be separated from the dependency structure in a multivariate distribution \[2\]. This is given by Sklar’s theorem \[29\].

**Theorem 1 (Sklar, 1973)** Given $u^q$ random variables with cdfs $F(y^q)$, $q = 1, \ldots, Q$, and joint cdf $F(y^1, \ldots, y^Q)$, there exist a unique copula $C$ such that for all $u^q$:

$$C(u^1, \ldots, u^Q) = F(F^{-1}(u^1), \ldots, F^{-1}(u^Q))$$

Conversely, given any distribution functions $F_1, \ldots, F_Q$ and copula $C$,

$$F(y^1, \ldots, y^Q) = C(F(y^1), \ldots, F(y^Q)),$$

is a $Q$-variate distribution function on $y^1, \ldots, y^Q$ with marginal distribution functions $F$.

This result allows us to construct a joint distribution by specifying the marginal distributions and the dependency structure separately \[2\]. This offers one the critical flexibility necessary for any multivariate output context: it is possible to simultaneously model complex marginal densities with potentially arbitrary multivariate output dependency structures without the need to specify the two in some complexly intertwined, hard-to-interpret and hard-to-learn model. Note that while the copula representation separates the two contexts (marginal and joint) the two remain tied through Eq. 3.

When the random variables are discrete, as is the case with the AU intensity levels, only a weaker version of Theorem 1 holds: there always exists a copula that satisfies Eq. 4, but it is no longer guaranteed to be unique \[29\]. Nevertheless, we can still construct the joint distribution for discrete variables as:

$$\Pr(y^1 = l^1, \ldots, y^Q = l^Q) =$$

$$\Pr(y^1 = l^1 < \psi^q_{l^1} < \psi^q_{l^2} < \ldots < \psi^q_{l^Q}) =$$

$$\sum_{c_{Q-1} = 0} \sum_{c_{Q-2} = 0} \cdots \sum_{c_{1} = 0} (1) F\left(\psi^q_{l^1-c_1} - c_{Q-1}, \ldots, \psi^q_{l^Q-c_{Q-1}}\right)$$

$$= \sum_{c_{Q-1} = 0} \sum_{c_{Q-2} = 0} \cdots \sum_{c_{1} = 0} (1) C\left(\psi^q_{l^1-c_1}, \ldots, \psi^q_{l^Q-c_{Q-1}}\right)$$

where $u^q_{l^q-c_q} = F(y^q_{l^q-c_q})$, $c_q \in \{0, 1\}$, is defined in Sec.2.1, and $\theta$ are the copula parameters, as defined below. It is important to note two critical aspects here. First, Eq. 5 captures dependency structures among the discrete outputs by correlating their error terms $\epsilon^1, \ldots, \epsilon^Q$ via the copula. Secondly, the joint density model induced by the copula is conditioned on the covariates $x$, i.e., $F(y^1, \ldots, y^Q) \leftarrow F(y^1, \ldots, y^Q \mid x)$. This, in contrast to the models in \[27, 16\] that rely solely on the AU labels, allows the covariates to directly influence the dependence structure of AUs.

Under this formulation, the probability of a particular label combination $y$ is determined by the volume of the axis-parallel hyper-rectangular subregion of $[0, 1]^Q$ induced by vertices $(u^1_{l^1}, \ldots, u^Q_{l^Q})$ and $(u^1_{l^1-1}, \ldots, u^Q_{l^Q-1})$ corresponding to that label combination. For the copula introduced in Eq. 5, this involves evaluation of $2^Q$ cdfs. As an example, for $Q = 2$ the model this reduces to:
\[ 
\Pr(y^1 = l^1, y^2 = l^2) = F(z_{l^1}^1, z_{l^2}^2) + F(z_{l^1-1}^1, z_{l^2-1}^2) - F(z_{l^1-1}^1, z_{l^2}^2) - F(z_{l^1}^1, z_{l^2-1}^2) 
\]

This evaluation becomes computationally expensive and impractical for \( Q > 5 \) due to the number of cdfs \((2^5)^+\) that need be evaluated. In Sec. 2.3, we propose a computationally more astute model, which avoids the exponential explosion induced by arbitrary \( Q \).

One specific benefit of copulas is that they can model different forms of (non-linear) dependency using simple parametric models for \( C(\cdot) \). In this paper, we limit our consideration to the commonly used Frank copula [9] from the class of Archimedean copulas, defined as:

\[
C_\theta(u^1, \ldots, u^Q) = -\frac{1}{\theta} \ln \left( 1 + \frac{Q}{\theta} \sum_{q=1}^{Q} (e^{-\theta u^q} - 1) \right). 
\]

The dependence parameter \( \theta \in (-\infty, +\infty) \setminus \{0\} \), and the perfect positive/negative dependence is obtained if \( \theta \to \pm\infty \). When \( \theta \to 0 \), we recover the ordinal model in Eq.2 (Frank copula becomes the independence copula [9] that is equivalent to the product of ordinal models for each AU). Although various copula functions (e.g., Clayton, Gumbel, etc.) are available for modeling different dependency structures, we choose Frank copula in this paper for two reasons. First, it has a simple closed-form, in contrast to, e.g., the Gaussian copula [2], which, in general, requires the intractable computation of multivariate Gaussian cdfs. Secondly, Frank copula is particularly suitable for the target task as it allows modeling of both positive and negative dependencies, while also capturing dependency in both the left and right tails (i.e., when different AUs are activated either at low intensity, or at high intensity levels together).

### 2.3. Copula Ordinal Regression

As mentioned in Sec.2.2, the joint modeling of multiple AUs using the model in Eq.5 is possible. However, this becomes prohibitively expensive as the number of outputs (i.e., AUs) increases. For instance, for 10 AUs, as commonly coded in face datasets, this would involve \( 2^{10} \) evaluations of the copula function. We mitigate this by approximating the learning of the joint pdf in Eq.5 using the bivariate joint distributions capturing dependencies of AU pairs. To this end, we use the Conditional Random Field (CRF) [15] framework. Formally, we introduce a random field with an associated graph \( G = (V, C) \), where nodes \( v \in V, |V| = Q \), correspond to individual AUs and cliques \( c \in C \) correspond to subsets of dependent AUs modeled using the copula functions. The joint probability distribution of \( Q \) intensity random variables is then defined as:

\[
P(y|x, \Omega) = \frac{1}{Z} \prod_{c \in C} \Psi(y_c|x),
\]

where \( Z \) is the partition function, \( y_c \) is the subset of random variables in clique \( c \), \( \Psi(\cdot) \) is the conditional potential on the labels in this clique, explained below, and \( \Omega = \{ \emptyset, \Omega \} \) are the model parameters.\(^1\)

In this setting specifically, we only consider unary and binary cliques, modeling individual independent AUs and pairs of AUs. In other words, \( C = V \cup E \), where \( E \) is the set of edges in \( G \). Hence,

\[
\Psi(y_c|x) = \begin{cases} 
\Pr(y^r|x), & c = r \in V \\
\Pr(y^r, y^s|x)\gamma, & c = (r, s) \in E 
\end{cases}
\]

where the unary term is the traditional independent AU ordinal regression model defined in Sec. 2.1 and the pairwise term is specified in Eq. 6. Note that the unary terms depend only on the \( \theta_r \) parameters of the ordinal regression model, while the edge potentials depend also on the copula association parameter \( \theta_{rs} \) that models the dependency of \( (r, s) \) pair of outputs. Furthermore, the weight \( \gamma \) is chosen so as to balance the magnitude of the cliques.

While modeling only bivariate distributions may seem a natural way of representing the joint distribution, we model the marginals via the unary potentials for two reasons. First, while the marginals focus on independent classification of target AU intensity, the bivariate copulas focus on encoding the dependence between the intensity levels of two AUs. Thus, by including the copulas in the potential function, a more discriminative classifier for the AU intensity levels is expected. Secondly, in the case when there is no dependence between AUs, in an ideal case \( \theta_{rs} \to 0 \), and Frank copula converges to the independence copula [9]. Yet, due to numerical instability, parameter estimation can be fragile in this case, leading to poor performance of the learned classifier. We control this by having the marginals in the unary potentials.

The most critical aspect in evaluation of the joint distribution in Eq. 8 is computation of the partition function. This is an \( np \)-complete problem, and thus, exact inference in general case is intractable. This is true in our case as it involves the integration over all possible AUs and their intensity levels, i.e., typically \( 6^{10} \) computations. However, approximate methods based on Markov chain Monte Carlo (MCMC) and loopy belief propagation (LBP) for parameter learning have been proposed. Since our joint distribution can be decomposed as a product of (unnormalized) likelihood terms, we resort to a simpler approach - the composite marginal likelihood (CML) [30]. CML decomposes the

\(^1\)For simplicity, we often drop the dependency on \( \Omega \) in notations.
multi-label classification problem into a set of simpler and easier-to-learn subproblems, making the parameter learning extremely efficient for subproblems [32]. By using the notion of CML, our learning objective can be written as:

\[
NCL = - \sum_{i=1}^{N} \left[ \sum_{x \in V} \ln \Pr(y_i^x | x_i) + \gamma \sum_{(r,s) \in E} \ln \Pr(y_i^r, y_i^s | x_i) \right],
\]

thus, avoiding the costly computation of the partition function. Here, \(N\) is the number of training instances. Note that under appropriate regularity conditions, the maximum composite likelihood estimator converges in distribution to true value of the model parameters (see [30] for details).

**Estimation of the AU pairs.** Modeling the fully connected graph (i.e., \(Q \times (Q-1)/2\) bivariate copulas) is impractical as not all AU exhibit a dependence pattern (e.g., AU16 (lower lip depressor) and AU17 (chin raiser) do not co-occur). In CRF and MRF models, the cliques (i.e., the edges) are typically determined from the precision matrix rather than from the correlation matrix \(S\). This is because the precision matrix unravels partial correlations among the variables, while the correlation matrix focuses on marginal correlations [10]. Important advantage of using partial correlations to infer AU dependencies is that, in contrast to marginal correlations, AUs that are correlated through another AU are ignored, therefore, avoiding the redundant modeling. To select the edges in the AU dependency graph, we exploit the partial correlations using a sparse estimate of the precision matrix \(\Upsilon\) computed from \(S\). The aim is to reduce the number of the model parameters by not accounting for the ‘weak’ dependencies among AUs. To this end, we first empirically estimate \(S\). Then, to obtain a sparse representation of \(S\), we employ the graphical lasso estimation [8] to solve the following convex optimization:

\[
(\Upsilon, \tilde{S}) = \min_{\Upsilon > 0} - \ln \det(\Upsilon) + tr(\tilde{S} \Upsilon) + \kappa||\tilde{\Upsilon}||_1, \tag{11}
\]

where \(\kappa\) is the regularization parameter.\(^2\) Finally, the edge set \(E\) is defined by keeping the edges satisfying the condition: \(E = \{(r,s) : |\Upsilon_{r,s}| > \delta\}\). \(\delta = 0.05\) is chosen so that only the pairs of AUs with strong partial correlations are kept, resulting in a model with significantly fewer parameters [23]. The learned graphs are depicted in Fig. 3.

2.4. Learning and Inference

The parameter optimization in the model is performed by minimizing \(NCL\) (Eq. 10) w.r.t. \(\Omega\). For this, we employ the Conjugate gradient method with line search [25].

**Re-parametrization.** The gradient-based learning proposed above has to be accomplished while respecting two sets of constraints: (i) the order constraints on \(\psi\): \(\psi_{j-1} \leq \psi_j \) for \(j = 1, \ldots, L\), and (ii) the positive scale constraint on \(\sigma: \{\sigma > 0\}\). To avoid constrained optimization, we introduce a re-parametrization of \(\psi\) using displacement variables \(\delta_r\), where \(\psi_j = \psi_1 + \sum_{k=1}^{j-1} \delta_r^k \) for \(j = 2, \ldots, L - 1\). The positiveness constraint for \(\sigma\) is simply handled by introducing the free parameter \(\sigma_0\) where \(\sigma = \sigma_0^2\). Thus, the unconstrained parameters of the ordinal marginals are \(\{\beta, \psi_1, \delta_1, \ldots, \delta_{L-2}, \sigma_0\}\), and they are defined separately for each of the \(Q\) ordinal marginals, and stored in \(\varphi\).

**Training.** During training, we seek to find optimal parameters \(\Omega^*\) by solving the regularized optimization problem

\[
\Omega^* = \arg \min_{\Omega = \{\varphi, \theta\}} NCL(\varphi, \theta) + \lambda R_\varphi, \tag{12}
\]

where \(NCL\) is given by Eq.10, \(R_\varphi\) is the standard \(L_2\) regularizer of the projection \(\beta\) and \(\sigma_0\), and \(\lambda\) is the regularization parameter. No specific regularization is necessary for the threshold parameters as they are automatically adjusted according to the score \(\beta x^T\).

Solving for the parameters \(\Omega = \{\varphi, \theta\}\) directly is possible, however, by noticing that the copula parameters \(\theta\) are independent of the node potentials in the \(NCL\), we can alternate between optimization of the marginals \(\varphi\) and the copula association \(\theta\). In this way, we detangle learning of the marginal model parameters from the joint copula parameters. Consequently, we reduce chances of falling into a local minimum due to the large number of parameters to be learned simultaneously. To this end, we propose a block-descent two-step optimization. We briefly describe the learning strategy.

**Algorithm 1 Copula Ordinal Regression Learning**

| Input: Training data \(D = \{(x_i, y_i)\}_{i=1}^{N}\) |
| Output: Model parameters \(\Omega = \{\varphi, \theta\}\) |
| **Initialization:** |
| \(\forall (r,s) \in E \rightarrow \theta_{rs} = \text{sign}(\text{corr}(y_r^r, y_s^s))\) |
| \(\forall r \in V \rightarrow \varphi_r = \arg \min_{\varphi_r^{N}} - \sum_{i=1}^{N} \ln \Pr(y_r^r | x_i, \varphi') + \lambda_r ||\varphi'||^2\) |
| repeat |
| \(\theta\)-step: \(\forall (r,s) \in E \rightarrow \theta_{rs} = \arg \min_{\theta_r} - \sum_{i=1}^{N} \ln \Pr(y_r^r, y_s^s | x_i, \theta')\) |
| \(\varphi\)-step: \(\forall r \in V \rightarrow \varphi_r = \arg \min_{\varphi_r^{N}} - \sum_{i=1}^{N} NCL_\varphi + \lambda_r ||\varphi'||^2\) |
| until convergence of \(NCL\) (Eq. 10) |

Initially, we form an independence model by setting \(E = \emptyset\) that treats each AU independently. After learning the parameters of the ordinal marginals \(\{\varphi\}\), we either consider a fully connected graph (COR-Full) or apply \textit{Glasso} optimization to infer the sparse graph, i.e., to identify the pairs of AUs that we later model with the copula functions (COR-L). During the \(\theta\)-step, we cycle through \(E\) and independently optimize the parameters of the bivariate copula function for each pair \((r, s) \in E\). Note that this can be
performed efficiently using parallel estimation of the association parameters \( \theta_{rs} \). Given the newly estimated copula parameters, in the \( \varphi \)-step, we minimize the objective function in Eq.12 w.r.t. the parameters of the ordinal marginals, i.e., \( \varphi \). Specifically, we optimize the marginal parameters of each AU (\( \varphi^g \)) by using the unary and edge potentials where the target AU is present. We do so in parallel for all AUs. After the \( \varphi \)-step, we refine the association parameters \( \theta \). We continue iterating between these two steps until convergence of the NCL objective function. In our experiments, the algorithm converged in less than 5 iterations. The advantage of the proposed learning approach over direct optimization is two-fold: (i) the estimation of the association and marginal parameters can be parallelized, thus leading to the computational complexity similar to that of marginal models. (ii) In the \( \varphi \)-step, we tune the regularization parameter \( \lambda \) separately for each AU, using the balanced intensity levels for that AU (i.e., a subset of \( N \) training examples where the number of 0 intensity levels is balanced with the intensity 1). Note that in the case of the joint optimization, a single \( \lambda \) need be used, since cross-validation of AU-specific \( \lambda \) is infeasible. This process is summarized in Alg.1.

**Inference.** The inference of test data in undirected graphical models is in general \( np \)-hard problem due to the need to evaluate all possible label configurations. Because of this, we resort to one of the most popular approximate decoders based on the message-passing and dual decomposition algorithms. Specifically, we employed the AD3 decomposition algorithm [4] where the original \( np \)-hard problem is divided into a set of subproblems which are solved independently using local message-passing, and their solutions are then combined to compute a global update. In our experiments, this algorithm achieved a near-real time joint decoding of 10+ target AUs in the inference step.

3. Experiments

**Data.** We evaluate the proposed model on two benchmark datasets: UNBC-MacMaster Shoulder Pain Expression Archive (PAIN) [18] and Denver Intensity of Spontaneous Facial Actions (DISFA) [22]. The PAIN dataset contains video recordings of 25 patients suffering from chronic shoulder pain while performing a range of arm motion exercises, while DISFA contains video recordings of 27 subjects while watching YouTube videos. Each frame is coded in terms of the AU intensity on a six-point ordinal scale. For the experiments presented here, we used all 12 AUs from DISFA, and 10 AUs from PAIN (see the AU numbers in Fig. 3). Since these data contain predominantly expressionless faces (i.e., 0 intensity level), the image frames with at least two active AUs (intensity levels > 1) were used. Also, because the intensity of the target AUs are extremely imbalanced in these data, we merged levels 5 and 6 as for some AUs as only few examples of the highest intensity levels were present. The resulting distribution of the used intensity levels is depicted in Fig. 2.

**Features.** We used the geometric facial features in our experiments, as in [13]. Namely, we used the locations of 49 out of 66 fiducial facial points (provided by the database creators) extracted from facial images in each dataset, using the 2D Active Appearance Model (2D-AAM) [21]. We removed the points from the chin line, as these do not affect the estimation of target AUs. We then registered the 49 facial points to a reference face (average points in each dataset) using an affine transformation. To reduce the dimensionality of the features, we applied PCA, retaining 97% of the energy. This resulted in approximately 20 dimensional feature vectors.

**Evaluation metrics.** Since the goal is AU intensity estimation to measure the performance of the compared approaches we use Pearson correlation coefficient (CORR). CORR is commonly used to measure the linear association between predicted and actual labels, but it ignores their scale. For this reason, we also report the Mean Squared Error (MSE), which is commonly used to measure regression and ordinal classification performance [14, 26]. It also encodes how inconsistent the classifier is in regard to the relative order of the classes, which is important when doing the intensity estimation. We also report Intra-class Correlation (ICC(3,1)), which is commonly used in behavioral sciences to measure agreement between annotators (in our case, the AU intensity labels and model predictions).

**Evaluation procedure.** We compare the performance of the proposed COR model learned in three setting: (i) COR-Full - using the fully connected graph (thus, modeling all pairs of AUs), (ii) COR-LD - using the sparse lasso graph of AU pairs. Both COR-Full and COR-LD are optimized using the direct optimization of the model parameters. (iii) COR-LIT - is the COR model with sparse lasso graph and proposed two-step learning approach. The learned sparse-lasso graphs are depicted in Fig. 3. We also compare these approaches to the standard ordinal regression (SOR) model [1], which uses the same marginal distribution as in the node potentials of our COR models. We also report results ob-
Figure 3: The global AU relations depicted in terms of correlation coefficients. The negative corr is depicted in red, and positive corr in blue, while their magnitude is proportional to the thickness of the line. Note that glasso removes the majority of AU pairs from the precision matrix, preserving only the strongest partial correlations. These are later modeled in the proposed COR-L model using the copula functions. The values of the learned association parameters $\theta$ (using COR-LIT) in most cases resemble the correlations of target pairs encoded in $S$, as expected.
Table 1: The intensity estimation results on the DISFA & PAIN datasets for different AUs.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>FAU</th>
<th>DISFA</th>
<th>PAIN</th>
<th>avg.</th>
<th>avg.</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1</td>
<td>2</td>
<td>4</td>
<td>5</td>
<td>6</td>
</tr>
<tr>
<td>COR-LIT</td>
<td>0.38</td>
<td>0.61</td>
<td>0.37</td>
<td>0.65</td>
<td>0.55</td>
</tr>
<tr>
<td>COR-LD</td>
<td>0.38</td>
<td>0.61</td>
<td>0.37</td>
<td>0.65</td>
<td>0.51</td>
</tr>
<tr>
<td>COR-Full</td>
<td>0.29</td>
<td>0.62</td>
<td>0.27</td>
<td>0.69</td>
<td>0.51</td>
</tr>
<tr>
<td>SOR</td>
<td>0.24</td>
<td>0.54</td>
<td>0.29</td>
<td>0.59</td>
<td>0.53</td>
</tr>
<tr>
<td>SVM</td>
<td>0.29</td>
<td>0.47</td>
<td>0.31</td>
<td>0.61</td>
<td>0.48</td>
</tr>
<tr>
<td>MLR</td>
<td>0.28</td>
<td>0.45</td>
<td>0.30</td>
<td>0.54</td>
<td>0.45</td>
</tr>
<tr>
<td>LT [13]</td>
<td>0.29</td>
<td>0.44</td>
<td>0.26</td>
<td>0.33</td>
<td>0.64</td>
</tr>
<tr>
<td>COR-LIT</td>
<td>0.48</td>
<td>0.63</td>
<td>0.41</td>
<td>0.70</td>
<td>0.66</td>
</tr>
<tr>
<td>COR-LD</td>
<td>0.45</td>
<td>0.63</td>
<td>0.41</td>
<td>0.70</td>
<td>0.53</td>
</tr>
<tr>
<td>COR-Full</td>
<td>0.32</td>
<td>0.63</td>
<td>0.33</td>
<td>0.71</td>
<td>0.53</td>
</tr>
<tr>
<td>SOR</td>
<td>0.27</td>
<td>0.61</td>
<td>0.33</td>
<td>0.69</td>
<td>0.51</td>
</tr>
<tr>
<td>SVM</td>
<td>0.32</td>
<td>0.51</td>
<td>0.36</td>
<td>0.62</td>
<td>0.50</td>
</tr>
<tr>
<td>MLR</td>
<td>0.31</td>
<td>0.51</td>
<td>0.35</td>
<td>0.58</td>
<td>0.49</td>
</tr>
<tr>
<td>LT [13]</td>
<td>0.33</td>
<td>0.46</td>
<td>0.38</td>
<td>0.41</td>
<td>0.66</td>
</tr>
</tbody>
</table>

(a) (left) Co-occurrence of AU1 and AU2 intensity labels, (middle) co-occurrence of their independent predictions, (right) co-occurrence of their joint predictions.

(b) Intensity thresholds for (left) AU1 and (right) AU2. Note that with the learned thresholds, the marginal model for AU1 can never correctly predict levels 1&3, which is overcome by the joint inference in COR model.

Figure 4: Comparison between SOR and COR-LIT models on AU1&AU2 on the DISFA dataset.

seen, AU1 marginal model is incapable of predicting levels 1&3, due to the highly imbalanced data. Yet, due to the strong learned association between AU1&2 (see Fig.3), the joint model overcomes this. Taken together, these results show: (i) that it is important to account for dependencies among the intensity levels of different AUs, (ii) that joint ordinal modeling of AU intensity bridges the limitations of the static nominal classifiers, originally designed for binary classification. Additional qualitative results and a demo-video demonstrating the performance of the proposed COR model are provided in the supplementary material.

4. Conclusions

We proposed a novel Copula Ordinal Regression model for joint modeling and estimation of intensities of AUs from facial images. We showed that by endowing the model with separate but coupled marginal and dependency components, we can successfully capture correlations between different facial features and co-occurrences of various AUs. This approach generalizes prior methods that rely on independent models by using an efficient parametric and flexible representation of the copula functions tied together through a CRF model. The proposed model outperforms related independent models and the state-of-the-art approach for joint intensity estimation of AUs.

Acknowledgments

This work has been funded by the European Community Horizon 2020 under grant agreement no. 645094 (SEWA). The work by R. Walecki is further supported by the European Community Horizon 2020 under grant agreement no. 688835 (DE-ENIGMA). The work of V. Pavlovic has been funded by the National Science Foundation under Grant no. IIS0916812.
References