Variational Gaussian Process Auto-Encoder for Ordinal Prediction of Facial Action Units

Stefanos Eleftheriadis* Ognjen Rudovic* Marc Peter Deisenroth* Maja Pantic*†
{stefanos, orudovic, m.deisenroth, m.pantic}@imperial.ac.uk

*Department of Computing, Imperial College London, UK
†EEMCS, University of Twente, The Netherlands

Abstract. We address the task of simultaneous feature fusion and modeling of discrete ordinal outputs. We propose a novel Gaussian process (GP) auto-encoder modeling approach. In particular, we introduce GP *encoders* to project multiple observed features onto a latent space, while GP *decoders* are responsible for reconstructing the original features. Inference is performed in a novel variational framework, where the recovered latent representations are further constrained by the ordinal output labels. In this way, we seamlessly integrate the ordinal structure in the learned manifold, while attaining robust fusion of the input features. We demonstrate the representation abilities of our model on benchmark datasets from machine learning and affect analysis. We further evaluate the model on the tasks of feature fusion and joint ordinal prediction of facial action units. Our experiments demonstrate the benefits of the proposed approach compared to the state of the art.

1 Introduction

Automated analysis of facial expressions has attracted significant attention because of its practical importance in psychology studies, human-computer interfaces, marketing research, and entertainment, among others [1]. The most objective way to describe facial expressions is by means of the facial action coding system (FACS) [2]. This is the most comprehensive anatomically-based system that can be used to describe virtually all possible facial expressions in terms of 30+ facial muscle movements, named action units (AUs). FACS also defines rules for scoring the intensity of each AU in the range from absent to maximal intensity on a six-point ordinal scale. Therefore, FACS is critical for high-level interpretation of facial expressions. For instance, the high intensity of AU12 (lip corner puller), as in full-blown smiles, may indicate joy. Conversely, its low intensity may indicate fake smiles as in the case of sarcasm.

The machine analysis of AU intensities is challenging mainly due to the complexity and subtlety of human facial behavior as well as individual differences in expressiveness and variations in head-pose, illumination, occlusions, etc. [3]. These sources of variation are typically accounted for at the feature level by means of geometric- and appearance-based features, capturing the geometry and texture changes in a face, respectively. Furthermore, some AUs usually appear

in combination with other AUs. 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 [4]. Furthermore, co-occurring AUs can be non-additive, *e.g.*, if one AU masks another a new and distinct set of appearances is created [2]. Thus, combining different facial features while accounting for AU co-occurrences in a common framework is expected to result in a robust and more accurate estimation of target AUs intensity.

Most existing approaches to AU intensity estimation model each AU independently and cast it as a classification [4,5,6,7,8] or regression [9,10,11,12] task. While classification seems to be a natural choice to handle the problem, the related literature fails to account for the ordinal nature of the target intensity levels (misclassification of different levels is equally penalized). The regression-based approaches model the intensity levels on a continuous scale, which is sub-optimal when dealing with discrete outputs. Similarly, the models that do attempt multiple AU intenisty estimation (*e.g.*, [13,14,15,16,17]) adopt the same sub-optimal approach to deal with the nature of the output as the independent methods. However, they have showed improved performance in the target task due to the modeling of AU co-occurrences. Apart from a few exceptions that treat each AU independently [9,10,7], none of the aforementioned approaches addresses the task of joint output modeling (*i.e.*, multiple AUs) while accounting for different modalities in the input (*i.e.*, fusion of geometric and appearance features). These limitations can naturally be addressed by following recent advances in manifold learning [18,19,20] and, in particular, using the framework of Gaussian processes (GPs) [21]. Within this framework, the problem of feature fusion is transformed to that of learning from multiple views, while continuous-valued predictions can be handled efficiently, for more than one output. However, as with the regression-based models described above, these models treat the ordinal labels as continuous values. This also limits their potential to unravel an 'ordinal' manifold, needed to facilitate estimation of target ordinal intensities.

In this work, we propose a novel manifold-based GP approach based on the Bayesian GP latent variable model (B-GPLVM) [22] that performs simultaneously the feature fusion and joint estimation of the AU ordinal intensity. Specifically, we propse the variational GP auto-encoder (VGP-AE), which is composed of a probabilistic *recognition* model, used to project the observed features onto the manifold, and a generative model, used for their reconstruction. This, in contrast to existing work (*e.g.*, [23]) that applies deterministic back-mappings, allows us to explicitly model the uncertainty in the projections onto the learned manifold. Additionally, we endow the proposed VGP-AE with the ordinal outputs [24]. The fusion of the information from the input features and learning of the joint ordinal output is performed simultaneously in a joint Bayesian framework. In this way, we seamlessly integrate the ordinal structure into the recovered manifold while attaining robust fusion of the target features. To the best of our knowledge, this is the first approach that achieves simultaneous feature fusion and joint AU intensity estimation in the context of facial behavior analysis.

2   Related Work on AU Intensity Estimation

To date, most existing work on automated analysis focuses on the detection of AU activations [25,26,27,28]. The problem of AU intensity estimation is relatively new in the field. Most of the research in this area focuses on independent modeling of AU intensities [4,5,6,7,8,9,10,11,12]. Only recently, joint estimation of the intensity levels has been addressed [13,14,15,16,17]. This is motivated by the fact that intensity annotations are difficult to obtain (due to the tedious process of manually coding) and that AU levels are highly imbalanced. Thus, by imposing the structure on the output in terms of AU co-occurrences robust intensity estimation is expected.

Toward this direction, [13] proposed a two-stage learning strategy, where a multi-class support vector machine (SVM) is first trained for each AU independently. Then, the structure modeling is handled via a dynamic Bayesian network, which captures the semantic relationship among the AU-specific SVMs. In a similar fashion, [14] used support vector regressors (SVR) and a Markov random field (MRF). However, these two-stage approaches are sub-optimal for the target task as the regressors/classifiers and the AU relations are learned independently. To overcome this, [15] proposed to learn latent trees that encode both the input features and (multiple) output AU labels. The structure of the latent variables is modeled using a tree-like graph. However, in the presence of high-dimensional inputs and multiple AUs, this method becomes prohibitively expensive. Moreover, the authors show that with this approach the fusion of different features does not benefit the estimation of AU intensity, achieving similar performance to when individual modalities are used. More recently, [17] proposed a sparse learning approach that uses the notion of robust principal component analysis [29] to decompose expression from facial identity. Then, joint intensity estimation of multiple AUs is performed via a regression model based on dictionary learning. However, this approach can deal with a single modality only. [16] casts the joint AU intensity estimation as a multi-task learning problem based on kernel regression (MLKR). However, in their formulation of the model, the use of MLKR does not scale to high-dimensional features, let alone when using features of different modalities (*e.g.*, geometric and appearance).

The work presented in this paper advances the current state of the art in several aspects: (1) The proposed VGP-AE can efficiently perform the fusion of multiple modalities by means of a shared manifold; (2) Automatic feature selection is implicitly performed via the manifold. The recovered latent representations are used as input to multiple *ordinal* regressors [24], which are concurrently learned in a joint Bayesian framework; (3) GPs allow us to efficiently deal with high-dimensional input and output variables without significantly affecting the model's complexity.

3   Variational Gaussian Process Auto-Encoder

We assume that we have access to a training data set $\mathcal{D} = \{\boldsymbol{Y}, \boldsymbol{Z}\}$, which is comprised of V observed input channels $\boldsymbol{Y} = \{\boldsymbol{Y}^{(v)}\}_{v=1}^{V}$, and the associated output

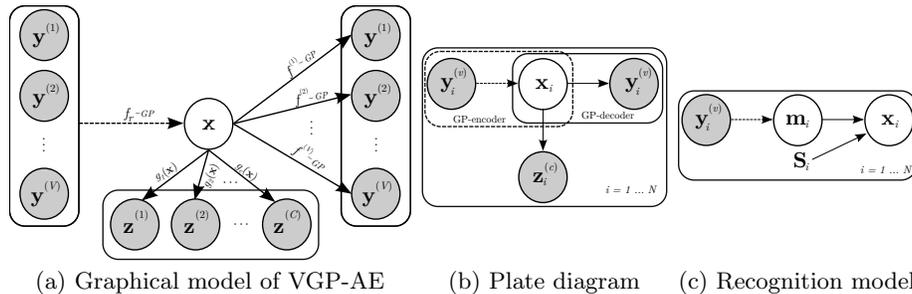(a) Graphical model of VGP-AE   (b) Plate diagram   (c) Recognition model

Fig. 1. The proposed VGP-AE. (a) $f^{(v)}$ and f_r are the GP-decoder and GP-encoder, respectively. The projection of the latent variable $\boldsymbol{x}$ to the labels' ordinal plane is facilitated through the ordinal regression $g(\boldsymbol{x})$. (b) Compact representation of the model. (c) The proposed recognition model (GP-encoder) with the intermediate variable $\boldsymbol{m}$.

labels $\boldsymbol{Z}$. Each input channel consists of N i.i.d. samples $\boldsymbol{Y}^{(v)} = \{\boldsymbol{y}_i^{(v)}\}_{i=1}^N$, where $\boldsymbol{y}_i^{(v)} \in \mathbb{R}^{D_v}$ denotes corresponding facial features. $\boldsymbol{Z} = \{\boldsymbol{z}_i\}_{i=1}^N$ is the common label representation, where $z_{ic} \in \{1, \ldots, S\}$ denotes the discrete, ordinal state of the c-th output (*i.e.*, AU intensity level), $c = 1, \ldots, C$. We are interested in simultaneously addressing the tasks of feature fusion and ordinal prediction of the multiple outputs. For this purpose, we propose an approach that resembles recent work of generative models [30,31]. In these models, auto-encoders are employed to learn compact representations of the input data. In a standard auto-encoding setting, the encoding/decoding functions are modeled via neural networks. Here we replace these functions with probabilistic non-parametric mappings, significantly reducing the number of optimized parameters, and naturally modeling the uncertainty in the mappings. The proposed approach can be regarded as a B-GPLVM (generative model) with a fast inference mechanism based on the non-parametric, probabilistic mapping (recognition model). To achieve this, we impose GP priors on both models, and hence, obtain a well-defined GP-*encoder*, in accordance to the GP-*decoder*.

3.1   The Model

Within the above setting, we assume that the observed features $\boldsymbol{Y}^{(v)}$ are generated by a random process, involving a latent (unobserved) set of variables $\boldsymbol{X} = \{\boldsymbol{x}_i\}_{i=1}^N, \boldsymbol{x}_i \in \mathbb{R}^q$, with $q \ll D_v$. The data pairs $\mathcal{D} = \{\boldsymbol{Y}, \boldsymbol{Z}\}$ are assumed to be conditionally independent given the latent variables, *i.e.*, $\boldsymbol{Y} \perp\!\!\!\perp \boldsymbol{Z} | \boldsymbol{X}$. The random process of recovering the latent variables has two distinctive stages: (a) a latent variable $\boldsymbol{x}_i$ is generated from some general prior distribution $p(\boldsymbol{x}) = \mathcal{N}(\boldsymbol{0}, \boldsymbol{I})$, and further projected to the labels' ordinal plane via $p(\boldsymbol{z}|\boldsymbol{x})$; (b) an observed input $\boldsymbol{y}_i^{(v)}$ is generated from the conditional distribution $p(\boldsymbol{y}^{(v)}|\boldsymbol{x})$. This process is described in Fig. 1(a),(b). Using this approach, we can now perform classification in the lower-dimensional space of $\boldsymbol{X}$. However, this requires access to the intractable true posterior $p(\boldsymbol{x}|\boldsymbol{y}^{(v)})$.

To constrain the distribution of the latent variables we follow [30,31] and introduce the *recognition* model $p_r(\boldsymbol{x}|\boldsymbol{y}^{(v)})$. Hence, we end up with a supervised auto-encoder setting

$$\boldsymbol{y}_i^{(v)}|\boldsymbol{x}_i = f^{(v)}(\boldsymbol{x}_i; \boldsymbol{\theta}^{(v)}) + \epsilon^{(v)}, \quad \boldsymbol{x}_i|\boldsymbol{y}_i^{(v)} = f_r(\boldsymbol{y}_i^{(v)}; \boldsymbol{\theta}_r) + \epsilon_r, \quad \boldsymbol{z}_i|\boldsymbol{x}_i = g(\boldsymbol{x}_i; \boldsymbol{W}), \quad (1)$$

where the latent space is further encouraged to reflect the structure of the output labels. Here, $\epsilon^{(v)} \sim \mathcal{N}(\boldsymbol{0}, \sigma_v^2 \boldsymbol{I})$, $\epsilon_r \sim \mathcal{N}(\boldsymbol{0}, \sigma_r^2 \boldsymbol{I})$. We place GP priors on $f^{(v)}, f_r$ with corresponding hyper-parameters $\boldsymbol{\theta}^{(v)}, \boldsymbol{\theta}_r$.[1] g denotes the ordinal regression that transforms the latent variables to the labels' ordinal plane, via $\boldsymbol{W} = \{\boldsymbol{w}_c\}_{c=1}^C, \boldsymbol{w}_c \in \mathbb{R}^q$.

In the following, we detail how to learn the GP auto-encoder in Eq. (1) by deriving a variational approximation to the log-marginal likelihood

$$\log p(\boldsymbol{Y}, \boldsymbol{Z}) = \log \int p(\boldsymbol{Z}|\boldsymbol{X}) \prod_v p(\boldsymbol{Y}^{(v)}|\boldsymbol{X}) p(\boldsymbol{X}) d\boldsymbol{X}. \quad (2)$$

3.2 Deriving the Lower Bound

We exploit the conditional independence property of $\boldsymbol{Y} \perp\!\!\!\perp \boldsymbol{Z}|\boldsymbol{X}$ and focus our analysis on the GP auto-encoder. The ordinal information from the labels is incorporated in the presented variational framework in Sec. 3.3. As in [28], we place GP priors on $f^{(v)}, f_r$, and after integrating out the mapping functions, we obtain the conditionals

$$p(\boldsymbol{Y}^{(v)}|\boldsymbol{X}) = \mathcal{N}(\boldsymbol{0}, \boldsymbol{K}^{(v)} + \sigma_v^2 \boldsymbol{I}), \qquad p_r(\boldsymbol{X}|\boldsymbol{Y}) = \mathcal{N}(\boldsymbol{0}, \boldsymbol{K}_r + \sigma_r^2 \boldsymbol{I}), \quad (3)$$

where $\boldsymbol{K}^{(v)} = k^{(v)}(\boldsymbol{X}, \boldsymbol{X})$ and $\boldsymbol{K}_r = \sum_v k_r^{(v)}(\boldsymbol{Y}^{(v)}, \boldsymbol{Y}^{(v)})$ are the kernels associated with each process. Note that in the recognition model the relevant kernel allows us to easily combine multiple features via the sum of the individual kernel functions. Training of the recognition model consists of maximizing the conditional $p_r(\boldsymbol{X}|\boldsymbol{Y})$ w.r.t. the kernel hyper-parameters $\boldsymbol{\theta}_r$. For the generative model we maximize the marginal likelihood (labels $\boldsymbol{Z}$ are omitted here)

$$p(\boldsymbol{Y}) = \int \prod_{v=1}^V p(\boldsymbol{Y}^{(v)}|\boldsymbol{X}) p(\boldsymbol{X}) d\boldsymbol{X}. \quad (4)$$

Since the above integral is intractable, we resort to approximations. Our main interest is to recover a Bayesian non-parametric solution for both the GP encoder and decoder. We first need to break the circular dependence between $\boldsymbol{Y}^{(v)}$ and $\boldsymbol{X}$ in order to train the two GPs simultaneously.

GP-encoder. We decouple $\boldsymbol{X}$ and $\boldsymbol{Y}$ by introducing an intermediate variable $\boldsymbol{M} = \{\boldsymbol{m}_i\}_{i=1}^N$, so that the recognition model becomes $\boldsymbol{y}^{(v)} \to \boldsymbol{m} \to \boldsymbol{x}$. The GP operates on $\boldsymbol{y}^{(v)}, \boldsymbol{m}$, while $\boldsymbol{x}$ is the noisy observations of $\boldsymbol{m}$. This process is described in Fig. 1(c). We follow a mean field approximation and introduce

---

[1] The subscript r indicates that the process facilitates the recognition model.

the variational distribution $q(\boldsymbol{X}|\boldsymbol{M}) = \prod_i q_i(\boldsymbol{x}_i|\boldsymbol{m}_i) = \prod_i \mathcal{N}(\boldsymbol{m}_i, \boldsymbol{S}_i)$. Here, $\boldsymbol{m}_i, \boldsymbol{S}_i \in \mathbb{R}^q$ are variational parameters[2] of q_i. We define $\boldsymbol{M}$ by employing the cavity distribution of the leave-one-out solution of GP [21]

$$p(\boldsymbol{M}|\boldsymbol{Y}) = \prod_i p(\boldsymbol{m}_i|\boldsymbol{Y}, \boldsymbol{M}_{\setminus i}) = \prod_i \mathcal{N}(\hat{\boldsymbol{m}}_i, \hat{\sigma}_i^2 \boldsymbol{I}), \tag{5}$$

where the subscript $\setminus i$ means 'all datapoints except i', and the mean and variance of the Gaussian are given by [21]

$$\hat{\boldsymbol{m}}_i = \boldsymbol{m}_i - \left[\boldsymbol{K}_r^{-1}\boldsymbol{M}\right]_i / \left[\boldsymbol{K}_r^{-1}\right]_{ii}, \qquad \hat{\sigma}_i^2 = 1/\left[\boldsymbol{K}_r^{-1}\right]_{ii}. \tag{6}$$

We now integrate out the intermediate layer and propagate the uncertainty of the GP mapping to the latent variable $\boldsymbol{X}$, which yields the variational distribution

$$q(\boldsymbol{X}|\boldsymbol{Y}) = \prod_i \mathcal{N}(\hat{\boldsymbol{m}}_i, \boldsymbol{S}_i + \hat{\sigma}_i^2 \boldsymbol{I}). \tag{7}$$

GP-decoder. The proposed recognition model, *i.e.*, the variational distribution of Eq. (7), can be employed to approximate the intractable marginal likelihood of Eq. (4). By introducing the variational distribution as an approximation to the true posterior, and after applying the Jensen's inequality, we obtain the lower bound to the log-marginal likelihood (again, labels $\boldsymbol{Z}$ are omitted)

$$\log p(\boldsymbol{Y}) \geq \mathcal{F}_1 = \sum_v \mathbb{E}_{q(\boldsymbol{X}|\boldsymbol{Y})}\left[\log p(\boldsymbol{Y}^{(v)}|\boldsymbol{X})\right] - KL(q(\boldsymbol{X}|\boldsymbol{Y})||p(\boldsymbol{X})). \tag{8}$$

Training our model consists of maximizing the lower bound of Eq. (8) w.r.t. the variational parameters $\boldsymbol{M}, \boldsymbol{S}$ and the hyper-parameters of the kernels $\boldsymbol{K}^{(v)}, \boldsymbol{K}_r$. Further details are given in Sec. 3.4.

3.3   Incorporating Ordinal Variables

In the previous section, we presented the recognition model that we employ to learn a nonlinear manifold from the observed inputs. In the following, we further constrain this manifold by imposing an ordinal structure. This is attained by introducing ordinal variables that account for C ordinal levels of AUs. We use the notion of ordinal regression [24] and, in particular, the ordinal threshold model that imposes the monotonically increasing structure of the discrete output labels to the continuous manifold. Formally, the non-linear mapping between the manifold $\boldsymbol{X}$ and the ordinal outputs $\boldsymbol{Z}$ is modeled as

$$p(\boldsymbol{Z}|g(\boldsymbol{X})) = \prod_{i,c} p(z_{ic}|g_c(\boldsymbol{x}_i)), \; p(z_{ic} = s|g_c(\boldsymbol{x}_i)) = \begin{cases} 1 & \text{if } g_c(\boldsymbol{x}_i) \in (\gamma_{c,s-1}, \gamma_{c,s}] \\ 0 & \text{otherwise,} \end{cases} \tag{9}$$

where $i = 1, \ldots, N$ indexes the training data. $\gamma_{c,0} = -\infty \leq \cdots \leq \gamma_{c,S} = +\infty$ are the thresholds or cut-off points that partition the real line into $s = 1, \ldots, S$ contiguous intervals. These intervals map the real function value $g_c(\boldsymbol{x})$ into the

---

[2] For simplicity we assume an isotropic (diagonal) covariance across the dimensions.

discrete variable s, corresponding to each of S intensity levels of an AU, while enforcing the ordinal constraints. The threshold model $p(z_{ic} = s|g_c(\boldsymbol{x}_i))$ is used for ideally noise-free cases. Here, we assume that the latent functions $g_c(\cdot)^3$ are corrupted by Gaussian noise, leading to the following formulation

$$g_c(\boldsymbol{x}_i) = \boldsymbol{w}_c^T \boldsymbol{x}_i + \epsilon_g, \quad \epsilon_g \sim \mathcal{N}(0, \sigma_g^2). \tag{10}$$

By integrating out the noisy projections from Eq. (9) (see [32] for details), we arrive at the ordinal log-likelihood

$$\log p(\boldsymbol{Z}|\boldsymbol{X}, \boldsymbol{W}) = \sum_{i,c} \mathbb{I}(z_{ic} = s) \log \left(\varPhi \left(\frac{\gamma_{c,s} - \boldsymbol{w}_c^T \boldsymbol{x}_i}{\sigma_g} \right) - \varPhi \left(\frac{\gamma_{c,s-1} - \boldsymbol{w}_c^T \boldsymbol{x}_i}{\sigma_g} \right) \right), \tag{11}$$

where $\varPhi(\cdot)$ is the Gaussian cumulative density function, and $\mathbb{I}(\cdot)$ is the indicator function. Finally, by using the ordinal likelihood defined in Eq. (11), we obtain the final lower bound of our log-marginal likelihood

$$\log p(\boldsymbol{Y}, \boldsymbol{Z}|\boldsymbol{W}) \geq \mathcal{F}_2 = \sum_v \mathbb{E}_{q(\boldsymbol{X}|\boldsymbol{Y})} \big[\log p(\boldsymbol{Y}^{(v)}|\boldsymbol{X}) \big] - KL(q(\boldsymbol{X}|\boldsymbol{Y})||p(\boldsymbol{X}))$$
$$+ \sum_{i,c} \mathbb{I}(z_{ic} = s) \mathbb{E}_{q(\boldsymbol{X}|\boldsymbol{Y})} \left[\log \left(\varPhi \left(\frac{\gamma_{c,s} - \boldsymbol{w}_c^T \boldsymbol{x}_i}{\sigma_g} \right) - \varPhi \left(\frac{\gamma_{c,s-1} - \boldsymbol{w}_c^T \boldsymbol{x}_i}{\sigma_g} \right) \right) \right]. \tag{12}$$

3.4   Learning and Inference

Training our model consists of maximizing the lower bound of Eq. (12) w.r.t. the variational parameters $\{\boldsymbol{S}, \boldsymbol{M}\}$, the hyper-parameters $\{\boldsymbol{\theta}^{(v)}, \sigma_v, \boldsymbol{\theta}_r^{(v)}, \sigma_r\}$ of the GP mappings, and the parameters $\{\boldsymbol{W}, \gamma, \sigma_g\}$ of the ordinal classifier. For the kernel of the GP-decoder we use the radial basis function (RBF) with automatic relevance determination (ARD), which can effectively estimate the dimensionality of the latent space [18]. For the kernel of the GP-encoder we use the isotropic RBF for each observed input. To utilize a joint optimization scheme, we use stochastic backpropagation [30,31], where the re-parameterization trick is applied in Eq. (12). Thus, we can obtain the Monte Carlo estimate of the expectation of the GP auto-encoder from

$$\mathbb{E}_{q(\boldsymbol{X}|\boldsymbol{Y})} \left[\log p(\boldsymbol{Y}^{(v)}|\boldsymbol{X}) \right] = \sum_i \mathbb{E}_{\mathcal{N}(\boldsymbol{\xi}|\boldsymbol{0}, \boldsymbol{I})} \left[\log p(\boldsymbol{y}_i^{(v)}|\hat{\boldsymbol{m}}_i + (\boldsymbol{S}_i^{1/2} + \hat{\sigma}_i \boldsymbol{I})\boldsymbol{\xi}) \right]. \tag{13}$$

The expectation of the ordinal classifier is computed in a similar manner. The advantage of Eq. (13) is twofold: (i) It allows for an efficient computation of the lower bound even when using arbitrary kernel functions (in contrast to [18]); (ii) It provides an efficient, low-variance estimator of the gradient [30]. The extra approximation (via the expectation) in the gradient step requires stochastic gradient descent. We use AdaDelta [33] for this purpose.

Inference in the proposed method is straightforward: The test data $\boldsymbol{y}_*^{(v)}$, are first projected onto the manifold using the trained GP-encoder. In the second step, we apply the ordinal classifier to the obtained latent position.

---

[3] Note that we adopt here a linear model for $g_c(\cdot)$ as it operates on a low-dimensional non-linear manifold $\boldsymbol{X}$, already obtained by the GP auto-encoder.

3.5   Relation to Prior Work on Gaussian Processes

Our auto-encoder approach is inspired by neural-network counterparts proposed in [30,31], where probabilistic distributions are defined for the input and output mapping functions. In the GP literature, auto-encoders are closely related to the notion of 'back-constraints'. Back-constraints were introduced in [34] as a deterministic, parametric mapping (commonly a multi-layer perceptron (MLP)) that pairs the latent variables of the GPLVM [35] with the observations. This mapping facilitates a fast inference mechanism and enforces structure preservation in the manifold. The same mechanism has been used to constrain the shared GPLVM [36], from one view in [37] and multiple views in [38].

Back-constraints have been recently introduced to the B-GPLVM [22]. In [39] the authors proposed to approximate the true posterior of the latent space by introducing a variational distribution conditioned on some unobserved inputs. However, those inputs are not related to the observation space considered in this paper (*i.e.*, the outputs Y of the GPLVM). In [23] the variational posterior of the latent space is constrained by using the trick of the parametric deterministic mapping from [34]. Finally, in [28], the authors replaced the variational approximation with a Monte Carlo expectation-maximization algorithm. Samples were obtained from the GP mapping from the observed inputs to the manifold.

Our proposed VGP-AE advances the current literature in many aspects: (1) We introduce a GP mapping for our recognition model. Hence, can model different uncertainty levels per input, which allows us to learn more confident latent representations. (2) The use of the non-parametric GPs also allows us to model complex structures at a lesser expense than the MLP (fewer parameters). Thus, it is less prone to overfitting and scales better to high-dimensional data. (3) Compared to [39] our probabilistic recognition model facilitates a low-dimensional projection of our observed features, while the variational constraint in [39] does not constitute a probabilistic mapping. (4) We learn the GP encoders/decoders in a joint optimization, while [28] train the two models in an alternating scheme.

4   Experiments

We empirically assess the structure learning abilities of the proposed VGP-AE as well as its efficacy when dealing with data of ordinal nature.

4.1   Experimental Protocol

Datasets. We first show the qualitative evaluation of the proposed VGP-AE on the MNIST [40] benchmark dataset of images of handwritten digits. We use it to assess the properties of the auto-endoced manifold. We then show the performance of VGP-AE on two benchmark datasets of facial affect: DISFA [6], and BP4D [41] (using the publicly available data subset from the FERA2015 [8] challenge). Specifically, DISFA contains video recordings of 27 subjects while watching YouTube videos. Each frame is coded in terms of the intensity of 12 AUs, on a six-point ordinal scale. The FERA2015 database includes video of 41 participants. There are 21 subjects in the training and 20 subjects in the

development partition. The dataset contains intensity annotations for 5 AUs.

Features. In the experiment on MNIST dataset, we use the normalized raw pixel intensities as input, resulting in a 784D feature vector. For DISFA and FERA2015, we use both geometric and appearance features. Specifically, DISFA and FERA2015 datasets come with frame-by-frame annotations of 66 and 49 facial landmarks, respectively. After removing the contour landmarks from DISFA annotations, we end up with the same set of 49 facial points. We register the images to a reference face using an affine transform based on these points. We then extract Local Binary Patterns (LBP) histograms [42] with 59 bins from patches centered around each registered point. Hence, we obtain 98D (geometric) and 2891D (appearance) feature vectors, commonly used in modeling of facial affect.

Evaluation. As evaluation measures, we use the negative log-predictive density (NLPD) to assess the generative ability (reconstruction part) of our model. For the task of ordinal classification, we report the mean squared error (MSE) and the intra-class correlation (ICC(3,1)) [43]. These are the standard measures for ordinal data. The MSE measures the classifier's consistency regarding the relative order of the classes. ICC is a measure of agreement between annotators (in our case, the ground truth of the AU intensity and the model's predictions). Finally, we adopt the subject-independent setting: for FERA2015 we report the results on the subjects of the development set, while for DISFA we perform a 9-fold (3 subjects per fold) cross-validation procedure.

Models. We compare the proposed VGP-AE to the state of the art GP manifold learning methods that perform multi-input multi-output inference. These include: (i) manifold relevance determination (MRD) [18], a regression model based on variational inference, (ii) variational auto-encoded deep GP (VAE-DGP) [23], which uses a recognition model based on an MLP to constrain the learning of MRD, and (iii) multi-task latent GP (MT-LGP) [19], which uses the same MLP-based recognition model and a maximum likelihood learning approach. We also compare to the variational GP for ordinal regression (vGPOR) [44]. As a baseline, we use the standard GP [21] with a shared covariance function among the multi-outputs. We also compare to the single-output ordinal threshold model (SOR) [24]. Finally, we compare to state of the art methods for joint estimation of AU intensity based on MRFs [14] and latent trees (LT) [15], respectively. For the single input (no fusion) methods (GP, vGPOR, SOR, LT, MRF), we concatenate the two feature sets. The parameters of each method were tuned as described in the corresponding papers. For the GP subspace methods, we used the RBF kernel with ARD, and initialized with the 20D manifold. For the GP regression methods, we used the standard RBF. For the sparse variational GP methods (vGPOR, MRD, VAE-DGP) we used 200 inducing points, and 20 hidden units for the MLP in the recognition models of VAE-DGP and MT-LGP.

4.2   Assessing the Recognition Model

In the following, we qualitatively assess the benefits of the proposed recognition model in the task of manifold recovery from the MNIST dataset. We select an
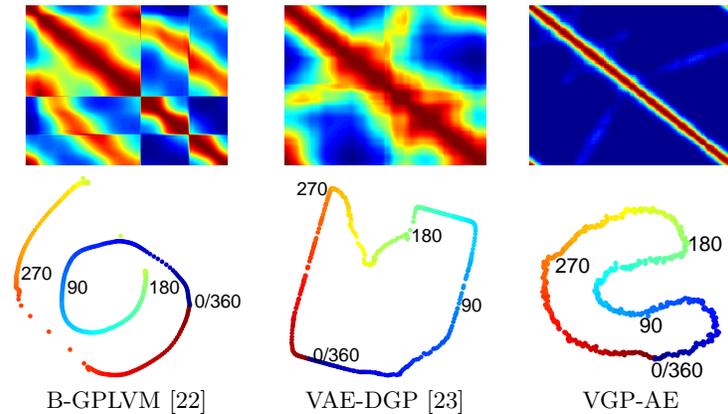
Fig. 2. Recovering the structure of a rotated '1' from MNIST. The learned kernel matrices (upper row) and 2D manifolds (lower row) obtained from B-GPLVM (left), VAE-DGP (middle) and the proposed VGP-AE (right), initialized from the same random instance.

image depicting the digit '1' and rotate it around $360°$. This results in a set of images of '1's rotated at a step of $1°$. Our goal is to infer the true structure of the data, for which we know *a priori* that it should correspond to a diagonal-like kernel and a circular manifold. However, the challenge arises from the symmetry of digit '1', which is almost identical at opposite degrees (*e.g.*, $0°$ and $180°$). The results are depicted in Fig. 2. Note that since we do not deal with the classification task we exclude the ordinal component in VGP-AE. We compare the learned manifold structure to the B-GPLVM [22], which does not model the back-projection to the latent space, and a single layer VAE-DGP, where the back-projections are modeled using MLP. In Fig. 2(upper row), we see from the learned kernels that the B-GPLVM is unable to fully unravel the dissimilarity between the 'inverted' images, resulting also in a non-smooth kernel with a discontinuity at $180°$ and $270°$. By contrast, the VAE-DGP benefits from the recognition model and manages to resolve this to some extent. Yet, the recovered kernel still suffers from a discontinuity around $180°$. On the other hand, the proposed VGP-AE, by using the more general recognition model based on GPs (infinitely wide MLP), succeeds to accurately discover the true underlying manifold, also resulting in a more smooth, almost ideal kernel. These observations are further supported by the instances of the learned 2D manifolds in Fig. 2(lower row). B-GPLVM learns a disconnected manifold with 'jumps' at $180°$ and $270°$. However, both the VAE-DGP and proposed VGP-AE recover a circular manifold, with the manifold recovered by VGP-AE being more symmetric.

4.3   Convergence Analysis

We next demonstrate the convergence of VGP-AE in the task of AU intensity estimation on FERA2015. Fig. 3(a) shows the effect of learning the ordinal clas-
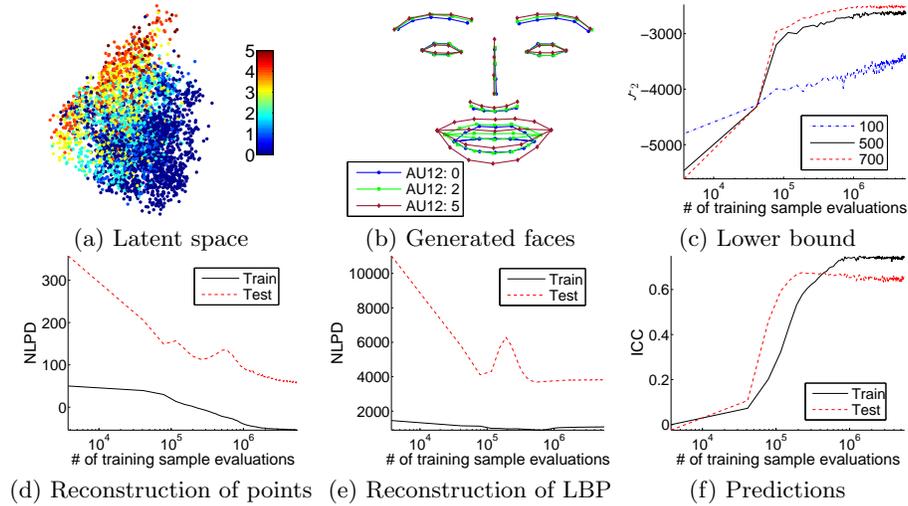
Fig. 3. Convergence analysis of the proposed method on FERA2015. (a) The recovered latent space with ordinal information from AU12, and (b) reconstructed face shapes sampled from different regions of the manifold. (c) the estimated average variational lower bound, $\mathcal{F}_2$, per datapoint, for different batch sizes. The model's reconstruction capacity for the points (d) and LBP (e) features, measured by the NLPD. (f) the average ICC for the joint AU intensity estimation. The horizontal axis corresponds to the amount of training points evaluated after 1500 epochs of the stochastic optimization.

sifier and the auto-encoded manifold within the joint optimization framework. It can be clearly seen from the recovered space that the information from the labels has been correctly encoded in the manifold, which now has an ordinal structure (the depicted coloring accounts for the 'ordinality' of AU12). As depicted in Fig. 3(b), we can accurately reconstruct face shapes with different AU intensities, by sampling from different regions of the space. Fig. 3(c) shows the convergence of the proposed method when optimizing the lower bound $\mathcal{F}_2$ of Eq. (12) for different batch sizes of the stochastic optimization. With a small batch size (100 datapoints) the model cannot estimate the structure of the inputs well. Hence, it approximates the log-marginal likelihood less accurately. By increasing the batch size to 500, the model converges to a better solution and optimization becomes more stable since the curve becomes smoother over the iterations. Further increase of the batch size does not have a considerable effect.

In Fig. 3(d)–(e) we evaluate the generative part of the auto-encoder by measuring the model's ability to reconstruct both input features (points and LBPs) in terms of NLPD. First of all, it is clear that our Bayesian training prevents the model from overfitting, since the NLPD of the test data follows the trend of the training data. Furthermore, we can see that the model can reconstruct the geometric features better than the appearance, which is evidenced by the lower NLPD (around -50 for points and 1500 for LBPs). We partly attribute this

Table 1. Joint AU intensity estimation on DISFA and FERA2015.

	Dataset	DISFA													FERA2015					
	AU	1	2	4	5	6	9	12	15	17	20	25	26	Avg.	6	10	12	14	17	Avg.
ICC	VGP-AE	.48	.47	.62	.19	**.50**	**.42**	**.80**	.19	**.36**	.15	**.84**	.53	**.46**	**.75**	**.66**	**.88**	.47	**.49**	**.65**
	VAE-DGP [23]	.39	.34	.46	.13	.40	.31	.75	.14	.23	.14	.75	.45	.38	.72	.61	.82	.40	.38	.59
	MRD [18]	.46	.39	.43	.09	.28	.34	.71	.09	.30	.09	.73	.36	.36	.68	.59	.80	.38	.38	.57
	MT-LGP [19]	.41	.33	.28	.10	.23	.22	.56	.13	.26	.18	.65	.23	.30	.67	.61	.80	.37	.41	.57
	vGPOR [44]	**.53**	**.49**	.54	.21	.35	.40	.75	.18	.30	.16	.79	.39	.42	.74	.62	.84	**.48**	.35	.61
	GP [21]	.28	.13	.42	.03	.13	.23	.62	.08	.26	.19	.67	.23	.27	.69	.58	.81	.35	.38	.56
	SOR [24]	.25	.18	**.65**	.08	.46	.15	.77	.14	.24	.04	.82	**.57**	.36	.61	.50	.77	.28	.45	.52
	LT [15]	.28	.26	.44	.24	.50	.13	.69	.06	.21	.06	.62	.37	.32	.70	.59	.76	.30	.31	.53
	MRF [14]	.46	.38	.50	**.37**	.41	.34	.67	**.32**	.29	**.20**	.69	.46	.42	.64	.53	.79	.34	.46	.55
MSE	VGP-AE	.51	**.32**	1.13	.08	.56	.31	.47	.20	**.28**	.16	**.49**	.44	.41	**.82**	**1.28**	**.70**	**1.43**	**.77**	**1.00**
	VAE-DGP [23]	.40	.36	.95	.08	.48	.29	**.43**	.19	.32	.16	.76	**.44**	.41	.91	1.33	.81	1.46	.86	1.07
	MRD [18]	.42	.38	1.31	.08	.56	**.27**	.47	.20	.36	.18	.82	.53	.46	1.00	1.39	.83	1.64	.88	1.15
	MT-LGP [19]	.40	.35	1.25	.08	.60	.30	.73	.18	.36	.16	1.19	.67	.52	.97	1.31	.81	1.58	.84	1.10
	vGPOR [44]	.38	.34	.95	**.06**	.57	**.27**	**.43**	.18	.33	.18	.65	.53	.41	1.00	1.54	.76	1.78	1.11	1.24
	GP [21]	.52	.51	1.13	.13	.65	.36	.61	.23	.38	.20	.94	.66	.53	.94	1.40	.76	1.62	.88	1.12
	SOR [24]	.47	.40	1.13	.07	.63	.37	.55	.21	.35	.21	.71	.61	.48	1.44	1.82	1.08	2.58	1.01	1.59
	LT [15]	.44	.38	**.93**	**.06**	**.36**	.32	.46	.16	.29	**.15**	.97	**.44**	.41	.89	1.33	.91	1.48	.85	1.09
	MRF [14]	**.37**	.35	.94	**.06**	.45	.29	.46	**.13**	.32	.16	.77	**.44**	**.40**	1.20	1.66	.86	2.19	.92	1.37

to the fact that the LBPs are of higher dimension and therefore more difficult to reconstruct. Another reason for this difference is that the model learns to reconstruct the part of the features that enclose the more relevant information regarding the task of classification. The latter is further supported by Fig. 3(e), where we see the progress of the average ICC during the optimization. In the beginning, the model has no information since the latent space is initialized randomly. As we progress the model fuses the information of the input features in the latent space and unravels the structure of the data. Thus, ICC starts rising and reaches its highest value, .65 on the test data. After that point the model does no longer benefit from the appearance features: it has reached the plateau.

4.4   Model Comparisons on Spontaneous Data of Facial Expressions

We compare the proposed approach to several methods on the spontaneous data from the DISFA and FERA2015 datasets. Table 1 summarizes the results. First, we observe that all methods perform significantly better (in terms of ICC) on the data from FERA2015 than on DISFA. This is mainly due to the fact that FERA2015 contains a much more balanced set of AUs (in terms of activations), and hence, all models (single- and multi-output) can learn the classifiers for the target task better. Furthermore, our proposed approach performs significantly better than the compared GP manifold learning methods, which treat the output labels as continuous variables. MRD lacks the modeling of back-projections. This results in learning a less smooth manifold of facial expressions, which affects its representation abilities, and hence, its predictions. On the other hand, the VAE-DGP learns explicitly the mapping from the observed features to the
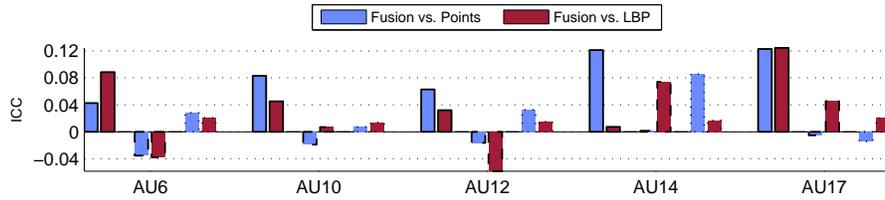
Fig. 4. Demonstration of the gain/loss from feature fusion for joint AU intensity estimation on FERA2015. Within each AU the first tuple (solid line) corresponds to the proposed VGP-AE, the second tuple (dashed line) to the VAE-DGP [23], and the third tuple (dotted line) to the vGPOR [44].

latent space in a deterministic and parametric fashion. Although this strategy is proven to be superior to unconstrained learning, it can be severely affected in cases where we have access to noisy and high-dimensional features. MT-LGP also models the back-mappings. However, it reports worse results, especially on DISFA. This drop in the performance is accounted to the non-Bayesian learning of the manifold, which constitutes the model more prone to overfitting.

Regarding the sparse ordinal regression instance of GPs, *i.e.*, vGPOR, we see that it manages to learn relatively accurate mappings between features and labels, and thus, performs close to our proposed method. However, it reports worse results since it cannot achieve the desirable fusion of the features without learning an intermediate latent space. The baseline methods, *i.e.*, GP and SOR, report lower results. The GP attains low scores due to handling the ordinal outputs in a continuous manner while the ordinal modeling helps SOR to report consistently better.

Finally, the proposed approach significantly outperforms the state of the art methods in the literature of AU intensity estimation, *i.e.*, LT and MRF. LT learns the label information in a generative manner, and treats them as extra feature dimensions. Although this approach can be beneficial in the presence of noisy features [15], it suffers from learning complicated and large tree structures when falsely detecting connections between features and AUs. Hence, it performs worse. The MRF performs on par to the proposed method on DISFA and achieves the best average MSE, but it is consistently worse on FERA2015. This inconsistency is due to its two-step learning strategy, which results in unraveling a graph that cannot explain simultaneously all different features and AUs.

In Fig. 4 we evaluate the attained fusion between the best performing methods on FERA2015, *i.e.*, the proposed VGP-AE, VAE-DGP [23] and vGPOR [44]. As we can see, the proposed approach (solid line, first tuple) manages to accurately fuse the information from the two input features in the learned manifold. Thus, it achieves higher ICC on all AUs compared to when the two modalities are used individually as input features. On the other hand, although vGPOR (third tuple, dotted line) reports also high ICC scores, it does not benefit from the presence of the two features: In most cases it cannot achieve a significant
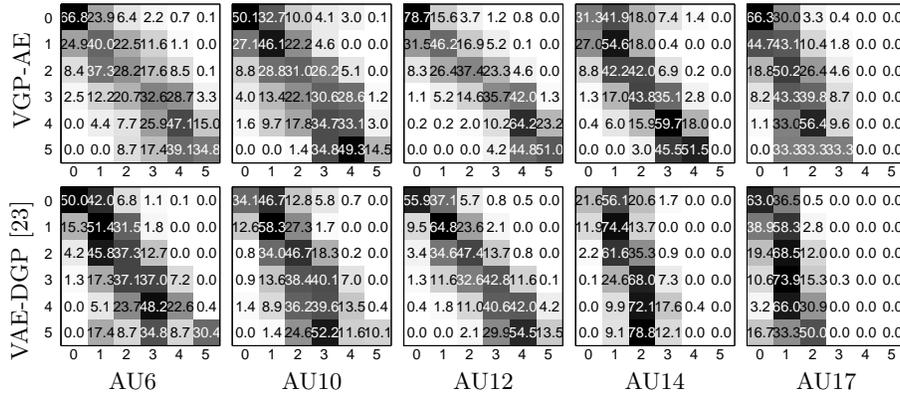
Fig. 5. Confusion matrices for predicting the $0-5$ intensity of all AUs on FERA2015, when performing fusion with VGP-AE (upper row) and VAE-DGP [23] (lower row).

increase compared to the individual inputs. Finally, VAE-DGP (middle tuple, dashed line) consistently attains better performance on all AUs with a single feature as input. This can be attributed to modeling the recognition model via the parametric MLP. The latter affects the learning of the manifold, especially when dealing with the high-dimensional noisy appearance features.

The above mentioned difference between our approach and the VAE-DGP is further evidenced in Fig. 5. The proposed fusion along with the novel non-parametric, probabilistic recognition model in our auto-encoder leads to less confusion between the ordinal states across all AUs. We further attribute this to the ordinal modeling of outputs in our VGP-AE, contrary to VAE-DGP that treats the output as continuous variables. This is especially pronounced in the case of the subtle AUs 14&17, where examples of high intensity levels are scarce.

5   Conclusion

We have presented a fully probabilistic auto-encoder, where GP mappings govern both the generative and the recognition models. The proposed variational GP auto-encoder is learned in a supervised manner, where the ordinal nature of the labels is imposed to the manifold. This allows the proposed approach to accurately learn the structure of the input data, while also remain competitive in the classification task. We have empirically evaluated our model on the task of facial feature fusion for joint intensity estimation of facial action units. The proposed model outperforms related GP methods and the state of the art approaches for the target task.

References

1. Bartlett, M., Whitehill, J.: Automated facial expression measurement: Recent applications to basic research in human behavior, learning, and education. Handbook of Face Perception. Oxford University Press, USA (2010)
2. Ekman, P., Friesen, W.V., Hager, J.C.: Facial action coding system. A Human Face, Salt Lake City, UT (2002)
3. Pantic, M.: Machine analysis of facial behaviour: Naturalistic and dynamic behaviour. Philosophical Transactions of the Royal Society B: Biological Sciences **364** (2009) 3505–3513
4. Rudovic, O., Pavlovic, V., Pantic, M.: Context-sensitive dynamic ordinal regression for intensity estimation of facial action units. IEEE TPAMI **37** (2015) 944–958
5. Mahoor, M.H., Cadavid, S., Messinger, D.S., Cohn, J.F.: A framework for automated measurement of the intensity of non-posed facial action units. In: IEEE CVPR-W. (2009) 74–80
6. Mavadati, S.M., Mahoor, M.H., Bartlett, K., Trinh, P., Cohn, J.F.: DISFA: A spontaneous facial action intensity database. IEEE TAC **4** (2013) 151–160
7. Ming, Z., Bugeau, A., Rouas, J.L., Shochi, T.: Facial action units intensity estimation by the fusion of features with multi-kernel support vector machine. In: IEEE FG. Volume 6. (2015) 1–6
8. Valstar, M.F., Almaev, T., Girard, J.M., McKeown, G., Mehu, M., Yin, L., Pantic, M., Cohn, J.F.: FERA 2015 - second facial expression recognition and analysis challenge. In: IEEE FG. Volume 6. (2015) 1–8
9. Savran, A., Sankur, B., Bilge, M.T.: Regression-based intensity estimation of facial action units. Image and Vision Computing **30** (2012) 774–784
10. Kaltwang, S., Rudovic, O., Pantic, M.: Continuous pain intensity estimation from facial expressions. In: Advances in Visual Computing. (2012) 368–377
11. Jeni, L.A., Girard, J.M., Cohn, J.F., De La Torre, F.: Continuous AU intensity estimation using localized, sparse facial feature space. In: IEEE FG. (2013) 1–7
12. Kaltwang, S., Todorovic, S., Pantic, M.: Doubly sparse relevance vector machine for continuous facial behavior estimation. IEEE TPAMI **38** (2015) 1748–1761
13. Li, Y., Mavadati, S.M., Mahoor, M.H., Ji, Q.: A unified probabilistic framework for measuring the intensity of spontaneous facial action units. In: IEEE FG. (2013)
14. Sandbach, G., Zafeiriou, S., Pantic, M.: Markov random field structures for facial action unit intensity estimation. In: IEEE ICCV-W. (2013) 738–745
15. Kaltwang, S., Todorovic, S., Pantic, M.: Latent trees for estimating intensity of facial action units. In: IEEE CVPR. (2015) 296–304
16. Nicolle, J., Bailly, K., Chetouani, M.: Facial action unit intensity prediction via hard multi-task metric learning for kernel regression. In: IEEE FG. (2015) 1–6
17. Mohammadi, M.R., Fatemizadeh, E., Mahoor, M.H.: Intensity estimation of spontaneous facial action units based on their sparsity properties. IEEE TCYB **46** (2016) 817–826
18. Damianou, A., Ek, C.H., Titsias, M., Lawrence, N.: Manifold relevance determination. In: ICML. (2012) 145–152
19. Urtasun, R., Quattoni, A., Lawrence, N., Darrell, T.: Transferring nonlinear representations using Gaussian processes with a shared latent space. Technical Report MIT-CSAIL-TR-08-020 (2008)
20. Calandra, R., Peters, J., Rasmussen, C.E., Deisenroth, M.P.: Manifold Gaussian processes for regression. In: IJCNN. (2016)

21. Rasmussen, C., Williams, C.: Gaussian processes for machine learning. Volume 1. MIT Press Cambridge, MA (2006)
22. Titsias, M.K., Lawrence, N.D.: Bayesian Gaussian process latent variable model. In: AISTATS. (2010) 844–851
23. Dai, Z., Damianou, A., González, J., Lawrence, N.: Variational auto-encoded deep Gaussian processes. In: ICLR. (2016)
24. Agresti, A.: Analysis of ordinal categorical data. John Wiley & Sons (2010)
25. Mahoor, M.H., Zhou, M., Veon, K.L., Mavadati, S.M., Cohn, J.F.: Facial action unit recognition with sparse representation. In: IEEE FG. (2011) 336–342
26. Chu, W.S., Torre, F.D.L., Cohn, J.F.: Selective transfer machine for personalized facial action unit detection. In: IEEE CVPR. (2013) 3515–3522
27. Zhao, K., Chu, W.S., De la Torre, F., Cohn, J.F., Zhang, H.: Joint patch and multi-label learning for facial action unit detection. In: IEEE CVPR. (2015)
28. Eleftheriadis, S., Rudovic, O., Pantic, M.: Multi-conditional latent variable model for joint facial action unit detection. In: IEEE ICCV. (2015) 3792–3800
29. Candès, E.J., Li, X., Ma, Y., Wright, J.: Robust principal component analysis? Journal of the ACM **58** (2011)
30. Kingma, D.P., Welling, M.: Auto-encoding variational Bayes. In: ICLR. (2013)
31. Rezende, D.J., Mohamed, S., Wierstra, D.: Stochastic backpropagation and approximate inference in deep generative models. In: ICML. (2014) 1278–1286
32. Chu, W., Ghahramani, Z.: Gaussian processes for ordinal regression. In: JMLR. (2005) 1019–1041
33. Zeiler, M.D.: ADADELTA: an adaptive learning rate method. arXiv preprint arXiv:1212.5701 (2012)
34. Lawrence, N.D., Candela, J.Q.: Local distance preservation in the GP-LVM through back constraints. In: ICML. Volume 148. (2006) 513–520
35. Lawrence, N.: Probabilistic non-linear principal component analysis with Gaussian process latent variable models. JMLR **6** (2005) 1783–1816
36. Shon, A., Grochow, K., Hertzmann, A., Rao, R.: Learning shared latent structure for image synthesis and robotic imitation. NIPS **18** (2006) 1233–1240
37. Ek, C.H., Torr, P.H., Lawrence, N.D.: Gaussian process latent variable models for human pose estimation. In: MLMI. Springer (2008) 132–143
38. Eleftheriadis, S., Rudovic, O., Pantic, M.: Discriminative shared Gaussian processes for multiview and view-invariant facial expression recognition. IEEE TIP **24** (2015) 189–204
39. Damianou, A., Lawrence, N.: Semi-described and semi-supervised learning with Gaussian processes. In: UAI. (2015)
40. LeCun, Y., Cortes, C., Burges, C.J.: The MNIST database of handwritten digits (1998)
41. Zhang, X., Yin, L., Cohn, J.F., Canavan, S., Reale, M., Horowitz, A., Liu, P., Girard, J.M.: BP4D-spontaneous: a high-resolution spontaneous 3d dynamic facial expression database. Image and Vision Computing **32** (2014) 692–706
42. Ojala, T., Pietikainen, M., Maenpaa, T.: Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. IEEE TPAMI **24** (2002) 971–987
43. Shrout, P.E., Fleiss, J.L.: Intraclass correlations: uses in assessing rater reliability. Psychological bulletin **86** (1979) 420
44. Sheth, R., Wang, Y., Khardon, R.: Sparse variational inference for generalized GP models. In: ICML. (2015) 1302–1311