This CVPR paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the version available on IEEE Xplore.

Copula Ordinal Regression for Joint Estimation of Facial Action Unit Intensity

Robert Walecki*, 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].

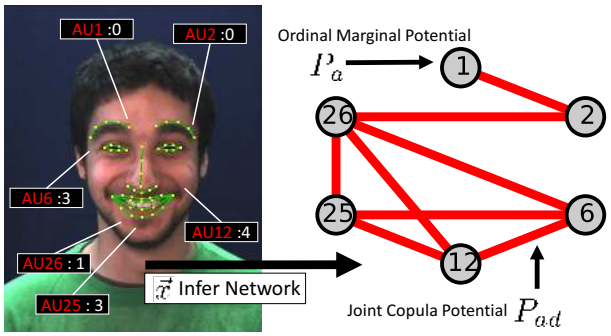


Figure 1: The AU intensity estimation with the proposed Copula Ordinal Regression using the random fields framework. The pruning of the edges in the fully connected graph is accomplished by learning the sparse graph of AU relationships using graph lasso. P_a and P_{ad} denote the node potentials (modeled using marginal ordinal models for each AU) and the edge potentials (modeled using Copula functions accounting for dependencies among the pairs of target AUs), while the $\vec{x}$ represents the input features (a set of fiducial facial points), in the proposed random field model.

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].

To date, most of the work on automated analysis of AUs has focused on detection of the presence/absence of AUs (*e.g*., [17, 3, 20]) instead of their full range intensity estimation. Furthermore, few methods attempted joint modeling of AUs activations (*e.g*., [33, 7]). However, these methods can deal only with the $binary$ classification problems, and, thus, are not applicable to the joint estimation of intensity of multiple AUs. Because the AU intensity estimation is a relatively new problem in the field, few works have addressed it so far. Most of these works perform independent estimation of the AU intensity using either classification-based approach [19, 22, 26] or regression-based approach [12, 11]. To the best of our knowledge, the only methods that attempt joint estimation of multiple AUs intensity are reported in [27, 16, 13]. The methods [27, 16] perform a two stage joint modeling of AU intensity. Specifically, in [27], the scores of the pre-learned regressors, such as Support Vector Regression, are fed into a set of Markov Random Field trees, used to model dependencies of subsets of AUs. Similarly, [16] models AU dependencies using a Dynamic Bayesian Network (DBN) approach, which feeds as inputs the AU-specific spectral regressors. The current state-of-the-art approach for the joint modeling of the AU intensity [13] formulates a generative MRF model, called Latent Tree (LT). In contrast to the two works mentioned above, this method can deal with the highly noisy and missing input features due to its generative component. Nevertheless, there are several critical limitations of the proposed approaches. The model outputs in [27] are treated as continuous, despite the fact that the intensity levels are defined on an ordinal (discrete) scale. Furthermore, in performing the two-stage learning, [27, 16] fail to allow the input features to influence the learned AU dependencies. Although defined in a probabilistic manner, the LT approach [13] relies on a set of heuristics for the model to be computationally tractable for more than few AUs.

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 $\mathcal{D} = \{\mathbf{Y}, \mathbf{X}\}$. $\mathbf{Y} = [\mathbf{y}_1, \ldots, \mathbf{y}_i, \ldots, \mathbf{y}_N]^T$ is comprised of N instances of multivariate outputs stored in $\mathbf{y}_i = \{\mathbf{y}_i^1, \ldots \mathbf{y}_i^q, \ldots \mathbf{y}_i^Q\}$, where Q is the number of AUs, and $\mathbf{y}_i^q$ takes one of $\{1, ..., L^q\}$ discrete intensity levels of the q-th AU. Furthermore, $\mathbf{X} =$

$[\mathbf{x}_1, \ldots, \mathbf{x}_i, \ldots, \mathbf{x}_N]^T$ are input features (e.g., facial points) that correspond to the combinations of labels in $\mathbf{Y}$. Thus, our goal is to simultaneously estimate the combination of the intensity levels $\mathbf{y}^q$ of Q AUs, given the facial features $\mathbf{x}$. 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 $y^q_* \in \Re$ as a function of covariates x, and then relate this latent projection to the ordinal level (y^q) through the threshold bounds:

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

where $x \in \Re^D$, β^q is the ordinal projection vector, ψ^q_l is the lower bound threshold for count level l ($\psi^q_0 = -\infty < \psi^q_1 < \psi^q_2 ... < \psi^q_{L-1} < \psi^q_L = +\infty$). The error (noise) terms ε^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 $F(z^q) = \Pr(\varepsilon^q < z^q)$. 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) $F(z^q) = \Phi(z^q) = \int_{-\infty}^{z^q} \mathcal{N}(\xi; 0, 1) d\xi$. Then, classification in ordinal regression models is performed using the following $ordinal$ likelihood [1]:

$$l^* = \underset{l=1...L}{\operatorname{argmax}} \Pr(y^q = l | y^q_*) = \underset{l=1...L}{\operatorname{argmax}} F(z^q_l) - F(z^q_{l-1}), \quad (2)$$

where $z^q_k = \frac{(\psi^q_k - \beta^q x^T)}{\sigma^q}$ 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:*

$$C\left(u^1, \ldots, u^Q\right) = F\left(F^{-1}(u^1), \ldots, F^{-1}(u^Q)\right) \quad (3)$$

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

$$F(y^1, \ldots, y^Q) = \mathrm{C}(F(y^1), \ldots, F(y^Q)), \quad (4)$$

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:

$$\begin{aligned} &\Pr(y^1 = l^1, \ldots, y^Q = l^Q) = \\ &\Pr(\psi_{l^1-1} < y^1_* < \psi_{l^1}, \ldots, \psi_{l^Q-1} < y^Q_* < \psi_{l^Q}) \\ &= \sum_{c_1=0}^{1} \cdots \sum_{c_Q=0}^{1} (-1)^{c_1+...+c_Q} F(z^1_{l^1-c_1}, ..., z^Q_{l^Q-c_Q}) \\ &= \sum_{c_1=0}^{1} \cdots \sum_{c_Q=0}^{1} (-1)^{c_1+...+c_Q} C_\theta(u^1_{l^1-c_1}, ..., u^Q_{l^Q-c_Q}) \end{aligned} \quad (5)$$

where $u^q_{l^q-c_q} = F(z^q_{l^q-c_q}), c_q \in \{0, 1\}$, is defined in Sec.2.1, and θ 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 $\varepsilon^1, \ldots, \varepsilon^Q$ via the copula. Secondly, the joint density model induced by the copula is conditioned on the covariates x, i.e., $F(y^1, \ldots, y^Q) \leftarrow F(y^1, \ldots, y^Q | x)$. 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 $\mathbf{y}$ 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:

$$\begin{aligned}\Pr(y^1 = l^1, y^2 = l^2) = F(z^1_{l_1}, z^2_{l_2}) \\ +F(z^1_{l_1-1}, z^2_{l_2-1}) - F(z^1_{l_1-1}, z^2_{l_2}) - F(z^1_{l_1}, z^2_{l_2-1})\end{aligned} \quad (6)$$

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:

$$C_\theta(u^1, ..., u^Q) = -\frac{1}{\theta} \ln \left(1 + \frac{\prod_{q=1}^{Q} (e^{-\theta u^q} - 1)}{(e^{-\theta} - 1)^{Q-1}} \right). \quad (7)$$

The dependence parameter $\theta \in (-\infty, +\infty) \backslash \{0\}$, and the perfect positive/negative dependence is obtained if $\theta \rightarrow \pm\infty$. When $\theta \rightarrow 0$, 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 dependence 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 $\mathcal{G} = (V, \mathcal{C})$, where nodes $v \in V, |V| = Q$, correspond to individual AUs and cliques $c \in \mathcal{C}$ correspond to subsets of dependent AUs modeled using the copula functions. The joint probability distribution of Q intensity random variables is then defined as:

$$P(\mathbf{y}|x, \Omega) = \frac{1}{Z} \prod_{c \in \mathcal{C}} \Psi(\mathbf{y}_c|x), \quad (8)$$

where Z is the partition function, $\mathbf{y}_c$ is the subset of random variables in clique c, $\Psi(\cdot)$ is the conditional potential on the labels in this clique, explained below, and $\Omega = \{\vartheta, \theta\}$ 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, $\mathcal{C} = V \cup E$, where E is the set of edges in $\mathcal{G}$. Hence,

$$\Psi(\mathbf{y}_c|x) = \begin{cases} \Pr(y^r|x), & \begin{matrix} c = r \in V \\ \text{unary clique} \end{matrix} \\ \Pr(y^r, y^s|x)^\gamma, & \begin{matrix} c = (r, s) \in E \\ \text{pairwise clique} \end{matrix} \end{cases} \quad (9)$$

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 ϑ_r parameters of the ordinal regression model, while the edge potentials depend also on the copula association parameter θ_{rs} that models the dependency of (r, s) pair of outputs. Furthermore, the weight γ 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 also 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 $\theta_{rs} \rightarrow 0$, 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 Ω 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:

$$NCL = -\sum_{i=1}^{N}\left[\sum_{r\in V}\ln \Pr(y_i^r|x_i) + \gamma \sum_{(r,s)\in E}\ln \Pr(y_i^r, y_i^s|x_i)\right], \tag{10}$$

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 Υ 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:

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

where κ 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. Ω. 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 ψ: $\{\psi_{j-1} \leq \psi_j \text{ for } j = 1,\dots,L\}$, and (ii) the positive scale constraint on σ: $\{\sigma > 0\}$. To avoid constrained optimization, we introduce a re-parametrization of ψ using displacement variables δ_k, where $\psi_j = \psi_1 + \sum_{k=1}^{j-1}\delta_k^2$ for $j = 2,\dots,L-1$. The positiveness constraint for σ is simply handled by introducing the free parameter σ_0 where $\sigma = \sigma_0^2$. Thus, the unconstrained parameters of the ordinal marginals are $\{\beta, \psi_1, \delta_1, \dots, \delta_{L-2}, \sigma_0\}$, and they are defined separately for each of the Q ordinal marginals, and stored in φ.

Training. During training, we seek to find optimal parameters Ω^* by solving the regularized optimization problem

$$\Omega^* = \underset{\Omega=\{\varphi,\theta\}}{\arg\min}\quad NCL(\varphi,\theta) + \lambda R_\varphi, \tag{12}$$

where NCL is given by Eq.10, R_φ is the standard L_2 regularizer of the projection β and σ_0, and λ is the regularization parameter. No specific regularization is necessary for the threshold parameters as they are automatically adjusted according to the score $\beta x^\top$.

Solving for the parameters $\Omega = \{\varphi, \theta\}$ directly is possible, however, by noticing that the copula parameters θ are independent of the node potentials in the NCL, we can alternate between optimization of the marginals φ and the copula association θ. 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 $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^N$
Output: Model parameters $\Omega = \{\varphi, \theta\}$

Initialization:
$\forall (r,s) \in E \rightarrow \theta_{rs} = sign(corr(y^r, y^s))$
$\forall r \in V \rightarrow \varphi_r = \arg\min_{\varphi'} -\sum_{i=1}^{N\in AU_r} \ln Pr(y_i^r \mid x_i, \varphi') + \lambda_r \|\varphi'\|^2$

repeat

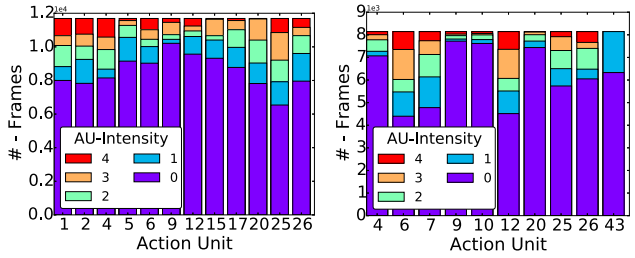
θ-step: $\forall (r,s) \in E \rightarrow \theta_{rs} = \arg\min_{\theta'} -\sum_{i=1}^{N} \ln Pr(y_i^r, y_i^s \mid x_i, \theta')$

φ-step: $\forall r \in V \rightarrow \varphi_r = \arg\min_{\varphi'} -\sum_{i=1}^{N\in AU_r} NCL_i + \lambda_r \|\varphi'\|^2$

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 *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 θ-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

[2] We used the Glasso Matlab code from [8].



(a) DISFA (b) Shoulder-Pain

Figure 2: Distribution of the AU intensity levels.

performed efficiently using parallel estimation of the association parameters θ_{rs}. Given the newly estimated copula parameters, in the φ-step, we minimize the objective function in Eq.12 w.r.t. the parameters of the ordinal marginals, i.e., φ. Specifically, we optimize the marginal parameters of each AU (φ^q) by using the unary and edge potentials where the target AU is present. We do so in parallel for all AUs. After the φ-step, we refine the association parameters θ. 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 φ-step, we tune the regularization parameter λ 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 λ need be used, since cross-validation of AU-specific λ 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.

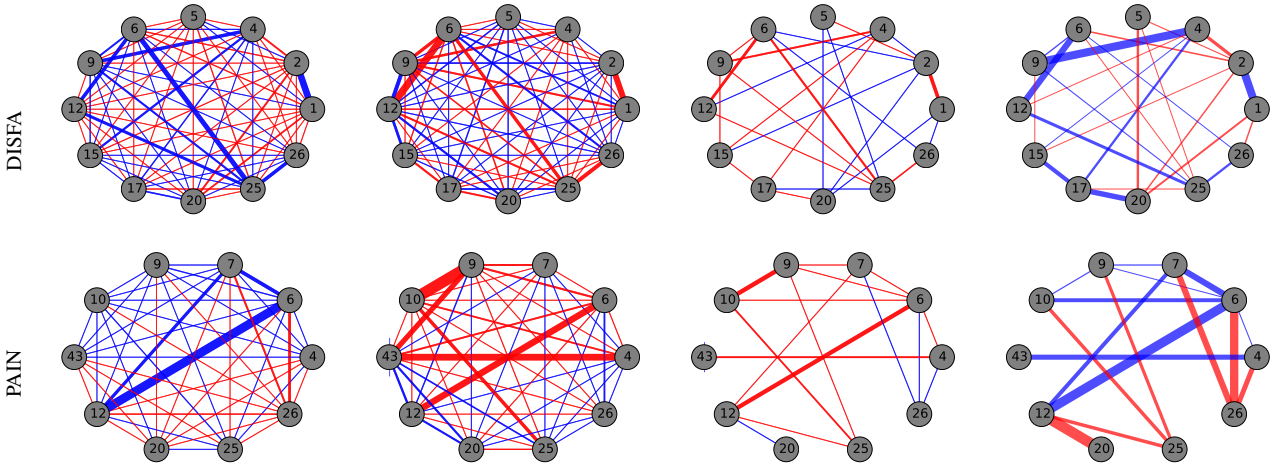
3. Experiments

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 Squard 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-



(a) Correlation (S) (b) Precision (S^{-1}) (c) Lasso Precision (Υ) (d) Learned association θ

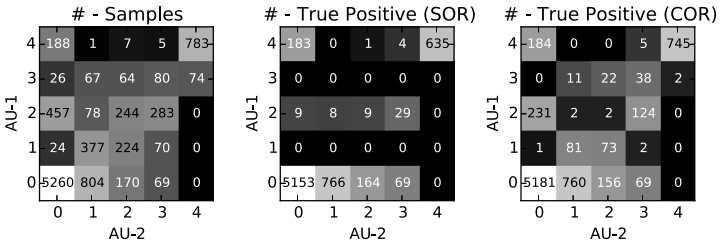
Figure 3: The global AU relations depicted in terms of correlation coefficients. The negative corr is depicted in red, and positive corr in blue, while their magnitude is proportional to the thickness of the line. Note that glasso removes the majority of AU pairs from the precision matrix, preserving 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 θ (using COR-LIT) in most cases resemble the correlations of target pairs encoded in S, as expected.

tained using the multiple Logistic Regression (MLR) [1] model - which ignores the class ordering and learns a separate projection β for each label. We also include the results attained by commonly used methods for AU intensity estimation, i.e., Support Vector Machines (SVM). SVM was used as the baseline on DISFA[22] and PAIN [18] datasets, by treating each of the intensity levels as a separate class. We apply the RBF kernel for SVM and optimize all hyper-parameters by a grid search as in the rest of the methods by searching over the range $\{10^{\pm 4}, 10^{\pm 3}, ..., 0\}$, and selecting the one that perform best on a validation set (20% of data not overlapping with test data). Note that these models support only a single output, therefore we train a separate model per AU. Finally, we compare our approach to the state-of-the-art for the target task - Latent Trees (LT-all) [13]. The authors of LT provided their source code, so all comparisons were performed in the same settings. In all our experiments, we applied a 5-fold cross validation procedure, with each fold containing data of different subjects.

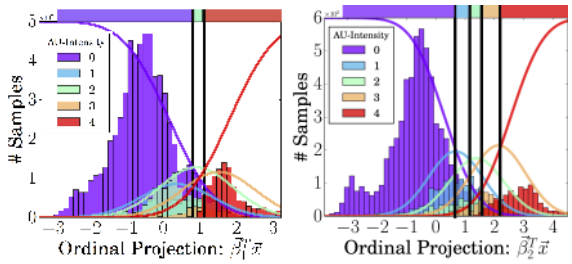
Table 1 shows the comparative results for different approaches evaluated on the DISFA and PAIN datasets. We make several observations: on average, both SVM and SOR achieve similar results, with the latter outperforming SVM in MSE, as expected. Also, SOR largely outperforms its non-ordinal counterpart, MLR, across all three measures. Compared to the state-of-the-art LT method, the independent output models achieve similar or better average performance in the evaluation setting. However, this method outperforms the afore-mentioned methods in MSE, despite the fact that it ignores the ordinal scale of target labels. Such performance of LT has also been observed by the authors [13], who showed that their approach shows significant improvements on highly noisy features due to its generative part. However, this robustness has not been obtained in our experiments on the target data. Compared to the proposed approaches, the COR models outperform the compared models on average. This is particularly evident in the ICC scores, where the average difference is 3% for COR-Full, and 6% for the COR-LIT. Similar trend can be observed in CORR measure, while in MSE this difference is less pronounced. Overall, we notice that the joint inference by the proposed models consistently outperforms marginal inference by the COR models, as expected. We attribute this to modeling of the AU dependencies through the copula functions. Next, we observe that both COR-LD & COR-LIT outperform (on average) COR-Full across all three measures, with COR-LIT performing the best. This is expected as both lasso-based models are less prone to overfitting, in contrast to the COR-FULL model. This also signals that only the partial correlations revealed by the sparse lasso are sufficient to improve the joint inference. On the other hand, comparing the COR-LIT & COR-LD, there is a slight difference on average. However, looking into CORR of AU6 & 9 in DISFA, and AU9 in PAIN, we see that the COR-LIT performs significantly better on these particular AUs. We found that this was due to its ability to tune the regularization parameters specifically for these two AUs, which, in the direct inference, is infeasible. In Fig. 4, we further demonstrate the benefit of joint inference over using the marginal (SOR) models for the target task. As can be

Table 1: The intensity estimation results on the DISFA & PAIN datasets for different AUs.

	Dataset	DISFA													PAIN										
	FAU	1	2	4	5	6	9	12	15	17	20	25	26	avg.	4	6	7	9	10	12	20	25	26	43	avg.
ICC (3,1)	COR-LIT	**0.38**	0.61	**0.37**	0.65	0.55	**0.39**	**0.58**	0.15	**0.22**	0.16	0.86	0.53	**0.46**	0.34	0.45	**0.42**	**0.45**	0.32	**0.41**	0.00	0.29	0.07	**0.54**	**0.33**
	COR-LD	**0.38**	0.61	**0.37**	0.65	0.51	0.38	**0.58**	0.15	0.21	0.16	0.86	0.53	0.45	0.34	**0.46**	0.41	0.38	0.32	**0.41**	0.00	0.29	0.07	**0.54**	0.32
	COR-Full	0.29	**0.62**	0.27	**0.69**	0.51	0.28	0.54	0.13	0.21	0.15	0.87	**0.54**	0.43	**0.35**	0.38	**0.42**	0.11	0.33	0.39	0.03	0.28	**0.09**	0.52	0.29
	SOR	0.24	0.54	0.29	0.59	0.53	0.27	0.49	0.18	0.14	0.17	0.79	0.51	0.39	0.31	0.39	0.36	0.01	0.24	0.38	0.03	0.17	0.00	0.46	0.24
	SVM	0.29	0.47	0.31	0.61	0.48	0.31	0.49	**0.22**	0.08	**0.19**	0.85	0.49	0.40	0.30	0.39	0.39	0.18	0.22	0.32	0.00	0.24	0.04	0.42	0.25
	MLR	0.28	0.45	0.30	0.54	0.45	0.30	0.44	0.16	0.06	0.21	0.71	0.45	0.36	0.14	0.36	0.38	0.09	0.18	0.24	**0.08**	0.24	0.04	0.45	0.22
	LT [13]	0.29	0.44	0.26	0.33	**0.64**	0.23	0.52	0.21	0.21	0.13	**0.88**	0.30	0.37	0.29	0.34	0.29	0.11	**0.39**	0.33	0.03	**0.39**	0.07	0.49	0.27
CORR	COR-LIT	**0.48**	**0.63**	**0.41**	0.70	**0.66**	**0.48**	**0.61**	0.18	**0.23**	0.19	0.87	0.56	**0.49**	**0.43**	**0.47**	**0.48**	**0.47**	0.34	**0.48**	0.00	0.32	0.12	**0.58**	**0.37**
	COR-LD	0.45	**0.63**	**0.41**	0.70	0.53	0.40	**0.61**	0.18	**0.23**	0.19	0.87	0.56	0.48	**0.43**	**0.47**	**0.48**	0.44	0.34	**0.48**	0.00	0.32	0.12	**0.58**	**0.37**
	COR-Full	0.32	**0.63**	0.33	**0.71**	0.53	0.32	0.57	0.15	0.21	**0.21**	0.88	**0.57**	0.45	**0.43**	0.40	0.47	0.14	0.34	0.44	**0.08**	0.28	**0.14**	0.54	0.33
	SOR	0.27	0.61	0.33	0.69	0.51	0.23	0.54	0.14	0.14	0.17	**0.89**	0.56	0.42	0.36	0.41	0.43	0.03	0.27	0.40	0.03	0.17	0.02	0.49	0.26
	SVM	0.32	0.51	0.36	0.62	0.50	0.29	0.50	0.22	0.09	**0.21**	0.85	0.50	0.41	0.30	0.39	0.41	0.19	0.23	0.33	0.00	0.23	0.04	0.45	0.26
	MLR	0.31	0.51	0.35	0.58	0.49	0.30	0.50	0.18	0.07	0.24	0.74	0.45	0.39	0.15	0.37	0.43	0.09	0.20	0.24	**0.08**	0.23	0.08	0.47	0.23
	LT [13]	0.33	0.46	0.38	0.41	**0.66**	0.23	0.56	**0.35**	0.14	0.12	**0.89**	0.29	0.40	0.31	0.43	0.32	0.12	**0.40**	0.33	0.03	**0.39**	0.09	0.49	0.29
MSE	COR-LIT	1.74	**1.09**	1.78	0.58	**0.68**	0.68	**0.43**	0.87	1.08	1.44	0.73	1.03	**1.01**	**0.52**	**2.57**	**1.51**	**0.26**	**0.14**	**2.18**	**0.32**	1.43	1.88	**0.14**	**1.10**
	COR-LD	**1.68**	**1.09**	2.19	0.58	0.70	0.68	**0.43**	1.07	1.08	1.69	0.73	1.13	1.09	0.71	**2.57**	1.62	0.31	0.21	**2.18**	0.38	1.73	1.88	**0.14**	1.17
	COR-Full	2.10	1.26	2.14	0.54	0.88	1.00	0.53	0.92	1.06	1.90	**0.49**	1.01	1.15	0.72	2.74	1.52	0.42	0.24	2.46	0.38	1.89	1.83	0.15	1.24
	SOR	2.24	1.37	2.19	**0.30**	0.98	1.00	0.50	0.98	1.05	1.69	0.47	**0.94**	1.14	0.84	2.80	1.53	0.47	0.29	2.85	0.38	1.95	1.87	0.17	1.31
	SVM	2.26	1.54	2.32	0.44	1.09	0.96	0.54	0.98	1.06	1.65	0.60	0.98	1.20	0.94	2.74	1.70	0.47	0.40	2.78	0.38	1.79	1.87	0.17	1.32
	MLR	1.96	1.51	2.45	0.55	1.05	0.97	0.71	0.98	1.06	1.66	1.02	1.53	1.29	1.03	2.76	1.87	0.47	0.43	2.98	0.38	1.84	1.87	0.17	1.38
	LT [13]	2.28	1.61	**1.61**	0.74	0.86	**0.67**	0.45	**0.82**	**0.85**	**1.29**	0.54	1.22	1.07	0.98	2.99	1.74	0.41	0.20	2.82	0.38	**1.35**	**1.69**	0.19	1.27



(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

[1] A. Agresti. Analysis of ordinal categorical data. *Wiley Series in Prob. and Stat.*, pages 1–287, 1984.

[2] P. Berkes, F. Wood, and J. W. Pillow. Characterizing neural dependencies with copula models. In *NIPS*, pages 129–136. 2009.

[3] W.-S. Chu, F. D. L. Torre, and J. F. Cohn. Selective transfer machine for personalized facial action unit detection. In *CVPR*, pages 3515–3522, 2013.

[4] D. Das, A. F. Martins, and N. A. Smith. An exact dual decomposition algorithm for shallow semantic parsing with constraints. In *Proc. of the 1st Joint Conf. on Lexical and Computational Semantics-Volume 1*, pages 209–217. Association for Computational Linguistics, 2012.

[5] A. C. Davison, S. Padoan, and M. Ribatet. Statistical modeling of spatial extremes. *Statistical Science*, pages 161–186, 2012.

[6] P. Ekman, W. V. Friesen, and J. C. Hager. Facial action coding system. *Manual: A Human Face*, 2002.

[7] S. Eleftheriadis, O. Rudovic, and M. Pantic. Multi-conditional latent variable model for joint facial action unit detection. In *ICCV*, 2015.

[8] J. Friedman, T. Hastie, and R. Tibshirani. Sparse inverse covariance estimation with the graphical lasso. *Biostatistics*, pages 432–441, 2008.

[9] C. Genest. Frank's family of bivariate distributions. *Biometrika*, pages 549–555, 1987.

[10] S. Horvath. Weighted network analysis: Applications in genomics and systems biology. *Springer Science and Business Media*, 2011.

[11] L. A. Jeni, J. M. Girard, J. F. Cohn, and F. D. L. Torre. Continuous au intensity estimation using localized, sparse facial feature space. *FG*, pages 1–7, 2013.

[12] S. Kaltwang, O. Rudovic, and M. Pantic. Continuous pain intensity estimation from facial expressions. In *AIVS*, pages 368–377, 2012.

[13] S. Kaltwang, S. Todorovic, and M. Pantic. Latent trees for estimating intensity of facial action units. In *CVPR*, 2015.

[14] M. Kim and V. Pavlovic. Structured output ordinal regression for dynamic facial emotion intensity prediction. *ECCV*, pages 649–662, 2010.

[15] J. D. Lafferty, A. McCallum, and F. C. N. Pereira. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *ICML*, pages 282–289, 2001.

[16] Y. Li, S. M. Mavadati, M. H. Mahoor, and Q. Ji. A unified probabilistic framework for measuring the intensity of spontaneous facial action units. In *FG*, pages 1–7, 2013.

[17] P. Lucey, J. F. Cohn, I. Matthews, S. Lucey, S. Sridharan, J. Howlett, and K. M. Prkachin. Automatically detecting pain in video through facial action units. *TSMCB*, pages 664–674, 2011.

[18] P. Lucey, J. F. Cohn, K. M. Prkachin, P. E. Solomon, and I. Matthews. Painful data: The unbc-mcmaster shoulder pain expression archive database. In *FG*, pages 57–64, 2011.

[19] M. Mahoor, S. Cadavid, D. Messinger, and J. Cohn. A framework for automated measurement of the intensity of non-posed facial action units. *CVPR*, pages 74–80, 2009.

[20] M. H. Mahoor, M. Zhou, K. L. Veon, S. M. Mavadati, and J. F. Cohn. Facial action unit recognition with sparse representation. In *FG*, pages 336–342, 2011.

[21] I. Matthews and S. Baker. Active appearance models revisited. *IJCV*, pages 135–164, 2004.

[22] S. M. Mavadati, M. H. Mahoor, K. Bartlett, P. Trinh, and J. F. Cohn. Disfa: A spontaneous facial action intensity database. *TAC*, pages 151–160, 2013.

[23] R. Mazumder and T. Hastie. Exact covariance thresholding into connected components for large-scale graphical lasso. *JMLR*, pages 781–794, 2012.

[24] J. E. Pessa, V. P. Zadoo, P. A. Garza, E. K. Adrian, A. I. Dewitt, and J. R. Garza. Double or bifid zygomaticus major muscle: anatomy, incidence, and clinical correlation. *Clinical Anatomy*, pages 310–313, 1998.

[25] C. Rasmussen and C. Williams. *Gaussian processes for machine learning*. The MIT Press, 2006.

[26] O. Rudovic, V. Pavlovic, and M. Pantic. Context-sensitive dynamic ordinal regression for intensity estimation of facial action units. *TPAMI*, pages 944–958, 2014.

[27] G. Sandbach, S. Zafeiriou, and M. Pantic. Markov random field structures for facial action unit intensity estimation. In *ICCV*, 2013.

[28] J. H. Shih and T. A. Louis. Inferences on the association parameter in copula models for bivariate survival data. *Biometrics*, pages 1384–1399, 1995.

[29] A. Sklar. Random variables, distribution functions, and copulas: a personal look backward and forward. *Lecture notes-monograph series*, pages 1–14, 1996.

[30] C. Varin, N. Reid, and D. Firth. An overview of composite likelihood methods. *Statistica Sinica*, pages 5–42, 2011.

[31] R. Winkelmann. *Econometric analysis of count data*. Springer Science & Business Media, 2003.

[32] Y. Zhang and J. Schneider. A composite likelihood view for multi-label classification. In *Int'l Conf. on Artificial Intelligence and Statistics*, pages 1407–1415, 2012.

[33] K. Zhao, W.-S. Chu, F. De la Torre, J. F. Cohn, and H. Zhang. Joint patch and multi-label learning for facial action unit detection. In *CVPR*, 2015.