Universitá degli Studi di Milano

Scuola di Dottorato in Scienze Matematiche

PhD School in Mathematical Sciences

Functional statistical learning methods applied to human emotion recognition from facial videos

by

Rongjiao Ji

Supervisor: **Alessandra Micheletti**
Co-Supervisor: **Nataša Krklec Jerinkić**
Industrial Co-Supervisor: **Zoranka Desnica**

August 2022

Abstract

The ever-growing fascination with automatically analyzing and understanding human behavior has inspired a profound focus on the evolution of facial expressions and the recognition of corresponding emotions. By harnessing functional statistical learning methods, we develop a comprehensive methodology that capitalizes on the dynamic properties of continuity and evolvability inherent in functional data extracted from facial videos, which possess distinct properties compared to the static facial images predominantly used in traditional research methods. Our approach employs multivariate function-on-scalar regression models and functional analysis of variance (FANOVA) to effectively separate shared information from group-specific influences and individual noise through paired group comparisons, even with limited sample sizes. The identified group patterns convey significant mean characteristics in grouped units and are further utilized as prior knowledge for multi-classification in a streamlined feature space, generating emotional agreement scores for incoming new samples. Both non-parametric and parametric multi-class classification methods are employed to assess the predictive capabilities of the multivariates. In summary, we seamlessly integrate the entire pipeline for various stages of training and testing processes within the domain of explainable automatic emotion recognition, unveiling compelling results and offering insightful interpretations that may shed new light on emotions and expressions.

Contents

Introduction

The increasing interest in the automatic analysis and understanding of human behavior has prompted researchers to concentrate on the development of facial expressions and the identification of related emotions. Traditional studies often analyze human emotions using extensive collections of static facial images, overlooking the time-dependent characteristics of continuity and changeability present in video data with limited sample sizes. Machine and deep learning techniques have shown great effectiveness in solving these complex problems; however, they have some drawbacks. The training of these models generally relies on very large datasets, as machine learning models often require thousands of parameters to be identified or tuned. Additionally, when using neural networks with multiple layers for artificial intelligence systems, the resulting models can become "black-box" learning systems, heavily dependent on the training set and difficult to interpret even for their designers.

Our approach to human emotion recognition aligns with the concept of Explainable Artificial Intelligence (XAI). We aim to investigate statistical learning methods that produce easily interpretable results, can be applied to new datasets, and could potentially contribute to the development of increasingly realistic virtual humans. Furthermore, our models should have a small number of parameters, enabling them to be trained even with limited datasets. This feature is crucial for industrial applications, as a virtual human acting as a personal assistant, for example, must quickly learn to interpret the emotions of a new customer during a short interaction period, necessitating the use of minimal data for training.

Briefly speaking, our objective is to identify underlying emotional patterns for different types of emotions and interpret the results in terms of primary facial movements that humans perform to express specific emotions. We will focus on seven specified emotions: calm, happy, sad, angry, fearful, disgusted, and surprised, aiming to solve multi-class classification problems. Our study concentrates on human facial expressions as the primary signal, excluding variables such as voice, ethnic differences, physical electric signals, and others. We will examine facial muscle movements independent of the environment, personal appearances, and background information.

This work addresses three main problems: 1) extracting and registering the expression evolution process at the same pronunciation speed, 2) detecting and exploring latent facial expression patterns common to specific human emotions, regardless of the performers, and 3) identifying the primary expressed emotion or quantifying the mixture of emotions from a given emotional facial video.

To address the practical challenges and achieve our research objectives, we employ a functional representation of data, treating measurements of facial muscle contractions collected discretely during a video as realizations of smooth, multivariate functional data. This approach reduces the dimension and complexity of the problem, representing infinite-dimensional functions as single data objects, and decreases the number of parameters to be identified in a statistical learning model while providing easily interpretable results.

When functional data come from different groups, it is reasonable to assume that each group's real phenomenon is characterized by a set of distinct underlying mean patterns describing the group's common attributes. Functional analysis of variance (FANOVA) models are appropriate for estimating and comparing functional group means and testing for significant differences. We are interested in detecting the functional mean patterns typical for each class (or emotion) and identifying which mean functions differ from others and when these differences occur. We utilize the point-wise permutation test for this purpose. Our methodology is first tested on simulated data to assess its effectiveness in identifying the correct underlying patterns and significant time zones where group-wise

patterns differ. We then apply the methodology to the RAVDESS dataset, consisting of videos in which professional actors perform different emotions. The outcome will be a set of functions describing which facial muscles are typically activated or deactivated and the order in which they must be engaged to reproduce a specific emotion.

The group-wise patterns are viewed as abstract intra-group insights derived from the data, representing filtered domain-dependent features specific to each group. This information can be utilized both as prior knowledge for emotion classification and to generate more realistic expressions in virtual characters. We explore the mapping of observed functions to a common low-dimensional representation space generated by the identified group-wise patterns across all classes. Subsequently, we leverage the generated scores as inputs to automatically classify a new set of functional data, gathered under conditions similar to our training set. We employ both non-parametric consensus voting schemes and parametric multinomial logistic regression methods for the multi-class classification task, assessing their predictive and explanatory abilities.

Addressing the two core questions of expression pattern detection and emotion recognition—how human facial expressions convey emotions and how to accurately classify newly observed facial videos into specific emotional categories—we propose a comprehensive model pipeline for training and testing the prediction accuracy of the model. Moreover, we assess actor performances to identify outliers and examine the effects of gender on detected emotional patterns. Importantly, the integrated approach proposed here ensures the model's applicability across different datasets for pattern detection and multi-class classification tasks, spanning various applications.

The structure of this thesis is as follows: Chapter 1 provides a brief introduction to the background knowledge of emotion recognition and expression detection from facial video data. Chapter 2 presents a general introduction to functional data analysis, covering its theoretical foundations, practical functional data construction, and functional curve registration methods. Chapter 3 illustrates our main methodology, which involves model construction and group-wise pattern detection through multiple multivariate function-

on-scalar regression. In Chapter 4, we concentrate on generating agreement scores using group-wise patterns and apply parametric and non-parametric classification methods for score-based multi-class classification. Chapter 5 brings together the entire pipeline for various stages of training and testing processes in the realm of automatic emotion recognition, revealing compelling results and providing insightful interpretations that unveil new perspectives on emotions and expressions.

Chapter 1

Emotion Recognition and Expression Detection from Facial Video Data

The study of human facial expressions never stops in our daily life while we communicate with others. Understanding emotions plays a crucial role in effective social communication among humans, influencing various aspects of life such as learning, innovation, creativity, motivation, decision-making, perception, and social interaction [21, 44]. While humans frequently assess the dynamic emotional states of others in daily conversations, interpreting the semantic meaning of human facial expressions and emotions poses a significant challenge for computers and "virtual humans" interacting with real people. Examples of relevant situations include machines functioning as functionaries, personal assistants, information providers, receptionists, or virtual humans for entertainment, video games, and virtual reality (e.g., in the Metaverse). Consequently, driven by diverse application needs, studies on the theoretical description and automatic detection of human facial expressions and emotions are gaining increasing attention in both academic research and technological development [30, 48, 140].

In collaboration with the Serbian company 3Lateral, which specializes in creating visual styles and designs for animation movies, this thesis is motivated by the detection of human emotions from expression evolution extracted from facial videos. We aim to

explore how to identify emotions by analyzing expressions and, furthermore, how to use this information to create more realistic and engaging virtual digital characters. Since human emotion detection serves as the primary motivating case study for developing mathematical and statistical techniques in this thesis, we dedicate this first chapter to introducing the context in which we will develop and apply our proposed methodologies in greater detail.

The structure of this chapter is as follows. First, in Section 1.1, we briefly describe the related literature in the context of emotion recognition from facial video data. Next, in Section 1.2, we introduce the concept of *facial expressions*, how they can be formally quantified (through *action units*), and detected in computer graphics using available software. Finally, in Section 1.3, we present the main dataset used for our study, RAVDESS, along with illustrations of its properties and visualizations.

1.1 A Review of Continuous Signal-based Emotion Recognition Approaches from Various Scopes

Emotions Descriptors. Many emotion theorists have claimed that there is a set of basic emotional categories. The most frequently used categories are Ekman's six basic emotions, in which the emotions can be grouped into six different categories: *happiness, sadness, surprise, disgust, anger,* and *fear* [37]. Additional categories such as amusement, boredom, excitement, and horror are less frequent and more challenging to detect, even for humans. In this research, we focus on the first six basic categories, plus the category "calm", in accordance with the dataset under study.

Besides the discrete categorical approach mentioned above, the *dimensional approach* offers an alternative to the discrete categorical method, dividing emotions into a 3D continuous space with measured variables of *arousal, valence*, and *dominance* [167]. Categorical emotional states can be mapped into the dimensional space [160], making this approach more suitable for emotion regression problems. However, we will not delve further into

the dimensional approach, as it is beyond the scope of our study.

Data Sources. While emotion detection from static images of human faces has been extensively studied [65, 66, 98], real-time emotion detection from facial videos remains an open problem. With the proliferation of smartphones and social media, video collections are rapidly expanding. Besides features on the human face related to facial expressions, signals on the human body, which respond to external emotional stimuli, can also be detected in various ways. This increasing availability of continuous data has spurred the development of modern intelligent systems, such as assistive devices and smart human-computer interfaces [14, 85]. Examples of relevant signals includ electrodermal activity (EDA) [44], electroencephalogram (EEG) [84], heart rate (HR) [137], facial electromyography (EMG) [116], the opening of the eyelids (EOG) [101] and sweating of the skin [8].

Classification Methods A standard procedure for video analysis mainly consists of two steps: video feature extraction and then emotion classification. First, several visual and audio features (e.g. action units, see Subsection 1.2.1) are extracted from videos to characterize the video content. Then, a general-purpose classifier is typically applied to recognize the expressed emotions [160]. Various machine learning methods have been investigated to model the relationship between video features and discrete emotional descriptors, including support vector machines [44, 110], Gaussian Mixture Models (GMMs) [118, 171], random forests [8], multi-layer feed-forward neural networks (NNs) [161] and hidden Markov models (HMMs) [102, 144].

Recently, deep learning-based models have outperformed traditional machine learning approaches, eliminating the need for a separate feature extraction phase. Some deep learning-based models are designed specifically for facial emotion detection tasks. In [9], the authors proposed using a Stacked Auto Encoder (SAE) to determine the best combination of muscles for describing a particular emotion, followed by a Softmax layer for multi-classification. Transfer learning techniques have been developed and applied for emotion recognition with multichannel data sources [106] and small datasets [115]. More

deep learning-based methods for emotion recognition from speech and visual information can be found in [4, 88, 94, 135].

1.2 Expression Detected from Facial Videos

To minimize the influence of individual facial appearance differences when analyzing expression evolution and corresponding emotions, researchers primarily focus on facial muscle movements. These movements, associated with the contraction or relaxation of specific facial muscles, can be encoded into *action units (AUs)* using the Facial Action Coding System (FACS) [39]. FACS is a widely-accepted standard in computer graphics for systematically categorizing physical expressions and extracting facial geometric features. As a result, FACS generates temporal profiles of each facial movement, breaking down AUs into continuous functions throughout a video [38, 133]. In this context, the AUs extracted from videos become the objects of study, suitable for any higher-order decision-making process on facial expression, including basic emotion recognition.

1.2.1 Using Facial Action Coding System to Encode the Facial Expressions into the Action Units

Facial Action Coding System (FACS) [39] is a system to taxonomize human facial movements by their appearance on the face. It has proven useful for psychologists and animators alike [40, 86, 133]. FACS encodes facial muscle movements by identifying subtle changes in facial appearance. Due to subjectivity and time-consuming limitations, FACS has been developed into an automated computational system that detects faces in videos, extracts facial movements, and deconstructs facial expressions into specific action units and their temporal segments [38]. In other words, FACS defines AUs.

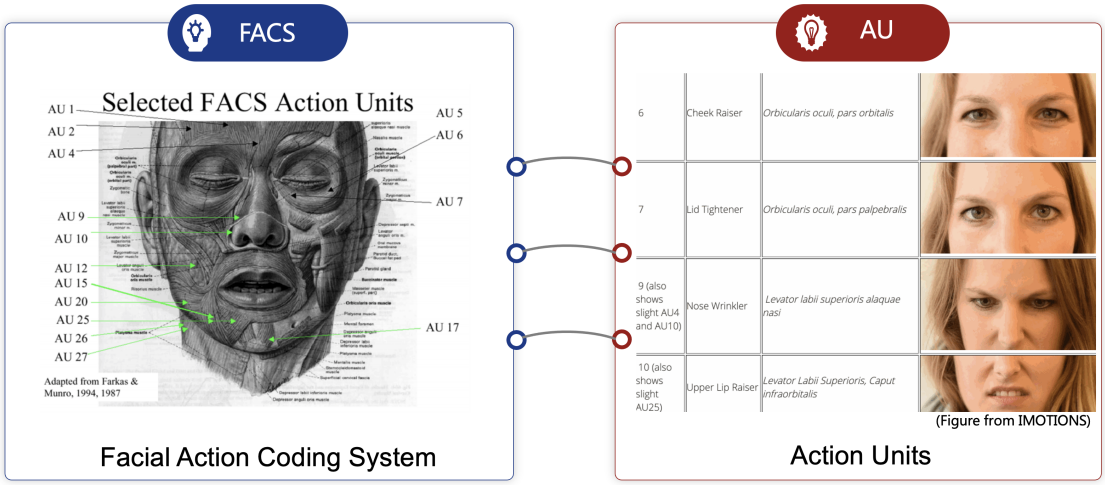


Figure 1.1: FACS and AUs. FACS: A common standard system to automatically categorize the physical expressions, and then produce temporal profiles of each facial movement. AU: A quantification of contraction or relaxation of one or more muscles of the face, deconstructed from facial expressions by FACS under temporal evolution. Credits to [3] and [2].

1.2.2 Softwares: OpenFace and Live Link Face

OpenFace is the first open-source tool for detecting AUs, introduced in 2016 and based on the FaceNet algorithm for automatic facial identification [5, 10]. Figure 1.2 [52] presents a screenshot of OpenFace in action, and the results of three exemplary extracted AUs from a video displaying the emotion of disgust are illustrated in Figure 1.3. OpenFace is widely used for facial landmark detection, head pose estimation, facial action unit recognition, and eye-gaze estimation [89, 153, 170]. Additionally, this tool offers real-time performance and can operate using a simple webcam without the need for specialized hardware [117].

In accordance with FACS rules, OpenFace can recognize and extract facial action units from facial images or videos [52, 159]. Figure 1.4 displays the user interface of OpenFace as it extracts AU values while processing a video frame by frame. We used OpenFace to extract the evolution of action units for the videos in our dataset. The software can extract 17 action units from each video, with each function having values ranging from $[0, 5]$ and sampled in each frame of the video.

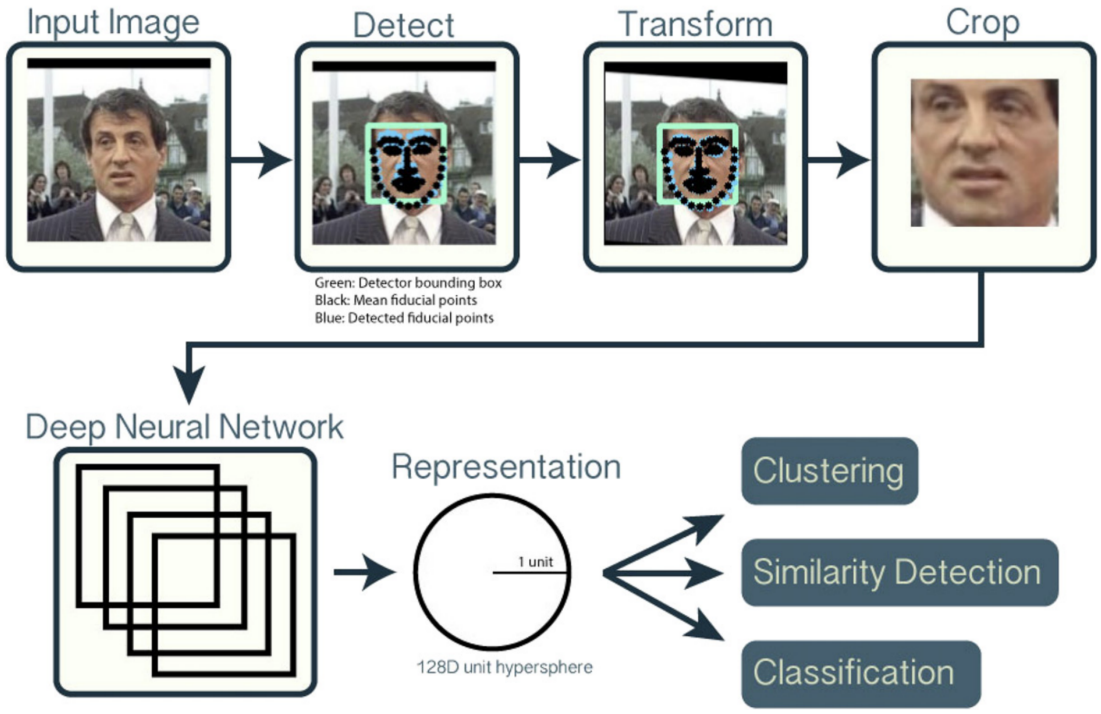


Figure 1.2: An example of the OpenFace procedure. Credits to [52].

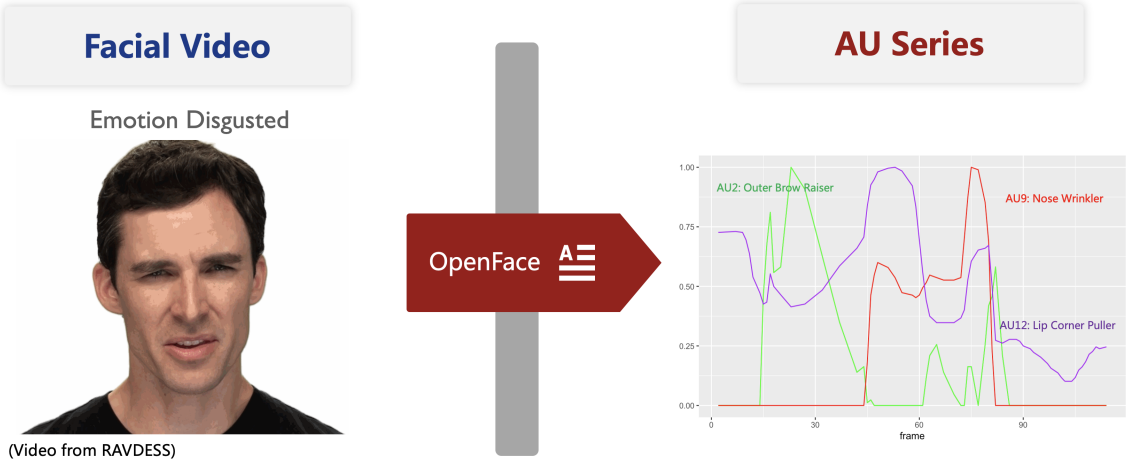


Figure 1.3: FACS helps to deconstruct facial expressions into action units.

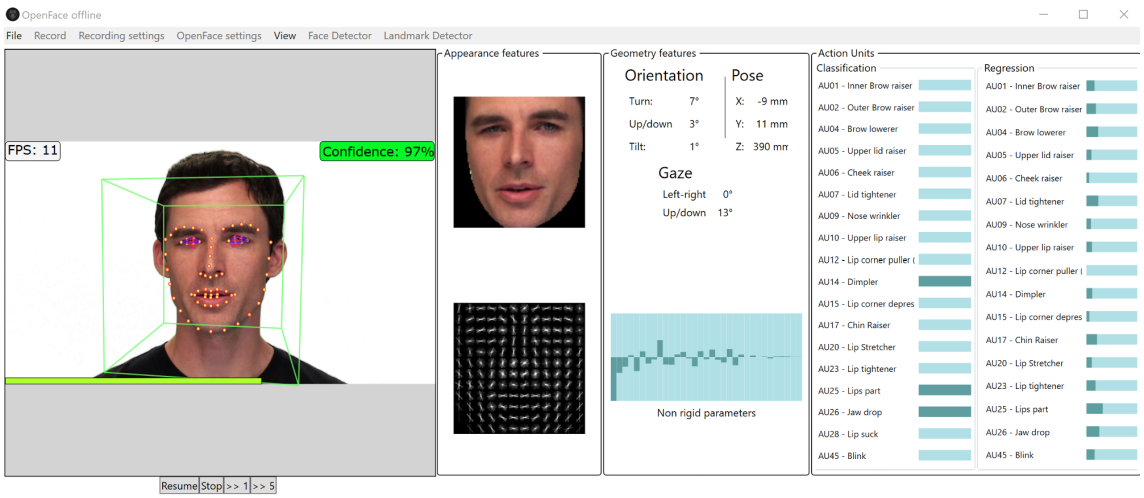


Figure 1.4: The input video and the output facial information for OpenFace.

Live Link Face utilizes Apple's TrueDepth Sensor from the latest iPhones to detect and 3D-map human face movements using a wire mesh mask via the *Live Link Face* App[1]. This app is an Unreal Engine tool, a 3D computer graphics game engine developed by Epic Games, which also includes our industrial partner 3Lateral as a member. Live Link Face can extract and analyze 52 Apple ARKit blend shapes of facial behavior (including head movement) at 60 frames per second [15]. Blendshapes are a standard approach for creating expressive facial animations in the digital production industry and are used as an alternative or complement to action units. The blend shape model is represented as a linear weighted sum of the target faces, which exemplify user-defined facial expressions or approximate facial muscle actions. The 52 different blend shapes detected by the app are automatically activated after Unreal detects incoming data from Live Link Face [53, 111]. Figure 1.5 illustrates an example of the animation generation process via Live Link Face.

Recent Live Link Face applications mainly involve blend shape extraction and avatar generation. For example, the app can fine-tune an avatar's face for more accurate digital human performance in social marketing strategies [111], analyze the facial behavior of an addressee in videos, and convert those behaviors to virtual characters [15], capture

[1] `https://apps.apple.com/us/app/live-link-face/id1495370836`.

facial movements of speakers to compare the influence of Prosody and Embodiment on the perceived naturalness of conversational agents' speech [36], contribute to a real-time tracking framework for developing and operating digital humans [156], and synchronize facial animation generation from speech for human-computer interaction [24].

Figure 1.5: An example of the use of Live Link Face procedure to generate virtual humans. Credits to [1].

Comparison between OpenFace and Live Link Face.

The primary advantages of OpenFace include its open-source nature, development in Python and Torch, compatibility with most operating systems (particularly Linux and OSX), and provision of a trained model without requiring specific hardware. OpenFace allows users to freely replace or alter its methods and parameters and does not necessitate extensive pre-processing steps for videos [10]. However, as noted in [52], OpenFace requires high-quality images for optimal efficiency.

Conversely, authors in [36] found that using Apple's TrueDepth Sensor via Live Link Face for recording provided better results than purely RGB-video-based solutions like OpenFace or speech-based facial animation like Oculus Lipsync. Given that the detected

features (facial blend shapes) in their study were used in the rendering process for generating virtual characters, the superior performance of 3D mesh-based techniques, such as those employed in Live Link Face software, is understandable.

In our research, and due to the nature of our dataset (RAVDESS dataset, see Section 1.3), we require software that can scan faces from pre-recorded videos, rather than real-time capture from live subjects. Additionally, our dataset consists of high-quality videos with uniform backgrounds, professional actors positioned centrally, and consistent criteria. Consequently, we focus on the 17 action units detected by OpenFace software and defer consideration of the 52 blend shapes detected by the Live Link Face app for future studies.

1.3 The RAVDESS Dataset

We base our research on human expression evolution and emotion identification on the analysis of the open-source RAVDESS dataset (Ryerson Audio-Visual Database of Emotional Speech and Song) [103]. RAVDESS is currently one of the best datasets for studying human emotions through facial videos, as it offers natural dynamic performances in standardized, high-quality videos with reasonable sample sizes. We can assume the facial muscle movements in these performances mimic real human actions under the same emotions.

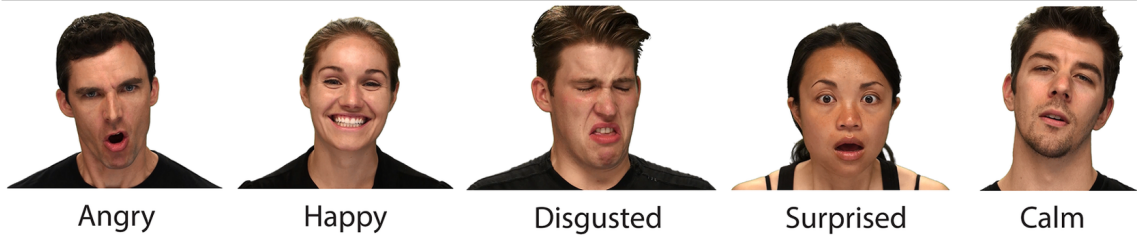
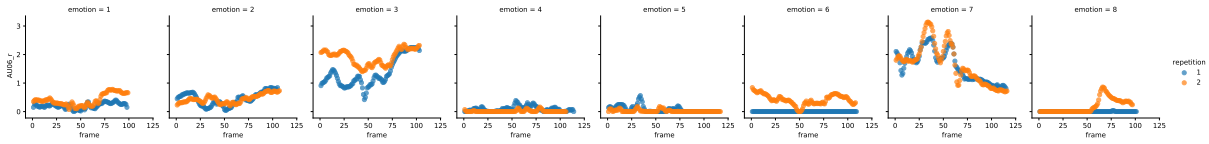


Figure 1.6: Photograms for some emotions in the RAVDESS dataset: Angry, Disgusted, Surprised, Calm. Emotion Sad and Fearful are also represented in the dataset, together with a neutral performance, for comparison studies.

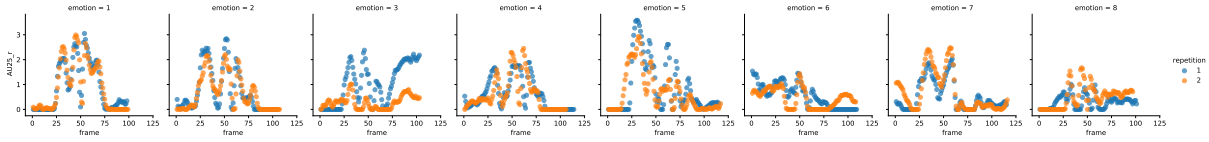
The dataset contains videos of 24 professional actors (12 female, 12 male) enacting

seven specified emotions: calm, happy, sad, angry, fearful, disgusted, and surprised (photograms for some emotions are shown in Figure 1.6). Additionally, a neutral performance is provided for each actor as a reference. The actors speak two predefined lexically-matched sentences, and for each emotion, two repetitions of the sentence with different emotional intensities are available. We focus on videos where 12 male actors pronounce the statement ("Kids are talking by the door") twice in a neutral North American accent with normal emotional intensity, using video-only modality and excluding audio information.

We use OpenFace software to extract the engagement degrees of action units from the selected RAVDESS videos. For one actor, depicted on the left side of Figure 1.3, we display the curves of AU06 in Figure 1.7a and AU25 in Figure 1.7b under different emotions. AU25 represents lip movement and is closely related to the pronunciation of the statement, serving as the reference variate for registration in Chapter 2. Furthermore, AU06 signifies cheek raiser movement and is identified as an important predictor for emotion recognition in Chapter 4. Additional plots for other actors can be found in A.2.



(a) AU06.



(b) AU25.

Figure 1.7: Examples of the evolution curves of AU06 and AU25 detected from the videos of one actor under different emotions.

1.3.1 Comparing RAVDESS to Alternative Facial Video Datasets

In [4], a review of the main datasets used for speech emotion recognition (SER) is presented. The authors categorize the datasets into three types: simulated, semi-natural, and

natural speech collections. Among the datasets listed in the paper, only a few contain visual signals, namely RAVDESS, IEMOCAP [19], and VAM [64].

As previously mentioned, the RAVDESS dataset offers rich variations in samples, including actor gender, intensity, number of repetitions, number of represented emotions, and the availability of a neutral performance for comparison. This is a crucial feature of RAVDESS, and only a few other datasets can claim to have such characteristics. RAVDESS is a simulated dataset with pre-designed, distinct emotions. However, the performed emotions may differ slightly from real situations in daily conversations, potentially leading to overfitting issues.

IEMOCAP [19] is a semi-natural English audiovisual dataset. Compared to simulated datasets, it has more natural performances but a limited number of sampled emotions due to the real constraints of less frequent emotions. Additionally, it is not an open-source dataset and is only available with a license.

VAM [64] is a natural audiovisual dataset based on dialogues from German TV talk shows. As such, it likely contains the most genuine emotions when compared to the other datasets. The videos are long and feature multiple emotions, so the sentences are cut and labeled separately with the corresponding emotions. However, the mixture of continuous and concurrent emotions in the same video and the dynamic variation in speech may result in inaccurate modeling. Furthermore, the limited data sources restrict the number of different emotions found in these videos. There may also be potential copyright and privacy issues when using this dataset.

The good qualities of the RAVDESS dataset are manifest: the data are evenly distributed in high frequency and evolve smoothly without rapid jitters. Nevertheless, the number of videos for each emotion in the dataset, and thus also the number of independent longitudinal data, is rather limited, since the action unit curves extracted from the same video are obviously correlated. Additionally, the number of action units observed in each video (that is 17) is comparable with the number of available videos for each emotion (that is 24). Therefore our dataset is composed by a small number of multivariate

longitudinal data of high dimension. This is the main obstacle to applying many classical time series methods in this study due to possible over-parametrization issues [62] and we will address this issue further in the next chapter.

Chapter 2

Data Preprocessing: Functional Data Format and Functional Curve Registration

In the previous chapter, we discussed action unit curves extracted from facial videos of actors' performances in the RAVDESS dataset, which are the focus of this thesis. We have discussed its positive qualities and the potential over-parametrization issue. Additionally, instead of predicting future values of our time series, we are interested in examining the "global shape" of the action unit curves to detect characteristics typical of each emotion for automatic classification. Functional data analysis (FDA) offers a suitable solution, as it represents the original discrete data as a set of continuous functions, focusing on similarities, differences, mean characteristics, and relationships between random functions.

This chapter will cover FDA's theoretical concepts and the functional data construction process. Researchers typically use smoothing and interpolation methods to construct functional data, assuming that the data belongs to a finite-dimensional functional space spanned by some functional basis. The second step involves registering the functions to prepare the data for further modeling processes. For our case study, we need to align the unregistered chronological timeline in the dataset into a common registered internal

timeline that follows a consistent pronunciation speed. This allows us to control, identify, and filter out the influence of speech and pronunciation, so we can study the specific influence of emotions.

This chapter is structured as follows: Section 2.1 provides a general introduction to functional data analysis, including a literature review, theoretical settings, and practical functional data construction. Section 2.2 describes functional curve registration methods and presents the registration results for the RAVDESS dataset.

2.1 Introduction to Functional Data Analysis

Functional data can represent data collected continuously over time for the same subjects in various natural, economic, or technological processes [125]. The central concept behind functional data is to consider point-wise collected data as noisy observations originating from a smooth random function that describes a real phenomenon [155, 157]. Functional Data Analysis (FDA) enables the in-depth analysis of properties of such smooth random functions for exploratory, confirmatory, or predictive data analysis [154] [7].

2.1.1 Literature review on Functional Data Analysis

The influential monograph by Ramsay and Silverman [125] serves as a starting point for research on statistical methods for random functions. This work described the primary features of FDA that can be applied to analyze continuous curves and illustrated the foundational theories of FDA with real case studies. Ferraty and Vieu [45] focused on functional observations from a non-parametric perspective, while Horváth and Kokoszka [74] provided an introduction to the Hilbert space approach to Functional Data Analysis and its associated theory. Additionally, Ullah [154] and Aneiros [7] offered systematic reviews of FDA applications, covering various fields where FDA has been applied.

Several standard statistical methods have been extended and adapted from classical point data to functional data to meet practical requirements. Functional data analysis

has been employed to explore common research problems, such as canonical correlation analysis [97], cluster analysis [169], discriminant analysis [83], [123], linear and nonlinear models [20], [43], the one-way ANOVA problem [32], principal component analysis [126], [16], regression models [12], [25] and variable selection [63].

Statistical methods for multivariate functional data have also been extended from univariate cases, enabling the simultaneous analysis of more than one function for a statistical unit. Examples include principal components analysis [13], [26], [68], [17], [47]; cluster analysis [148], [79], [82]; multivariate analysis of variance (MANOVA) approaches [60]; outlier detection methods [78], [78], [33]; variable selection and dimension reduction methods [61]; regression problems for multivariate functional data [27], [162], [57], [43], among others.

2.1.2 Functional Data Theoretical Setting and Functional Principal Component Analysis (FPCA)

We first introduce the conceptual setting in building functional data from the noisy observed time points, under the general multivariate and multi-class case. Let Ω be a probability space and $\mathbf{Y}$ be a L_2-continuous multivariate stochastic process defined on Ω, where $\mathbf{Y}$'s realization is a set of D curves, $D \geq 2$. Assume that each of these D curves stands for one scalar function in one dimension defined on a given finite interval $[0, T],\ 0 < T < \infty$, such that $\mathbf{Y}(t) = [y_1(t), \ldots, y_D(t)]^\intercal,\ t \in [0, T]$ (denoted sometimes as $\mathbf{Y} = [\mathbf{y}_1, \ldots, \mathbf{y}_D]^\intercal$ for simplicity). We additionally assume to have an independent sample of $\mathbf{Y}(t)$, observed as a set of multivariate curves. The mean function of $\mathbf{Y}(t)$ is defined as $\mathbb{E}[\mathbf{Y}(t)] = \boldsymbol{\mu}(t) = (\mu_1(t), \ldots, \mu_D(t))^\intercal$, $t \in [0, T]$, where $\mu_d(t) = \mathbb{E}[y_d(t)]$ for $d = 1, \ldots, D$. The covariance operator of $\mathbf{Y}$ is then defined as an integral operator $\mathcal{C}$ with kernel

$$\mathbf{C}(t, s) = \mathbb{E}[(\mathbf{Y}(t) - \boldsymbol{\mu}(t)) \otimes (\mathbf{Y}(s) - \boldsymbol{\mu}(s))], \forall t, s \in [0, T],$$

where $\otimes$ is the tensor product on $\mathbb{R}^D$. Thus, $\mathbf{C}(t,s)$ is a $D \times D$ matrix with elements $\mathbf{C}(t,s)[m,n] = Cov(\mathbf{Y}_m(t), \mathbf{Y}_n(s)), \forall m,n = 1,\ldots,D$. Under the hypothesis of L_2-continuity, $\mathcal{C}$ is a Hilbert-Schmidt operator, which, in other words, is compact and self-adjoint [81].

Meanwhile, the functional extension of principal component analysis (FPCA) [13] on the covariance operator $\mathcal{C}$ provides a countable set of positive eigenvalues $\{\eta_{j,j\geq 1}\}$ associated to a countable orthonormal basis of multivariate eigenfunctions $\{\mathbf{f}_{j,j\geq 1}(t)\}$, $\mathbf{f}_j(t) = [f_{j,1}(t),\ldots,f_{j,D}(t)]^\intercal$, such that

$$\mathbf{C}\mathbf{f}_j(t) = \eta_j \mathbf{f}_j(t),$$

with $\eta_1 \geq \eta_2 \geq \cdots \geq 0$ and $\langle \mathbf{f}_i(t), \mathbf{f}_j(t)\rangle_{\{L_2([0,T])\}^D} = \delta_{i,j}$ with $\delta_{i,j} = 1$ if $i = j$, and $\delta_{i,j} = 0$ otherwise. Assume that the principal component score ξ_j of $\mathbf{Y}(t)$ is a zero-mean random process, obtained as the projection of the centered $\mathbf{Y}(t)$ on the jth eigenfunction $\mathbf{f}_j(t)$, i.e.,

$$\xi_j = \int_0^T \langle \mathbf{Y}(t) - \boldsymbol{\mu}(t), \mathbf{f}_j(t)\rangle dt = \int_0^T \sum_{d=1}^{D} (y_d(t) - \mu_d(t)) f_{j,d}(t) dt. \tag{2.1}$$

Therefore, $\mathbf{Y}(t)$ varies around the mean function $\mu(t)$ with random amplitude variations in the directions of $\{\mathbf{f}_j\}_{j\geq 1}$. By applying a Karhunen-Loeve expansion [51] to $\mathbf{Y}(t)$, we obtain

$$\mathbf{Y}(t) = \boldsymbol{\mu}(t) + \sum_{j\geq 1} \xi_j \mathbf{f}_j(t) \text{ and } y_d(t) = \mu_d(t) + \sum_{j\geq 1} \xi_j f_{j,d}(t). \tag{2.2}$$

By substituting the series in (2.4) with the corresponding truncated series, we obtain the best approximation of the considered process $\mathbf{Y}(t)$, under the mean square criterion, in the finite-dimensional space of functional principal components.

2.1.3 Functional Data Construction

To reconstruct functional data from discrete data, researchers typically use smoothing and interpolation methods, assuming that functional data belong to a finite-dimensional functional space spanned by functional basis functions. This process allows functional data objects to be constructed from time series data by specifying a set of basis functions and coefficients defining their linear combination. Non-periodic time series often use B-spline basis functions, which are continuous piecewise polynomial functions, due to their computational efficiency and flexibility. An alternative is to use functional principal components as the basis for functional data reconstruction, which automatically considers data variability since the basis is generated through the covariance operator.

For the method based on splines, assume that the curve in dth dimension of kth sample $\mathbf{Y}_k(t)$ is $y_{d,k}(t)$ and lies in the span of a set of basis functions $\Theta(t) = (\theta_1(t), \ldots, \theta_q(t), \ldots, \theta_Q(t))^\intercal$ with basis size Q. Then the observed sample in one dimension and in the multivariate case can be expressed as a linear combination of the elements of $\Theta(t)$, with coefficients in vectors $\mathbf{a}_{d,k}, d = 1, ..., D$ and matrix $\mathbf{A}_k$, respectively, such that

$$y_{d,k}(t) = \sum_{q=1}^{Q} a_{d,k,q}\theta_q(t) = \mathbf{a}_{d,k}^\intercal \Theta(t), \text{ and } \mathbf{Y}_k(t) = \mathbf{A}_k\Theta(t), \text{ with } \mathbf{A}_k = [\mathbf{a}_{1,k}, \ldots, \mathbf{a}_{d,k}, \ldots, \mathbf{a}_{D,k}]^\intercal. \tag{2.3}$$

For the method based on functional principal components, by applying a Karhunen-Loeve expansion [51] to $\mathbf{Y}_k(t)$, we can obtain

$$y_{d,k}(t) = \mu_d(t) + \sum_{j\geq 1} \eta_{j,k} f_{j,d}(t) \text{ and } \mathbf{Y}_k(t) = \boldsymbol{\mu}(t) + \sum_{j\geq 1} \eta_{j,k}\mathbf{f}_j(t). \tag{2.4}$$

By approximating the series in (2.4) with the corresponding truncated series, we can obtain the best approximations of the considered process $\mathbf{Y}(t)$ in the space of functional principal components, given the mean and FPC's functions properly represented by a set of flexible basis functions such as B-spline in Equation 2.3.

We first use the method based on functional principal components in the registration

process where FPCA is involved to calculate the warping functions. More details are in Section 2.2. Then, for the modeling process in Chapter 3, to build the function-on-scalar regression and estimate the coefficients, we use the method based on B-splines. We show in Figure 2.1 examples of discrete points and constructed functions of AU25 for the first actor under different emotions, where the functions captured the main shapes of the discrete sequences and are smooth as expected. Since the construction method based on functional principal components basis also relies on a fundamental spline basis, we fixed a set of B-splines with 20 basis functions based on empirical experiments for both types of methods in the further study.

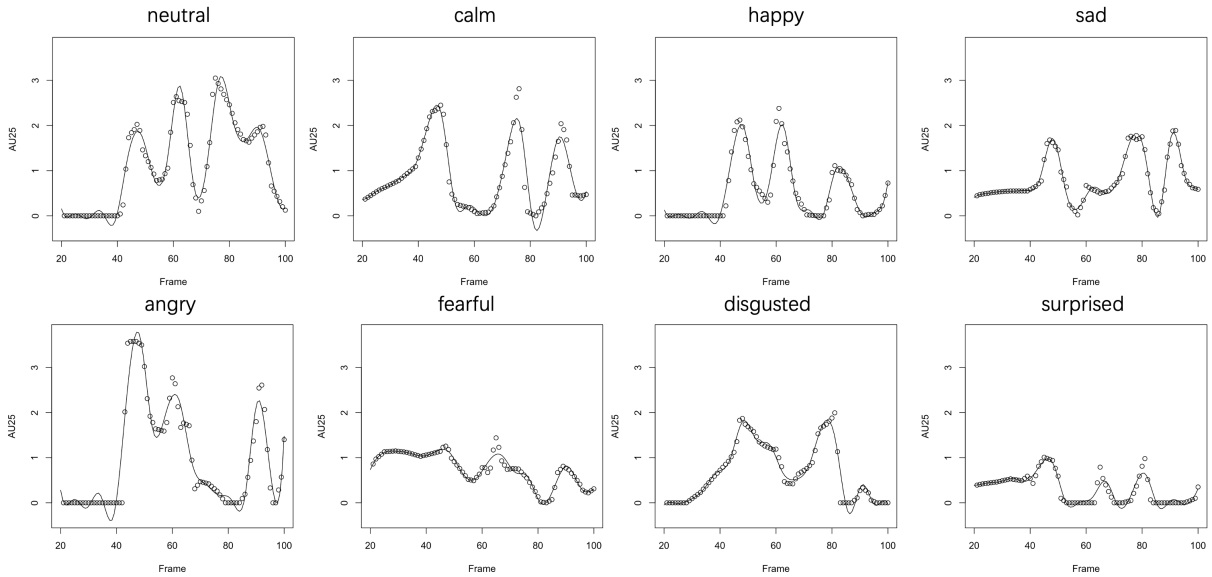


Figure 2.1: Examples of discrete points and constructed functions of AU25 for the first actor under all emotions.

2.2 Functional Curve Registration

Functional data samples are often not provided in a pre-registered form. It is more common for longitudinal data to be recorded in time frames of varying lengths and sampled at different time instants. This is the case in our driving study: the videos collected in the RAVDESS dataset have slightly different durations, and the actors pronounce the sentence

with slightly different starting times and speeds. These factors contribute to significant phase variability in the data. As a result, the functional data collected from the videos require a second preprocessing phase for alignment. The primary goal of this section is to align observed functions onto a common registered internal timeline by isolating the phase variability of the functions and preserving the amplitude variability. After functional data construction and registration, the data will be prepared for the subsequent chapter, where we identify underlying group-wise mean patterns and significant time zones where they differ (in our driving case study, groups are identified by different emotions).

This preprocessing phase can be omitted if the data are already registered, such as in the analysis of financial time series data observed within the same timeframe and at the same discrete time points.

2.2.1 Literature Review for Functional Registration Methods

When analyzing functional data that are not perfectly aligned, many researchers overlook the fact that there are two sources of potential variability in functional data: amplitude variability and phase variability. Phase variation typically manifests as a random change of time scale, while amplitude variability, separated from phase variability after registration, contains more statistically significant values for study. Therefore, if functional data are not aligned, they must be registered prior to any further modeling procedures.

There is a sizable literature on curve registration, mainly in three areas: landmark-based registration, metric-based registration, and model-based registration. In landmark-based registration, the performance of the model relies on the ability to select correctly the facial landmarks with distinguishable characteristics of points on the face in 2D or 3D space. This type of method focuses on detecting the occurrence of structural features (e.g. peaks and valleys) [54, 92, 126], and automatically identifying mathematical landmarks with quantified uncertainty [143]. Metric-based registration formulates an optimality criterion to estimate amplitude and phase components [127, 142, 158], frequently via a nonparametric regression method [56, 132]. The choice of the optimality criterion is crucial

and can drastically influence the registration result [107, 141, 175]. The Fisher–Rao (FR) metric and the square-root velocity function (SRVF) representation, introduced in [142] to overcome the computation difficulty of L_2 metric by using the Fisher–Rao distance, has been the basis for several recent approaches to registration [77, 150, 166]. Essentially, the FR distance between two functions is equivalent to the L_2 distance between their respective SRVF transformations [152]. Further, model-based registration can be viewed as an extension of metric-based approaches where registration is determined through a formal statistical model [29, 50, 145].

The model-based registration has an additional advantage that it can be directly linked to common inferential tasks. A popular approach is to use jointly functional principal component analysis (FPCA) when selecting and aligning the templates [93]. As we introduced in Section 2.1.2, FPCA is good at discovering dominant directions of variation and reducing feature dimensions in modeling functional observations. Therefore, it is a natural tool for identifying the features to which data is registered. If data is unregistered and the phase variability is ignored, the resulting model may fail to capture patterns presented in the functional data [93]. If the FPCA method is coupled with registration in regression models, simultaneous estimation of regression and warping parameters can be achieved in handling phase and amplitude variability [151, 165]. There are many studies operating under and expanding this framework. First, these methods estimate the template, and then estimate the warping functions for a given template; these steps are iterated until convergence [55, 67, 109, 151, 152]. They calculate a mean template and define an SRVF of the observed curves, similar to the metric-base registration. Since the SRVF uses the derivative of the observed curve, the data to be registered are required to be smooth. In contrast, rather than aligning pre-smoothed observed functional data, the method proposed in [165] registers observed curves using smooth templates based on the inverse warping functions. The inverse warping functions can be represented using the B-spline basis with coefficients via a pre-defined design matrix. We follow the

principal-components-based registration method proposed in [165] [1] due to its efficiency and registration ability, and illustrate briefly its theory in Section 2.2.2.

2.2.2 Iterative Process via FPCA and Warping Functions Estimation

Let us use $\mathbf{Y}$ to represent the L_2-continuous multivariate stochastic process under study and $y_{d,k}(t)$ to denote the evolution of one specific variate $d \in \{1, \ldots, D\}$ in sample $k \in \{1, \ldots, K\}$. Here we denote by $y_{d,k}(t)$ the registered version, defined over a common internal time t, while we denote by $y_{d,k}(t_k^*)$ the original curve, defined over an individual chronological and non-registered time t_k^* in the sample k. In the case of our motivating case study of emotion detection, all the action units from the same video are recorded synchronously through the video's chronological time domain. Therefore, in order to detect the common internal time t, we just need to register the curves in one dimension, that is for one AU, instead of repeating the registration in all dimensions. Since the action unit AU25, which represents the lips movement, is strongly related to the pronunciation of the statement, we decided to align the first AU25 into a common internal timeline, across the videos.

For simplicity, since we are focusing our attention on one specific variate, we omit the subscript d in our notations and we denote the univariate sample curves of interest as $\{y_k(t)\}_{k\in\{1,\ldots,K\}}$, for the registered version, and $\{y_k(t_k^*)\}_{k\in\{1,\ldots,K\}}$ for the original observed ones.

During the process of aligning observed functions into a common registered internal timeline, the crucial point is to isolate the phase variability of the original functions, while keeping the amplitude phase unchanged in the registered template functions to maintain the information under interest. Since the phase variation of one function is normally represented by a random change of time scale, we define the non-linear warping functions

[1] Codes are available in the R package "registr" [164].

$h_k : [0,T] \to [0,T], k = 1,\ldots,K$ to map the unregistered chronological time t_k^* onto registered internal time t, so that $h_k(t) = t_k^*$, $h_k^{-1}(t_k^*) = t$, with $E[h_k(t)] = t$.

Meanwhile, we register observed curves using smooth templates based on the inverse warping functions $h_k^{-1}(t_k^*)$, rather than aligning pre-smoothed observed functional data $\{y_k(t_k^*)\}$. The inverse warping functions $h_k^{-1}(t_k^*)$ can be represented using the B-spline basis $\boldsymbol{\Theta}$ of basis size Q with coefficients $\boldsymbol{\delta}_k$ via a design matrix $\boldsymbol{\Lambda}_h \in \mathbb{R}_{T\times Q}$, such that

$$h_k^{-1}(t_k^*) = \boldsymbol{\Lambda}_h(t_k^*)\boldsymbol{\delta}_k. \tag{2.5}$$

Generally speaking, we consider the joint estimation of FPCA and warping functions in the generative process which iterates between two steps, after giving initial guesses.

The subject-specific mean $\mu_k(t)$ is considered as the template in the registration. Moreover, each $y_k(t)$ can be decomposed into a smooth global mean $\mu_0(t)$, FPCs $\mathbf{f}_j(t)$ with the respective eigenvalues $\eta_j > 0$, and subject-specific FPC scores $\xi_{k,j} \sim N(0,\eta_j)$ are treated as random effects. We have

$$E[y_k(h_k^{-1}(t_k^*))|h_k^{-1}] = E[y_k(t)] = \mu_k(t) = \mu_0(t) + \sum_{j\geq 1}\xi_{k,j}\mathbf{f}_j(t). \tag{2.6}$$

Thus, the first step consists in using FPCA to calculate the template and principal component scores for each sample, conditioned on the current inverse warping function h_k^{-1} (and implicitly on its parameters $\boldsymbol{\delta}_k$), which maps the observed curve $y_k(t_k^*)$ to its template.

We now turn to the second step in our iterative algorithm, in which the inverse warping function h_k^{-1} is estimated for each subject conditionally on the template. This is done by maximizing the log-likelihood of the probability function of $y_k(t_k^*)$, given the current estimates of the subject-specific mean $\mu_k(t)$. In the Gaussian case, the estimate $\hat{h}_k(t_k^*)$ is obtained equivalently by maximizing the log-likelihood of the observed data over candidate warping functions, or by minimizing the usual L_2 loss. Therefore, we have

$$\hat{h}_k(t_k^*) = \arg\max_{h_k} -[l(h_k^{-1}|y_k, \mu_k)] = \arg\min_{h_k} \int (y_k(h_k^{-1}(t_k^*)) - \mu_k(t))^2 dt. \tag{2.7}$$

In practice, estimation through Equation 2.7 is complemented by the constraints $h_k(0) = 0$, $h_k(1) = 1$, and $h_k(t_k^*)$ is an increasing function.

Essentially, the warping function parameters $\boldsymbol{\delta}_k$ in (2.5) are estimated separately for each subject using a constrained optimization and loop over subjects. The constrained optimization can be made more efficient with an analytic form of the gradient, as detailed in [165].

These two steps are alternated until the errors in successive iterations converge. Through this process, the functions are automatically aligned using nonlinear time warping, and the obtained registered internal timeline retains only the amplitude information, which serves as a valuable tool for comparing and classifying functions.

2.2.3 Curve Registration Results for Action Units

The RAVDESS dataset's videos have different durations, and the actors pronounce the sentences with varying starting times and speeds. Therefore, we first need to align the curves by separating and removing phase variability from amplitude variability. As mentioned earlier, we align the observed AU25 curves across the videos into a common internal timeline. In this way, the warping functions estimated by registering AU25 in each video serve as a template to compress or stretch the time frames of the other AUs accordingly.

After curve alignment, it is possible that one registered time point targets several different AU values when the original timeline is compressed. In such cases, we perform linear interpolation, resulting in a "mean" approximation of these values as the single AU value at the registered time point. Then we can rebuild the dataset, using the commonly aligned frames as columns. During this process, many missing values appear in the dataset, so interpolation is needed again because each video has a different match to

the registered timeline.

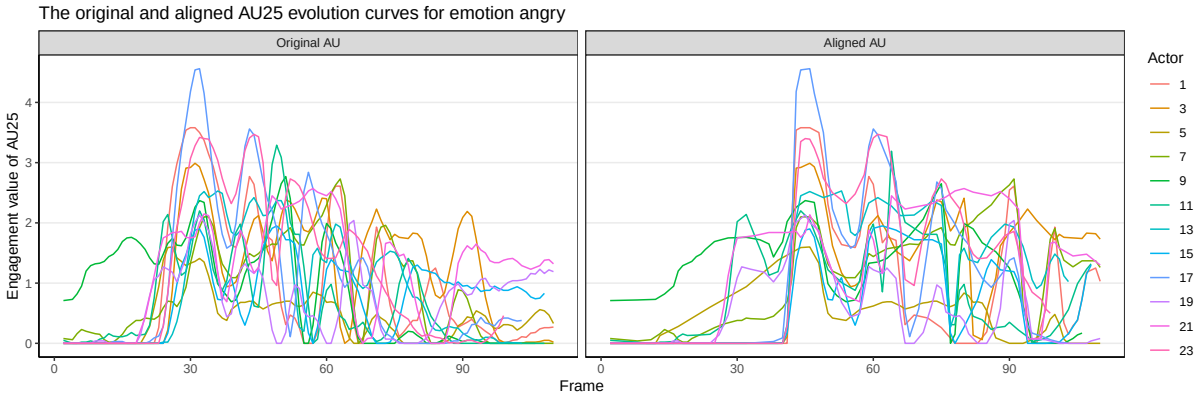


Figure 2.2: Example of registration of the curves of AU25 for the male actors representing the emotion angry in the corresponding videos.

Figure 2.2 illustrates the shapes of the curves of AU25, before and after the registration process for the emotion angry case, as an example (the registration results for all emotions can be found in Appendix A.3). After registration, the curves concentrate more closely together while preserving their overall shape. Moreover, the peaks and valleys across different emotions are aligned onto the same time points, which correspond to the pronunciation of some open-mouth syllables, enabling point-wise cross-group comparisons without the influence of time-space mismatch. After registration, all the AUs are represented by functions sampled in 120 time frames. Approximately, the first and last 20 frames are breaks before and after the speech, respectively, so the variation there does not influence further analysis. The study then focuses only on the frames between 20 to 100, ensuring that only meaningful information without empty breaks remains.

After completing the pre-processing steps, the functional data gathered from the videos are smoothed and aligned, making the data ready for formally identifying the underlying group-wise mean patterns and significant time zones in the next step. We treat the neutral performance as the control group and the seven emotions (calm, happy, sad, angry, fearful, disgusted, and surprised) as classification groups (with $G = 7$). We also consider the 17 registered AUs as the variates of our multivariate functional dataset (with $D = 17$). The group sample mean curves of the AUs are shown in Figure 2.3. To improve

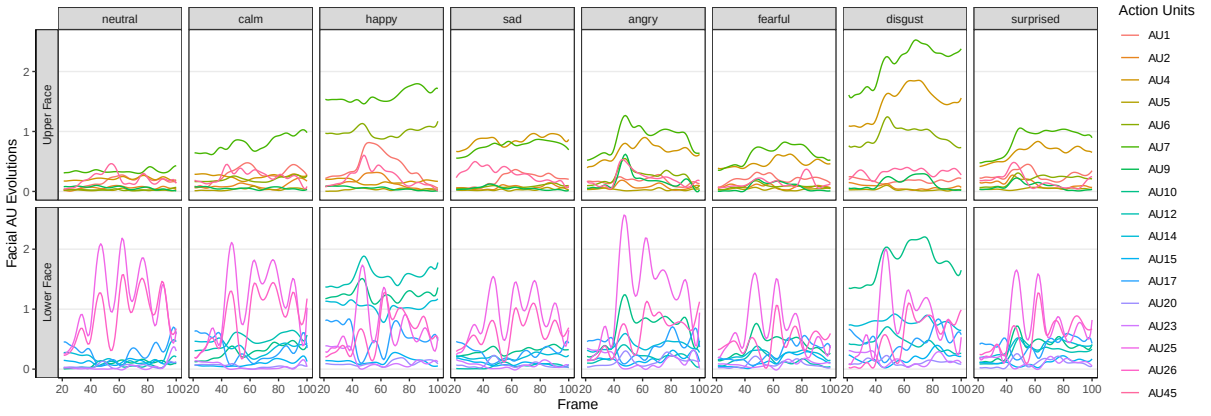


Figure 2.3: The group sample mean curves of 17 registered AUs (expression evolution curves) under eight emotions in the RAVDESS dataset. The top row reports the AUs related to the upper part of the face, while the bottom row reports the AUs related to the lower part of the face.

visualization, we divide the AUs into two separate groups: one group representing the upper face movements (AU01: Inner Brow Raiser, AU02: Outer Brow Raiser, AU04: Brow Lowerer, AU05: Upper Lid Raiser, AU06: Cheek Raiser, AU07: Lid Tightener, AU09: Nose Wrinkler, AU45: Blink) and another group representing the lower face movements (AU10: Upper Lip Raiser, AU12: Lip Corner Puller, AU14: Dimpler, AU15: Lip Corner Depressor, AU17: Chin Raiser, AU20: Lip Stretcher, AU23: Lip Tightener, AU25: Lips Part, AU26: Jaw Drop).

Chapter 3

Model Construction: Group-wise Pattern Detection through Multiple Multivariate Function-on-Scalar Regression

In cases where functional data originate from distinct groups, it is reasonable to presume that the authentic phenomena defining each group are characterized by diverse underlying mean patterns, which represent shared attributes within the group. Consequently, employing FANOVA models or, equivalently, function-on-scalar regression models, is appropriate for estimating and comparing the functional means of these groups, as well as for testing the existence of significant disparities. However, FANOVA tests merely reveal information concerning the presence of overall significant differences in the groups' means. In our driving case study, our objective is to discern which functional mean patterns are emblematic of each class (or emotion) and, therefore, to determine which mean functions deviate from the others and the temporal periods at which these differences emerge. With a reference class corresponding to the neutral emotion, our intention is to juxtapose the behavior of other classes against this baseline. To achieve this, we set up a technique

based on point-wise ANOVA and permutation tests to identify time zones where the observed patterns exhibit significant differences. Consequently, we define the group-wise patterns as the estimated group effect functions, filtered by the time zones in which they significantly deviate from the control neutral case.

The efficacy of the proposed methodology is initially assessed through simulated data. Sensitivity analysis regarding the primary parameters is executed by gauging the capacity to pinpoint the accurate underlying patterns and the precise time periods where the group-wise patterns exhibit significant variation. Subsequently, we apply the methodology in our real-world case study, based on the RAVDESS dataset. The evolution of expressions under various emotions of interest is initially registered as delineated in Chapter 2.2. Following this, we analyze the mean behaviors of the AUs curves according to the methodology we propose.

The structure of this chapter goes as follows. Section 3.1 illustrates the construction process of multiple multivariate function-on-scalar regression models. Section 3.2 expounds on the identification of group-wise typical patterns through permutation tests. In Section 3.3, we present the findings of the proposed methodology for both simulated data and the RAVDESS dataset, respectively.

3.1 Multiple Multivariate Function-on-Scalar Regression

Function-on-scalar regression, a less commonly known term, is one of the three fundamental functional regression models applied to numerous situations where observations manifest as entire curves or functions. Contrasting with function-on-function regression and scalar-on-function regression, function-on-scalar regression addresses instances where responses are functions and predictors are scalars. Application areas of function-on-scalar regression encompass various fields, including human physical activity states [58, 95], integration of features from wearable technology across multiple physical domains [35],

genetic analysis related to lung growth [42] or blood pressure and cardiovascular health [11], brain images such as diffusion tensor imaging data [100, 130, 136, 176], EEG experiments investigating brain activity [173], plant phenotype data on root bending and lunar effects [168], and fluorescence spectroscopy in a cervical pre-cancer study [178].

In this thesis, we restrict our attention to function-on-scalar regression. Moreover, we address multiple multivariate function-on-scalar regression, while "multiple" (also "multi-level", "multi-group", or "multi-class") refers to situations where subjects or samples are acquired from various groups with distinct labels and "multivariate", on the other hand, pertains to cases involving multiple covariates (or variables) measured on each experimental unit. In our driving case study of emotional pattern detection, we regard the AUs' evolutions as multivariate functional responses and the multiple labels of the corresponding video's emotion as the non-time-varying scalar predictors.

3.1.1 Model Construction

Consider a dataset consisting of registered functional data with known group memberships. This dataset serves as the training set to estimate the group-wise effect parameter functions of a general multiple multivariate function-on-scalar regression model. Let $y_{g,d,k}(t)$ represent the evolution of a specific variate $d \in \{1, \ldots, D\}$ in sample $k \in \{1, \ldots, K\}$ for group $g \in \{0, \ldots, G\}$. We treat $g = 0$ as the control group, and the evolution curves of this group are compared with those of other groups $\tilde{g} \in \{1, \ldots, G\}$ in G pair comparisons. In the context of emotion recognition, the neutral group (i.e., videos in which the predefined sentences are pronounced without expressing any emotion) corresponds to $g = 0$, which serves as the control group and is compared with other emotions $\tilde{g} \in \{1, \ldots, G\}$.

To investigate the spatio-temporal characteristics of the emotional responses, we construct a semiparametric functional mixed model with a focus on the delineation of group-wise differences. The proposed model captures the inherent variability within each group and facilitates the estimation of the temporal evolution patterns for each variate d and

group g. The evolution $y_{g,d,k}(t)$ can then be decomposed into three components:

$$y_{g,d,k}(t) = \mu_{d,0}(t) + \alpha_{d,g}(t) + \epsilon_{g,d,k}(t), \tag{3.1}$$

where $\mu_{d,0}(t)$ is the grand mean function independent of group membership, $\alpha_{d,g}(t)$ represents the additional effect of group g on the considered variate d, and $\epsilon_{g,d,k}(t)$ denotes the unexplained zero-mean variation specific to the k-th sample within group g. To uniquely identify the functional parameters, we impose the conventional constraint on additional effects, that is,

$$\sum_{g=0}^{G} \alpha_{d,g}(t) = 0, \forall t. \tag{3.2}$$

Analogous to [130], we express Model (3.1) as a linear regression model in a function-on-scalar format, enabling the application of the common functional least squares method to estimate the parameters. In this context, individual evolution functions are regressed on the scalars representing the individual's constant (non-time-varying) group memberships. By categorizing the functions originating from the same group, we can define a $((G+1)K+1) \times (G+2)$ design matrix $\mathbf{Z}_d$ as described in [125, Section 9.2], with appropriate 0 and 1 entries to describe functions' group memberships and the functional decomposition as in Equation (3.1). Therefore, we have

$$\mathbf{y}_d(t) = \mathbf{Z}_d \beta_d(t) + \epsilon_d(t), \forall t, \tag{3.3}$$

where

$$\mathbf{y}_d(t) = \begin{bmatrix} y_{0,d,1}(t) \\ \vdots \\ y_{0,d,K}(t) \\ y_{1,d,1}(t) \\ \vdots \\ y_{1,d,K}(t) \\ \vdots \\ y_{G,d,1}(t) \\ \vdots \\ y_{G,d,K}(t) \\ 0 \end{bmatrix}, \mathbf{Z}_d = \begin{bmatrix} 1 & 1 & 0 & 0 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \cdots & \vdots & \vdots \\ 1 & 1 & 0 & 0 & \cdots & 0 & 0 \\ 1 & 0 & 1 & 0 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \cdots & \vdots & \vdots \\ 1 & 0 & 1 & 0 & \cdots & 0 & 0 \\ & & & & \vdots & & \\ 1 & 0 & 0 & 0 & \cdots & 0 & 1 \\ \vdots & \vdots & \vdots & \vdots & \cdots & \vdots & \vdots \\ 1 & 0 & 0 & 0 & \cdots & 0 & 1 \\ 0 & 1 & 1 & 1 & \cdots & 1 & 1 \end{bmatrix}, \beta_d(t) = \begin{bmatrix} \mu_{d,0}(t) \\ \alpha_{d,0}(t) \\ \alpha_{d,1}(t) \\ \vdots \\ \alpha_{d,G}(t) \end{bmatrix} = \begin{bmatrix} \beta_{d,0}(t) \\ \beta_{d,1}(t) \\ \beta_{d,2}(t) \\ \vdots \\ \beta_{d,G+1}(t) \end{bmatrix}, \epsilon_d(t) = \begin{bmatrix} \epsilon_{0,d,1}(t) \\ \vdots \\ \epsilon_{0,d,K}(t) \\ \epsilon_{1,d,1}(t) \\ \vdots \\ \epsilon_{1,d,K}(t) \\ \vdots \\ \epsilon_{G,d,1}(t) \\ \vdots \\ \epsilon_{G,d,K}(t) \\ 0 \end{bmatrix}.$$

In this form, $\beta_d(t)$ represents the corresponding set of $G+2$ functional parameters, including the mean $\mu_{d,0}(t)$ and the $G+1$ additional groups effects $\{\alpha_{d,0}(t), \alpha_{d,1}(t), \ldots, \alpha_{d,G}(t)\}$. The term $\epsilon_d(t)$ contains the noise, possessing the same dimension as $\mathbf{y}_d(t)$. Moreover, the additional constraint (3.2) on the group-wise effects is implicitly incorporated into the final line of the model's matrix form in Equation (3.3). Following the notations and settings above, we can assemble all the D variates in a vector and concurrently solve the model in Equation (3.4) at once:

$$\mathbf{Y}(t) = \mathbf{Z}\beta(t) + \epsilon(t), \tag{3.4}$$

where

$$\mathbf{Y}(t) = \begin{bmatrix} \mathbf{y}_1(t) \\ \vdots \\ \mathbf{y}_d(t) \\ \vdots \\ \mathbf{y}_D(t) \end{bmatrix}, \mathbf{Z} = \begin{bmatrix} \mathbf{Z}_1 \\ \vdots \\ \mathbf{Z}_d \\ \vdots \\ \mathbf{Z}_D \end{bmatrix}^\intercal, \beta(t) = \begin{bmatrix} \beta_1(t) \\ \vdots \\ \beta_d(t) \\ \vdots \\ \beta_D(t) \end{bmatrix}, \epsilon(t) = \begin{bmatrix} \epsilon_1(t) \\ \vdots \\ \epsilon_d(t) \\ \vdots \\ \epsilon_D(t) \end{bmatrix}.$$

3.1.2 Model's Parameter Estimation in Functional Space

Using the functional basis reconstruction techniques shown in Equation (2.3), we assume that the observed response functions can be represented as $\mathbf{Y}(t) = \mathbf{A}\Theta(t)$, with known ba-

sis coefficient matrix $\mathbf{A}$, given a prefixed B-spline basis $\Theta(t) = [\theta_1(t), \ldots, \theta_q(t), \ldots, \theta_Q(t)]^\intercal$. Additionally, we assume that the parameter functions can be represented as $\beta(t) = \mathbf{B}\Theta(t)$ with the coefficient matrix $\mathbf{B}$ underestimation. In our application, we used a basis size $Q = 20$. In more detail, to specify better the entries of the matrices $\mathbf{A}$ and $\mathbf{B}$ in Equation (3.4) and to introduce some notations, we first assume that each curve $y_{g,d,k}(t),\ d \in \{1, \ldots, D\}$ for the k-th video can be expressed as a linear combination of the basic elements,

$$y_{g,d,k}(t) = \sum_{q=1}^{Q} a_{g,d,k,q}\theta_q(t) = \mathbf{a}_{g,d,k}^\intercal \Theta(t),\ \mathbf{a}_{g,d,k}^\intercal = [a_{g,d,k,1}, \ldots, a_{g,d,k,Q}].$$

Then, by combining the coefficients in a matrix form, we have

$$\mathbf{y}_{g,d}(t) = \mathbf{A}_{g,d}\Theta(t), \mathbf{A}_{g,d} = [\mathbf{a}_{g,d,1}, \ldots, \mathbf{a}_{g,d,K}]^\intercal = \begin{bmatrix} a_{g,d,1,1} & \ldots & a_{g,d,1,Q}, \\ \vdots & \vdots & \vdots \\ a_{g,d,K,1} & \ldots & a_{g,d,K,Q} \end{bmatrix},$$

$$\mathbf{Y}_g(t) = \mathbf{A}_g\Theta(t), \mathbf{A}_g = [\mathbf{A}_{g,1}, \ldots, \mathbf{A}_{g,d}, \ldots, \mathbf{A}_{g,D}]^\intercal,$$

$$\mathbf{Y} = \mathbf{A}\Theta(t), \mathbf{A} = [\mathbf{A}_1, \ldots, \mathbf{A}_g, \ldots, \mathbf{A}_G]^\intercal.$$

Similarly, the coefficient function $\beta(t) = [\beta_1(t), \ldots, \beta_D(t)]^\intercal$ can be decomposed on the same functional basis, such that

$$\beta(t) = \mathbf{B}\Theta(t),\ \text{where } \mathbf{B} = [\mathbf{B}_1, \ldots, \mathbf{B}_d, \ldots, \mathbf{B}_D]^\intercal, \mathbf{B}_d = \begin{bmatrix} b_{1,d,1} & \ldots & b_{1,d,Q}, \\ \vdots & \vdots & \vdots \\ b_{G,d,1} & \ldots & b_{G,d,Q} \end{bmatrix}.$$

Therefore, to estimate the parameters of Model (3.4) we need to find a matrix $\mathbf{B}$ which

minimizes the functional least squares loss function $L(\mathbf{B})$, given by

$$\min_{\mathbf{B}} L(\mathbf{B}) = \min_{\mathbf{B}} \int \|\mathbf{A}\Theta(t) - \mathbf{ZB}\Theta(t)\|_F^2 dt,$$

where $\|\cdot\|_F$ is the Frobenius norm, $\mathbf{A}\Theta(t)$ are the observed response values and $\mathbf{ZB}\Theta(t)$ are the estimated ones.

Let $J_{\theta\theta}$ be the $Q \times Q$ matrix of the inner products of each pair of functional bases, with its (i, j) entry given by $\int \theta_i(t)\theta_j(t)dt$. By solving this optimization problem on $R^{(G+1)DK\times T}$ space, the loss function can be expressed in the following way:

$$\begin{aligned}
L(\mathbf{B}) &= \int \|\mathbf{A}\Theta(t) - \mathbf{ZB}\Theta(t)\|_F^2 dt \\
&= \int \|(\mathbf{A} - \mathbf{ZB})\Theta(t)\|_F^2 dt \\
&= \int trace[((\mathbf{A} - \mathbf{ZB})\Theta(t))^\intercal((\mathbf{A} - \mathbf{ZB})\Theta(t))]dt \\
&= \int trace[\Theta(t)^\intercal(\mathbf{A} - \mathbf{ZB})^\intercal(\mathbf{A} - \mathbf{ZB})\Theta(t)]dt \\
&= trace[\int \Theta(t)^\intercal(\mathbf{A} - \mathbf{ZB})^\intercal(\mathbf{A} - \mathbf{ZB})\Theta(t)dt] \\
&= trace[(\mathbf{A} - \mathbf{ZB})^\intercal(\mathbf{A} - \mathbf{ZB}) \int \Theta(t)^\intercal\Theta(t)dt] \\
&= tr[((\mathbf{A} - \mathbf{ZB})J_{\theta\theta}^{\frac{1}{2}})^\intercal(\mathbf{A} - \mathbf{ZB})J_{\theta\theta}^{\frac{1}{2}} \\
&= vec[(\mathbf{A} - \mathbf{ZB})J_{\theta\theta}^{\frac{1}{2}}]^\intercal vec[(\mathbf{A} - \mathbf{ZB})J_{\theta\theta}^{\frac{1}{2}}] \\
&= \|vec[(\mathbf{A} - \mathbf{ZB})J_{\theta\theta}^{\frac{1}{2}}]^\intercal\|_2^2 \\
&= \|vec(J_{\theta\theta}^{\frac{1}{2}}\mathbf{A}^\intercal) - vec[J_{\theta\theta}^{\frac{1}{2}}\mathbf{B}^\intercal\mathbf{Z}^\intercal)\|^2 \\
&= \|vec(J_{\theta\theta}^{\frac{1}{2}}\mathbf{A}^\intercal) - (\mathbf{Z} \otimes J_{\theta\theta}^{\frac{1}{2}})vec(\mathbf{B}^\intercal)\|^2.
\end{aligned}$$

Here, $\otimes$ denotes the Kronecker product between matrices and $vec(\cdot)$ represents the column vector converted by stacking the columns of a matrix on top of one another. Assume that $\mathbf{Z} \otimes J_{\theta\theta}^{\frac{1}{2}}$ has full column rank, and then, by minimizing $L(\mathbf{B})$ with respect to $\mathbf{B}$, that is set to 0 the derivatives of $L(\mathbf{B})$, we can derive the least squares estimator of the coefficient

matrix $\hat{\mathbf{B}}$ in a vectorized form, as

$$vec(\hat{\mathbf{B}}^\intercal) = [(\mathbf{Z} \otimes J_{\theta\theta}^{\frac{1}{2}})^\intercal (\mathbf{Z} \otimes J_{\theta\theta}^{\frac{1}{2}})]^{-1} (\mathbf{Z} \otimes J_{\theta\theta}^{\frac{1}{2}})^\intercal vec(J_{\theta\theta}^{\frac{1}{2}} \mathbf{A}^\intercal).$$

The estimated coefficients are then used to construct the parameter functions. In the following, we denote the estimated common means across groups and the group-wise effects, obtained through the above procedure, as $\hat{\mu}_{d,0}(t)$ and $\hat{\alpha}_{d,g}(t), \forall d \in \{1, \ldots, D\}$ and $g \in \{0, \ldots, G\}$.

3.2 Group-wise Patterns Identification via FANOVA and Contrasts

Once the functional parameters are estimated, we can perform statistical inference and model validation by analyzing the estimated group effects and their uncertainty. To assess the statistical significance of the differences between the mean curves of the control group and the other groups, we first introduce a series of hypothesis tests has been proposed in the literature in Subsection 3.2.1.

Furthermore, the test statistic for examining the equality of the group means is considered using the division of pointwise between-subject variations and within-subject variations under corresponding degrees of freedom. Then it is compared to the critical value, a prefixed quantile of the permutation-based distribution of the test statistic under the null hypothesis, to determine whether to reject the null hypothesis at every time point. By employing this non-parametric, permutation-based approach, we can more accurately identify the specific time intervals where the emotional patterns differ significantly and the patterns which characterize each group in Subsection 3.2.2. This information is essential for the development of automatic recognition or classification and contributes to a deeper understanding of the underlying emotional dynamics in facial expressions.

3.2.1 Previous Studies on Equality Tests for Mean Curves

The one-way analysis of variance (ANOVA), a widely employed method for group comparisons, facilitates the comparison of the means of k populations [6, 41]. Originating from the study of correlation coefficients among groups or variables [46], ANOVA has evolved into a generalized method for comparing means of normal random vectors [80] and has been regarded as a special case of linear regression with deterministic covariates [112]. Functional ANOVA (FANOVA) extends ANOVA's capabilities by enabling the examination of random functions over time, highlighting the intrinsic differences among these functions [31, 172]. In the realm of functional data, FANOVA has been extensively explored for comparing the mean functions of two or more groups of functional samples. Horváth assessed the equality of operators defining linear functional models using proposed test statistics on magnetometer data [75] and compared the means of two temporally dependent populations in a fully functional manner [76]. Górecki [59] considered a comprehensive numerical comparison of various methods, while the textbook [172] provided a thorough review of FANOVA problems. FANOVA tests have found practical applications in diverse fields such as profile monitoring of signals in process control [122], chemometrics [146], economics [128], and environmental studies [108].

To delve deeper into the analysis of group-wise patterns, we draw particular attention to the pointwise F-test, which inspired our work. Based on the functional analysis of variance (FANOVA), a commonly employed approach for hypothesis testing in functional data analysis is the functional F-test [130]. Proposed by Ramsay and Silverman [129], this test seamlessly extends the classical F-test to functional data analysis. The test statistic for the group means-testing problem ($\mathcal{H}_0 : \mu_1(t) = \cdots = \mu_k(t)$) is defined as $F_n(t) = \frac{SSR_n(t)/(k-1)}{SSE_n(t)/(n-k)}, t \in T$, where $SSR_n(t)$ and $SSE_n(t)$ represent the pointwise between-subject and within-subject variations, respectively, k denotes the number of groups, and n signifies the total number of observations.

Various global tests, such as the GPF test [174], offer a less computationally intensive

globalized version of the pointwise F-test, and the global T_n and $Tmax$ tests for multivariate functional data [124], derived from the pointwise Hotelling T^2-test, have been introduced. These global tests circumvent the *multiple comparison problems* that arises when pointwise F-test statistics are significant for all $t \in T$ at a given significance level, but the one-way ANOVA problem is not overall significant at the same significance level.

The primary objective of global tests is to determine if the functional means of the compared groups differ, without delving into the specifics of where and how significant these differences are. To identify the location and magnitude of these differences, Pini and Vantini [120–122] proposed interval-wise testing for functional data. This approach relaxes the need to analyse together the entire temporal range of the data in the global test, and introduces adjusted and unadjusted p-value functions to select time domain portions where the two means exhibit significant differences. Catering to various functional hypotheses, they perform the test on each possible set of consecutive basis coefficients and calculate the corresponding adjusted/unadjusted p-values for each basis component. In the case of data embedded in $L_2(T) \cap C_0(T)$, the point-wise p-value coincides with the p-value of the permutation test based on the test statistic for all $t \in T$. In our motivating case study, the Action Unit (AU) data is continuous after the smoothing pre-processing step. Our point-wise testing can be considered a narrowed-down version of interval-wise testing, where the interval length is fixed to a one-time unit, and the number of intervals equals the number of frames in our data. Interval-wise testing could be a valuable extension to our selection strategy for detecting longer time zones in which differences between groups on single time points are not crucial, and do not need to be detected. However, this approach, in spite of being valuable in many applications, presents a trade-off when applied to our main case study of emotion recognition. While it can identify more stable and continuous time zones, it may overlook micro-expressions that happen suddenly and last for a very short duration (less than one second), occasionally going unnoticed even by humans, but which are quite relevant to identify an emotion. If point-wise testing reveals a significant difference between a neutral state and an emotion in a single time unit rather

than in a continuous interval, it could indicate the presence of a micro-expression. With interval-wise testing, such micro-expressions might be missed as other points within the interval could reduce the overall significance. For this reason we choose to develop our analyses using a pointwise test in time.

Given the restrictive assumptions of functional ANOVA tests concerning data distribution, it is suggested to pursue non-parametric techniques that require fewer assumptions about the data. In [129], the authors employed a non-parametric approach, a permutation-based method, to approximate the distribution of test statistics under the null hypothesis, thus allowing for a more flexible assessment of the differences between functional means. This approach is particularly advantageous when dealing with complex, multivariate functional data, as it does not rely on stringent assumptions about the data's distribution. The permutation test is literally a resampling method, calculating all possible values of the test statistic when the original observed data are permuted in paired groups [99, 134].

3.2.2 Group-wise Effect Tests and Significant Time Zone Detection via FANOVA and Permutation Tests

After constructing the function-on-scalar regression model and estimating the associated functional parameters, our objective is to examine whether and at what time points the group labels significantly influence the functional mean of the data. In our emotion detection application, our primary interest lies in determining which groups significantly deviate from the control one (representing a neutral emotional state in our case study), and the time frames in which the means differ significantly. To achieve this, contrasts are employed to compare the means of pairs of groups, defined as linear combinations of the regression coefficients with terms totaling zero.

For each variate $d \in \{1, \ldots, D\}$ and group $g \in \{0, \ldots, G\}$, assume that the k-th sample curve $y_{d,g,k}(t)$ at a fixed time t follows a $\mathbb{N}(\mu_{d,g}(t), \sigma_d^2(t)), k \in \{1, \ldots, K\}$ distribution, with group mean $\mu_{d,g}(t)$ and variance $\sigma_d^2(t)$, which are assumed to be constant across

groups and samples for the same variate. Consequently, the subgroup sample mean is $\overline{Y}_{d,g,\cdot}(t) = \frac{1}{K}\sum_{k=1}^{K} y_{d,g,k}(t) \sim \mathbb{N}(\mu_{d,g}(t), \frac{\sigma_d^2(t)}{K})$. Our goal is to construct a test for each pair of groups under comparison, namely between a control group $g = 0$ (whose data, in our setting, are occupying the first rows of $\mathbf{y}_d(t)$ in Model (3.3)) and another group under study $g = \tilde{g}$ for one variate d. In the i-th test ($i \in \{1, \ldots, DG\}$) we design a vector $\mathbf{a} = [0, \ldots, 0, 1, 0, \ldots, 0, -1, 0, \ldots, 0]$ of length $D(G+2)$ (with the same length of $\beta(t)$ in last section) with the entries $a_{d(G+2)+2} = 1$, $a_{d(G+2)+2+\tilde{g}} = -1$, and 0 in the remaining entries. As a result, the sum of weights in $\mathbf{a}$ equals zero. Denoting the vector of subgroup means $\mu_{d,g}(t)$ for all $d \in \{1, \ldots, D\}$ and $g \in \{0, \ldots, G\}$ as $\boldsymbol{\mu}(t)$, we can define the contrast in the i-th test as $c = \mathbf{a}^\intercal \boldsymbol{\mu}(t) = \alpha_{d,0}(t) - \alpha_{d,\tilde{g}}(t) = (\alpha_{d,0}(t) + \mu_{d,0}) - (\alpha_{d,\tilde{g}}(t) + \mu_{d,0})$. Thus, the difference in the subgroup sample means $\overline{Y}_{d,0,\cdot}(t) - \overline{Y}_{d,\tilde{g},\cdot}(t)$ serves as an unbiased estimator of the contrast c. In this context, we can define DG distinct tests, in which the null hypothesis at each time point t for the control group and the group under study $\tilde{g}$, for the variate d, is

$$\mathcal{H}_0 : \mathbf{a}^\intercal \boldsymbol{\mu}(t) = 0,$$

or, equivalently,

$$\mathcal{H}_0 : \alpha_{d,0}(t) - \alpha_{d,\tilde{g}}(t) = 0,$$

and the related F statistics under the null hypothesis is [22]

$$F_{d,\tilde{g}}(t) = \frac{(\overline{Y}_{d,0,\cdot}(t) - \overline{Y}_{d,\tilde{g},\cdot}(t))^2}{\frac{2}{K} S_p^2(t)},$$

where $S_p^2(t) = \frac{1}{N-DG}\sum_{g=1}^{G}\sum_{d=1}^{D}\sum_{k=1}^{K}(y_{d,g,k}(t) - \overline{y}_{d,g,\cdot}(t))^2$ is the total sample variance at time t, and $N = GDK$ is the total sample size. Under our assumptions of gaussianity for the functional data $y_{d,g,k}(t)$ and the additional assumptions regarding the variance, under the null hypothesis, for any given $t \in \mathcal{T}$, $F_{d,\tilde{g}}(t)$ follows an F-distribution with 1 and $N - DG$ degrees of freedom, denoted as $F_{1,N-DG}$.

The statistic $F_{d,\tilde{g}}(t)$ should be compared with the critical value, or $(1-\alpha)$-quantile,

$c_{1-\alpha}$ of an $F_{1,N-DG}$ distribution, for any predetermined significance level α. Thus, as usual, we reject the null hypothesis when $F_{d,\tilde{g}}(t) > c_{1-\alpha}$. Now we define $\mathcal{T}^* = \{t \in \mathbb{R}_+ | F_{d,\tilde{g}}(t) > c_{1-\alpha}\}$, which represents the set of time intervals in which we observed a significant difference in the mean behavior of group $\tilde{g}$ from the control group, at level α. We then exploit $\mathcal{T}^*$ to define the *typical mean patterns* characterizing each group as $\tilde{\alpha}_{d,g}(t)$, where $\tilde{\alpha}_{d,g}(t) = \hat{\alpha}_{d,g}(t)$ when $t \in \mathcal{T}^*$ and $\tilde{\alpha}_{d,g}(t) = 0$ when $t \notin \mathcal{T}^*$. Therefore, in our setting, the functions $\tilde{\alpha}_{d,g}(t)$ reflect the mean behavior that new functional data must exhibit, in the identified specific time intervals $\mathcal{T}^*$, to belong to group $\tilde{g}$ and to be distinguished from the control group.

In fact, since the distribution of the $F_{d,\tilde{g}}(t)$ statistics is a Fisher F only under some quite restrictive assumptions such as data normality and constant variance, as previously mentioned, we applied a nonparametric permutation test to recover the critical values. The permutation test is not based on any particular assumption on the distribution of the data, apart from their exchangeability or permutation invariance under the null hypothesis [90]. It aims to approximate the distribution of the reference statistics under $\mathcal{H}_0$. One permutation is obtained by reshuffling the labels of the samples of the two groups under comparison. Then the F statistic $F_{d,\tilde{g}}(t)$ is computed on each permuted sample, and by repeating the procedure n times, the sample distribution of the F statistics under $\mathcal{H}_0$ is recovered [28]. The decision rule does not change: given the quantile $c_{1-\alpha,n}$ of the permutation distribution, the null hypothesis is rejected when $F_{d,\tilde{g}}(t) > c_{1-\alpha,n}$.

Note that the typical group-wise mean patterns $\tilde{\alpha}_{d,g}(t)$ can be easily interpreted in the context of the particular problem under study. For example, in the case of our motivating case study, positive (or, respectively, negative) values assumed by $\tilde{\alpha}_{d,g}(t)$ over large time intervals mean that the action unit d is strongly activated (respectively, deactivated) for a long time while expressing emotion g; successive positive values of $\tilde{\alpha}_{d_1,g}(t)$ and $\tilde{\alpha}_{d_2,g}(t)$ on different time intervals reveal that to express emotion g, one need first to activate action unit d_1 and then action unit d_2; when $\tilde{\alpha}_{d',g}(t)$ is almost always close to 0, it means that the action unit d' is not relevant in recognizing emotion g, and so on. Thus, the methodology

proposed here to extract the typical group patterns reflects our original aim to establish an interpretable statistical learning procedure.

3.3 Numerical Results on Simulated and Real Data

In this section, we present the evaluation of our proposed methodology on both simulated and real data. Our objective is to assess the method's capacity to accurately discern the underlying patterns and significant time zones in which the group-wise patterns exhibit considerable differences. Additionally, we investigate the method's capabilities and limitations under varying parameter configurations, such as the noise level in the data, the sample size, and the chosen significance level of the quantile employed in the permutation test. Note that the quantile in the permutation tests plays a pivotal role in addressing the "multiple testing" problem, as the values of functional data at distinct and close time instants are inherently correlated, rendering our tests non-independent. This may result in a modification of the overall significance of the permutation tests, necessitating the determination of an 'optimal' value for the chosen significance level, at least empirically. Subsequently, we apply our method to the primary case study of emotional pattern detection. The identified patterns can provide valuable insights into typical facial movements, i.e., the expressions, associated with each emotion, and facilitate an automatic classification of emotions expressed by a human face. It is worth mentioning that, in spite of the specific motivating case study here considered, the proposed methodology and the related software can be readily applied to a diverse range of similar situations across various fields of application.

3.3.1 Evaluation of Model Efficiency on Simulated Data

To explore the accuracy and efficiency of our proposed methodology, we generate synthetic functional data, partitioned into eight groups, analogous to our primary case study, and structured according to Model (3.1). Furthermore, we manipulate the model parameters

to conduct a sensitivity analysis of our methodology. The workflow of the employed strategy is depicted in Figure 3.1.

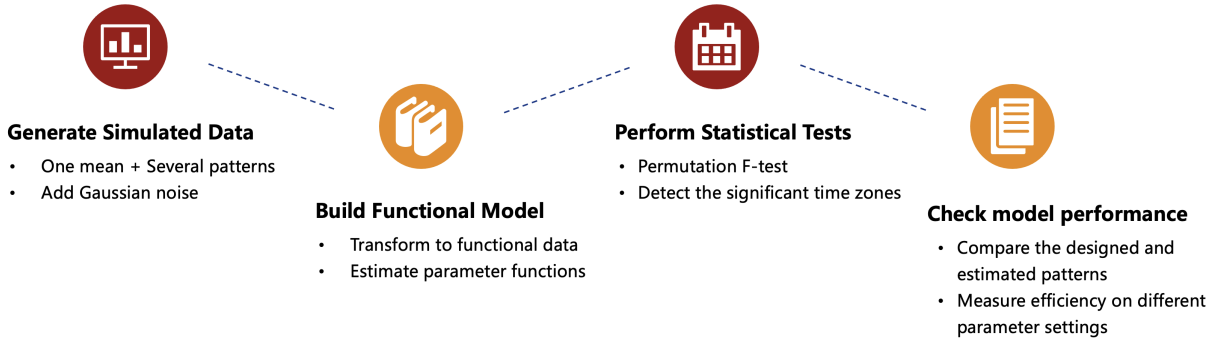
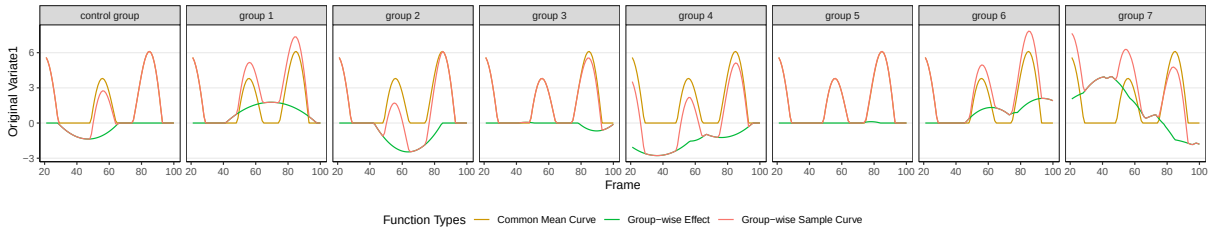


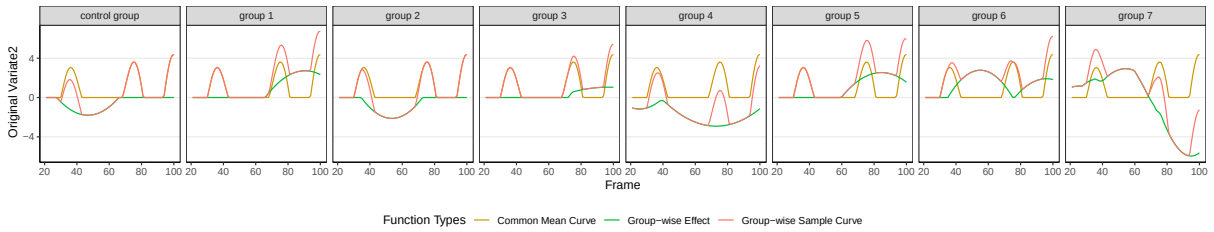
Figure 3.1: Simulation workflow in order to assess the accuracy and computational efficiency of the proposed methods.

Simulation settings. We simulate a dataset of 3-variate functional data divided into eight groups (thus, in our setting, $D = 3, G = 7$ and the group with index $g = 0$ serves as the control group, corresponding to the neutral expression in our motivating case study). The ground mean and additive effects for each group's functions are generated by combining a set of sine, cosine, and log curves with randomly-selected coefficients and cutoff points, using a fixed seed. The support for these functions ranges from 20 to 100, mirroring the action unit curves' settings in the primary case study. Figure 3.2 illustrates the main mean characteristics of our simulated patterns. The dark yellow curve represents the ground mean of each variate $\mu_{d,0}(t)$, the green curve denotes the additive effect attributable to the group $\alpha_{d,g}(t)$, which differs from zero exclusively in specific time zones, and the red curves signify the resulting mean group behavior $E(y_{g,d,k}(t)) = \mu_{d,0}(t) + \alpha_{d,g}(t)$.

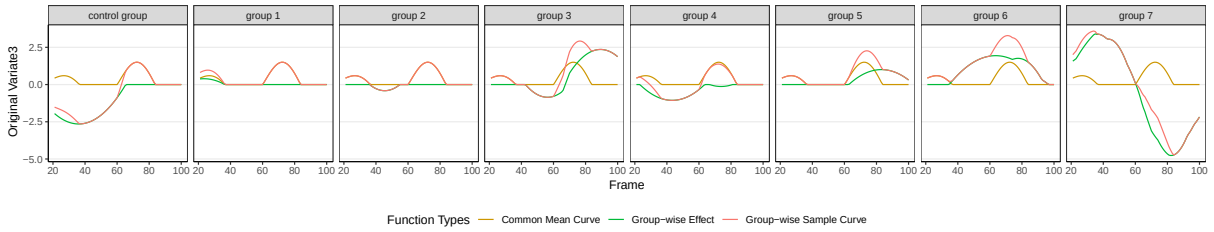
Eventually, to generate our data, a random Gaussian noise with zero mean is added to each mean function $E(y_{g,d,k}(t))$. The parameters under study include the standard deviation of the Gaussian noise (spanning in the discrete set $\{0.05, 0.5, 1, 1.5, 2, 2.5, 3, 3.5, 4, 4.5, 5, 10\}$); the simulated sample size of each group (ranging in the set $\{24, 50, 76, 100\}$, with 24 being the total number of data available in each group in our reference case study); and the quantile level considered in the permutation test to determine the critical value (varying



(a) Variate 1.



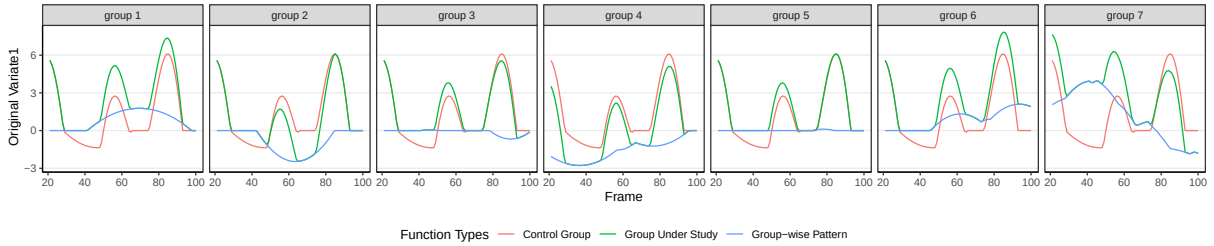
(b) Variate 2.



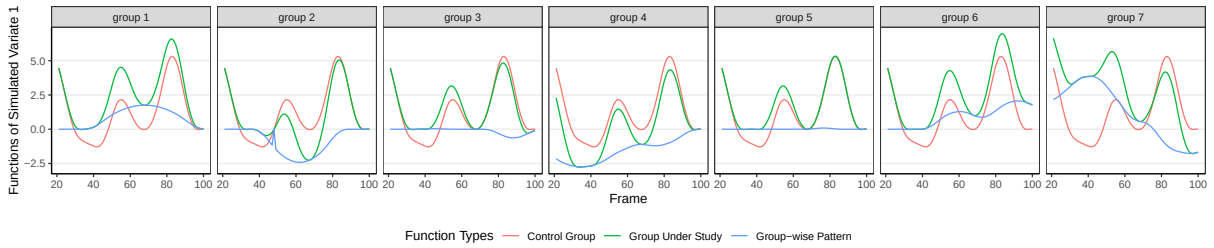
(c) Variate 3.

Figure 3.2: The group-wise typical mean functions of three variates in simulated data, composed as the sum of one common functional mean and group-wise effects. The dark yellow curve represents the ground mean of each variate $\mu_{d,0}(t)$, the green curve denotes the additive effect attributable to the group $\alpha_{d,g}(t)$, which differs from zero exclusively in specific time zones, and the red curves signify the resulting mean group behavior $E(y_{g,d,k}(t)) = \mu_{d,0}(t) + \alpha_{d,g}(t)$

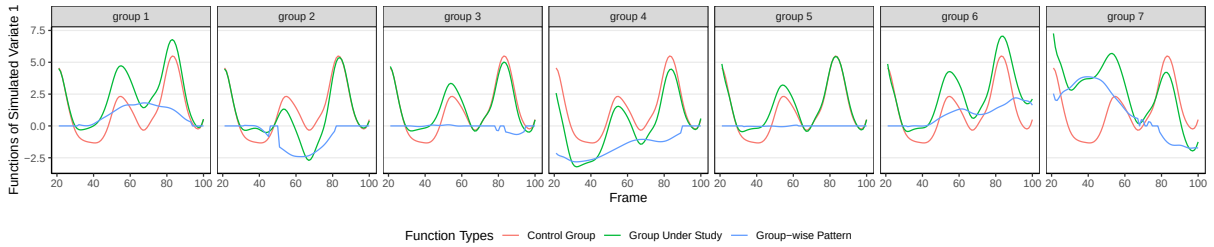
among $\{0.85, 0.875, 0.9, 0.925, 0.95, 0.975, 0.99, 0.999\}$). As in each permutation, we need to reshuffle the labels of the samples of the two groups under comparison, we set the number of permutations to 1000 to make sure the permutation is thorough enough to support further analysis.



(a) The means of the original simulated data



(b) Estimation results based on permutation test with the standard deviation of data set as 0.05.



(c) Estimation results based on permutation test with the standard deviation of data set as 2.

Figure 3.3: The results of the estimation of the functional mean and of the group-wise mean effects through the permutation test, using a sample size of 24 and a quantile $q = 1 - \alpha = 90\%$. The results for variate 1 only, as an example, and for two different noise levels are reported.

Results of FANOVA test methods. After generating the simulated data, we designate the first group with index $g = 0$ as the control group. This group serves as a reference, and we compare the other groups to it in a sequential manner. We apply the functional model introduced in the previous sections and estimate the functional parameters, which

include the common mean and the additional group-effect functions. Subsequently, we perform the permutation tests to identify the time zones where the typical group-wise pattern functions exhibit significant differences from the control group. The results for the first variate at different noise levels are presented in Figure 3.3. Additional results for the other two variates can be found in Appendix A.4.

From the results, we can observe that the permutation method is capable of detecting the major properties of the patterns even when larger noise levels are present, despite losing the smoothness due to noise interference. Essentially, our permutation-based FANOVA method effectively identifies the correct underlying patterns and pinpoints the time zones where the group means significantly differ.

Sensitivity analysis results for key parameters. In evaluate the performance of our proposed method under different values of primary parameters (i.e., Gaussian noise level, sample size, and quantile of the permutation test), we defined three indices: *pattern dissimilarity*, *significant time zone matching rate* and *multi-classification correctness*. Specifically, pattern dissimilarity P_D is computed as the point-wise Euclidean distance between the simulated and estimated means:

$$P_D = \sum_{g=1}^{G} \int_0^T ||\boldsymbol{\alpha}_g(t) - \hat{\boldsymbol{\alpha}}_g(t)||_2 dt$$

where $\boldsymbol{\alpha}_g(t) = [\alpha_{1,g}(t), \ldots, \alpha_{D,g}(t)]^\intercal$, and similarly $\hat{\boldsymbol{\alpha}}_g(t) = [\hat{\alpha}_{1,g}(t), \ldots, \hat{\alpha}_{D,g}(t)]^\intercal$. Figure 3.4a displays the values of P_D independent of parameter values.

The significant time zone matching rate S_{time} quantifies the accuracy of detecting time zones, where the group-wise typical patterns significantly differ, are correctly detected. Using $sign(\cdot)$ to denote the sign of a function, and ν to denote the 1-dimensional Lebesgue measure, we obtain

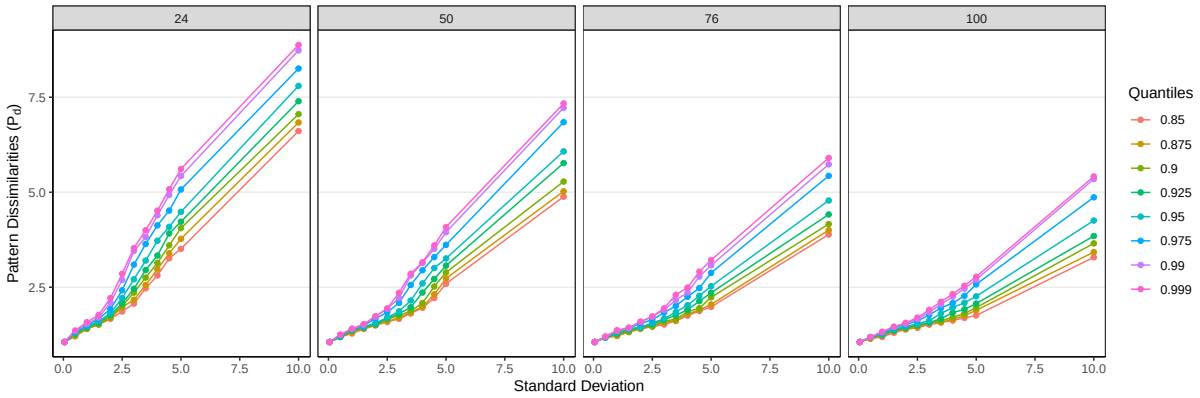
$$S_{time} = \sum_{g=1}^{G} \frac{\nu\{t \in [0,T] \mid sign(\boldsymbol{\alpha}_g(t)) = sign(\hat{\boldsymbol{\alpha}}_g(t)), \boldsymbol{\alpha}_g(t) \neq 0, \hat{\boldsymbol{\alpha}}_g(t) \neq 0\}}{\nu\{t \in [0,T] \mid \boldsymbol{\alpha}_g(t) \neq 0\}}$$

Therefore S_{time} represents the total proportion of time in which the estimated and true group patterns share the same sign and are not null. Figure 3.4b illustrates the variability of S_{time} with respect to the main parameters.

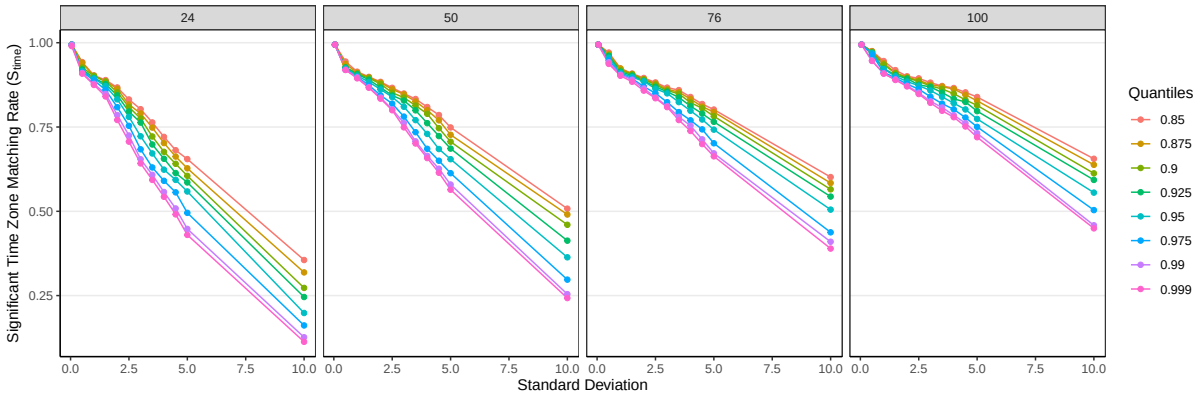
Eventually, we performed automatic multi-class classifications of the simulated dataset, utilizing a standard multinomial logistic regression approach based on "agreement scores" between the simulated data and identified group-specific patterns (see Chapter 4 for further details). The multi-classification correctness index shows the proportion of accurately classified samples for all three variates in a test set. Figure 3.4c presents the results of multi-classification correctness.

The findings reveal that, larger sample sizes enhance the detection of groups' typical pattern functions, particularly when noise levels are high. Nevertheless, with samples larger than 50 units, only minor improvements are observed, indicating that relatively small sample sizes already enable accurate pattern detection. The selection of a high quantile in the permutation test generally reduces the detection accuracy of both the mean typical pattern shape and the time zones of significant group-wise difference. Classification correctness mostly decreases when the quantile increases, although some deviations from this trend apear for small samples. Therefore, for small samples, specific choices regarding the permutation quantile must be made to optimize classification results.

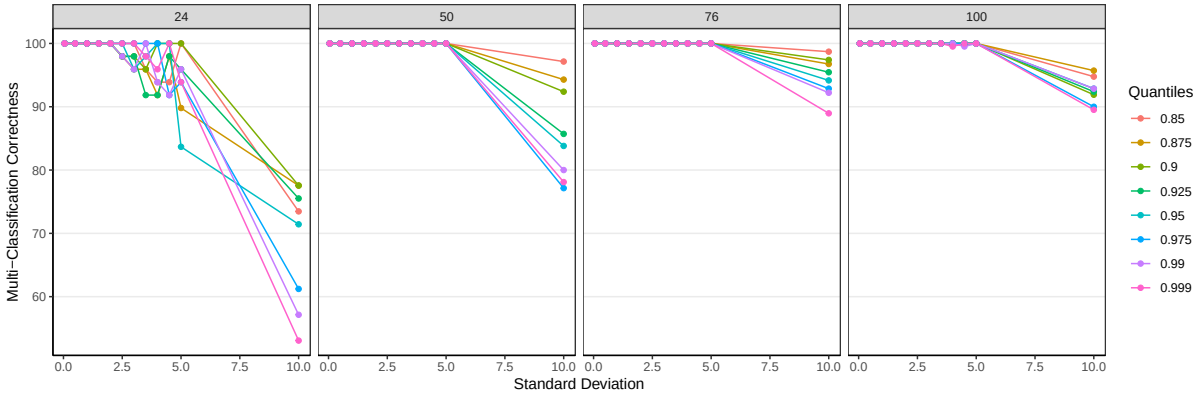
Focusing specifically on the case with a sample size of 24 for each group and a standard deviation of the Gaussian noise equal to 10, which closely resembles our driving case study, the highest multi-classification correctness rate remains around 80% when the quantiles are 0.9 and 0.875. This rate is acceptable considering it is a 7-group classification problem. As a result, in the following section, we choose a quantile of 0.9 for the permutation tests in the analysis of our motivating case study, since it provides better performance in pattern detection while maintaining the highest multi-classification correctness rate.



(a) Pattern Dissimilarities.



(b) Significant Time Zone Matching Rate.



(c) Multi-Classification Correctness.

Figure 3.4: Variability of the three quality indices with different parameters in the simulated data: standard deviation of the Gaussian noise, quantile level used in the permutation tests and sample size. The sample size is reported on the top of each picture.

3.3.2 Applications to Emotional Pattern Detection

The proposed methodology is primarily motivated and designed to detect human emotions from expression evolutions extracted from facial videos. Our analysis is grounded on the representative and open-sourced RAVDESS dataset, as described in Section 1.3. This dataset comprises seven different emotion types plus a neutral case, as represented in videos. We employ *action units (AU)*, continuous functions that encode specific facial muscle contractions. Our data possess a multivariate functional format, as multiple AUs evolve simultaneously and interdependently within a single video. The ultimate objective of this chapter is to identify the typical patterns characterizing expression evolution under various emotions of interest and to use this information as prior knowledge for better detecting latent emotions in unlabelled videos.

Results of functional modeling and FANOVA tests. Similarly to the study on simulations, after the registration, we canconstruct the FANOVA model and estimate the common mean functions and the additional group-wise typical mean patterns. Hereby, the common mean can encapsulate pronunciation and performance information, while the group-wise patterns are associated with specific emotional effects. Each AU is significantly activated or deactivated within particular time periods. Figure 3.5 shows the mean functions and detected group-wise patterns for the emotion angry and all action units. Additional plots for all emotions can be found in Appendix A.4.

A comparison of the detected overall common mean in the first column of Figure 3.5 with the sample mean curves for the neutral case in the first column of Figure 2.3 reveals a remarkably similar behavior, in particular for AU25 and AU26, which are related to the mouth in the lower part of the face. This observation confirms that the estimated overall means $\mu_{d,0}(t)$ represent the information related to the pronunciation of the sentence, which should be filtered out during the emotion recognition phase. Meanwhile, the detected mean group-wise patterns exhibit periods of strengthening and inhibition, corresponding to positive and negative curves in Figure 3.5, independence of each emotion. Therefore,

each AU's behavior is often unique under different emotion groups.

Examining our results more closely, Figure 3.6 illustrates the outcomes for AU06 (Cheek Raiser). Additional results for the other action units can be found in Appendix A.4. Figure 3.6a presents the estimated mean pattern for AU06 $\hat{\mu}_{d,0}(t)$ (in blue), compared with the typical emotional patterns $\hat{\mu}_{d,0}(t) + \hat{\alpha}_{d,g}(t)$ (in red), and the specific group-wise patterns (kernels) $\hat{\alpha}_{d,g}(t)$ (in green). Figure 3.6b displays the observed F-statistics curves (in red) alongside the point-wise 0.9 quantile (in blue) of the permutation distribution. The time intervals where the red curve exceeds the quantile arethe periods in which the patterns of the emotions significantly differ from the neutral case. We observe that the detected patterns of AU06 are distinct for different emotions, and their activation degrees are relatively high, indicating that AU06 is a crucial component for emotion detection.

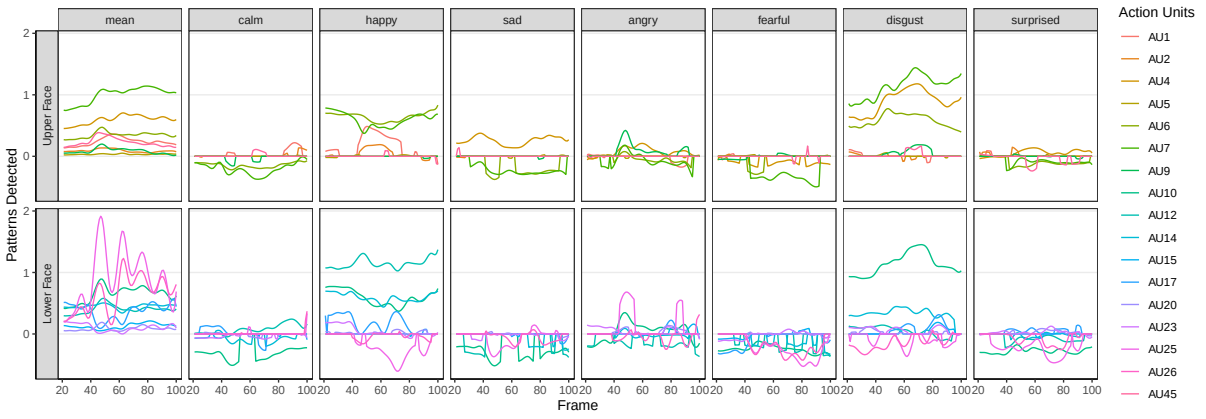
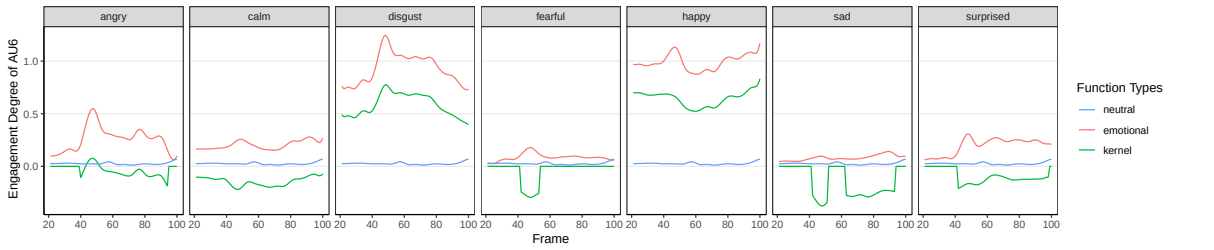
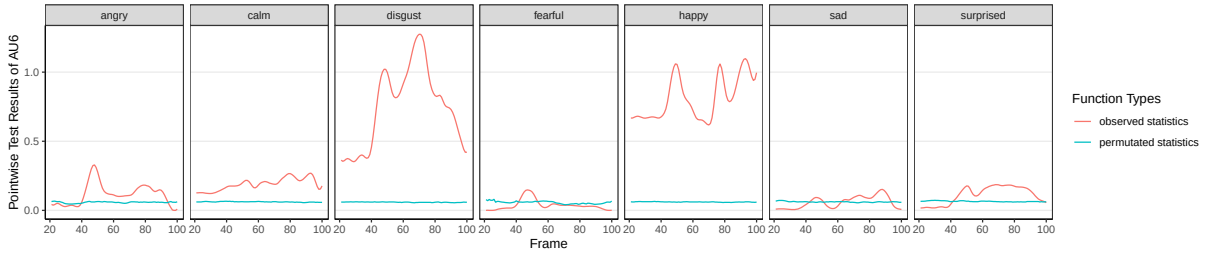


Figure 3.5: Mean and group-wise patterns detected for each emotion and action unit. The top row reports the patterns related to the upper part of the face, while the bottom row reports the patterns related to the lower part of the face.

Upon closer inspection of the results for one variate across all classes, Figure 3.7 depicts all variates for one class (emotion angry) in a heatmap version, while the results for other classes (emotions) can be found in Appendix A.4. The x-axis in the heatmap represents the range of frames (the time zone) of the facial video under investigation. As also portrayed in the column of emotion angry in Figure 3.5, many AUs oscillate between strengthening and inhibition with small amplitudes. AU25 (Lips part) is positively activated, possibly due to the stress of tones when people become angry, while AU12 (Lip Corner Puller) is



(a) Group-wise patterns estimated via group comparisons.



(b) FANOVA test results of observed and permutated statistics.

Figure 3.6: The Group-wise patterns of AU06 (Cheek Raiser) under neutral and seven emotions, together with the corresponding FANOVA test statistics. We present the estimated mean pattern for AU06 $\hat{\mu}_{d,0}(t)$ (in blue), compared with the typical emotional patterns $\hat{\mu}_{d,0}(t) + \hat{\alpha}_{d,g}(t)$ (in red), and the specific group-wise patterns (kernels) $\hat{\alpha}_{d,g}(t)$ (in green)

negatively activated as an angry person usually pushes down the lip corners.

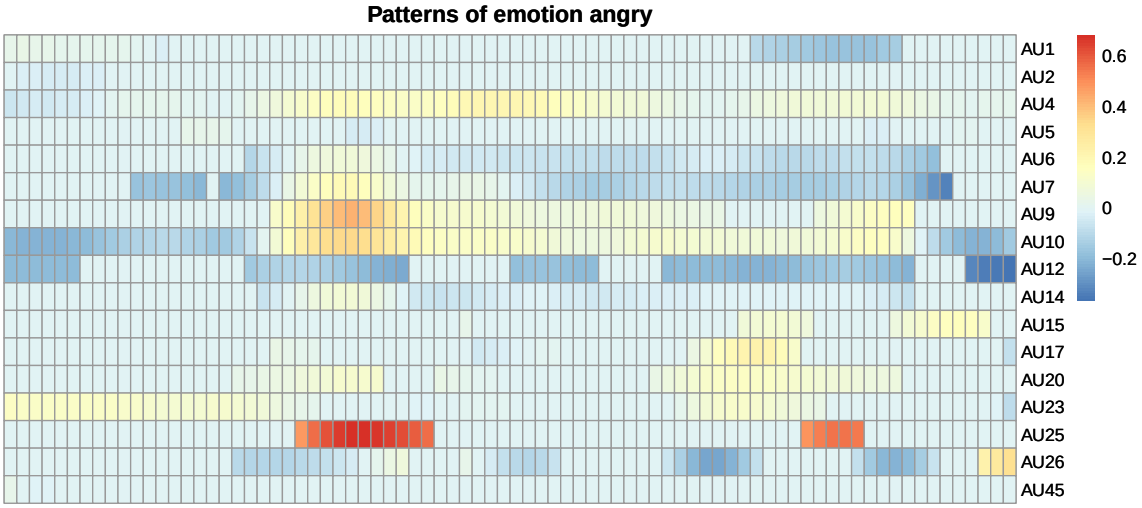


Figure 3.7: Heatmap of group-wise patterns detected for emotion angry.

For other emotions, Table 3.1 summarizes the significant deviation from the neutrality of all the action units. The two effects of strengthening or inhibition are further char-

acterized according to the different durations of the effect, labeled as short (S), medium (M), and long (L), as illustrated in the column titles in Table 3.1. We define an activation/deactivation for a set of consecutive time instants as a short-period (S) when it lasts less than 20% of the time range; a medium-period (M) when it lasts between 20% and 80% of the time range; a long-period (L) when it lasts more than 80% of the time range. Similar to the example of the emotion angry, we found that for emotions happy and disgust, many action units exhibit a strengthening effect over a large time range. Emotion sad sometimes induces a stronger inhibition of action units than in the neutral case. Emotion fearful has a more significant impact on the upper half of the face (brows, eyelids, and nose), while emotion calm is more related to the center of the face (Cheek Raiser, Lid Tightener, and Lip Corner Puller). Emotion surprised is the only emotion where AU45 is significantly influenced.

Table 3.1: Emotions and their corresponding influences on the action units

Emotions	**Selected Action Units**					
	Strengthened(S)	*Inhibited(S)*	*Strengthened(M)*	*Inhibited(M)*	*Strengthened(L)*	*Inhibited(L)*
Calm	02,05,26,45	09,15,17,20	01	10, 14,23	12	06,07
Happy	01,02,05,20,26	02,05,09,15,23	17,23	25	06,07,10,12,14	
Sad	05,20,25	02,05,09,12,14,15,17,20,23,25,26		06,10,25	04	07
Angry	01,04,06,07,09,14,15,17,25,26	01,02,23,26	04,09,10,20,23	06,07,10,12,14		
Fearful	05,09,15,17,23	02,06,10,12,14,15,17,20		04,07,09,23,25,26		
Disgust	02,05,15	02 ,05,12,25	09,12,17,20,23	26	04 ,06, 07,10,14	
Surprised	02, 09,17,20	09,14,15,20,23,45	04,23	06,10,12,25,26	07	

Short-period (S): happened less than 20% of time range; Medium-period (M): happened 20% $\sim$ 80% of time range; Long-period (L): happened more than 80% of time range.

Chapter 4

Agreement Scores Generation via Group-wise Patterns and Score-based Multi-class Classification

Statistical modeling predominantly serves as a powerful instrument for delineating data structures, elucidating causal relationships, and facilitating empirical predictions. In the preceding chapter, we established a methodology to characterize the mean behavior of a collection of multivariate curves and identify the prototypical patterns within distinct groups. These group-wise patterns represent abstract inter-group insights, derived from data and filtered through domain-dependent features specific to each group. The comprehensive framework of group-wise functional data analysis is furthered by addressing the third task, multi-classification, which involves classification problems with more than two classes. Given a set of new functional data sharing similar structures with our training set, our objective is to automatically categorize this set of curves into a target group by analyzing the information gleaned from the identified mean group-wise patterns. We employ and compare various methods to accomplish the multi-classification task and assess their predictive capabilities. In the context of our driving case study, the classification component targets automatic emotion recognition.

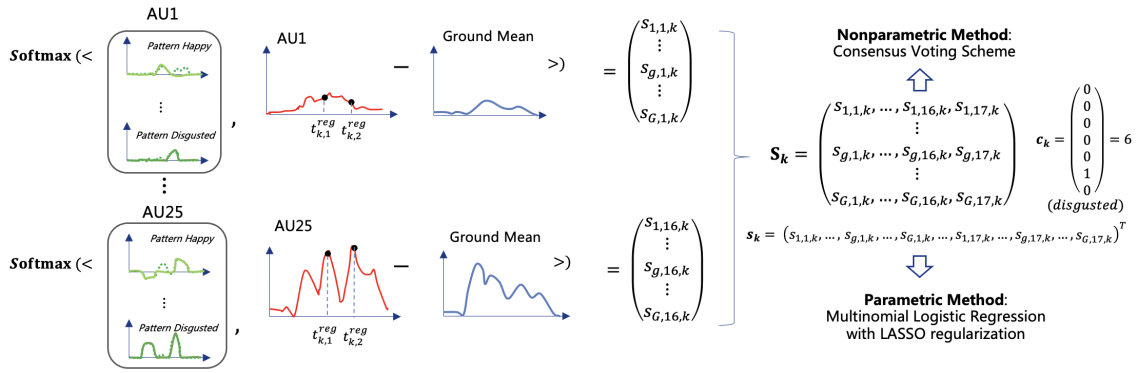


Figure 4.1: Flowchart Illustrating Agreement Scores Generation via Group-wise Patterns and Score-based Multi-class Classification.

The structure of this chapter is described below and is also illustrated in the flowchart shown in Figure 4.1. In Section 4.1, we describe the process of generating agreement scores based on group patterns, which serve as the extracted domain knowledge. Subsequently, we explore an ensemble strategy utilizing a consensus voting framework in Section 4.2, where each variate casts a vote for the class(es) with the highest score(s), and the final decision is reached through the consensus of all variates by identifying the class(es) with the majority of votes. In Section 4.3, we treat the entries of the vectorized score matrix as predictors in regular multinomial logistic regression; we also introduce the corresponding sparse version by employing LASSO penalized learning and illustrate the outcomes at the end about the relevant features of each group.

4.1 Agreement Scores Based on Group-wise Patterns as Domain Knowledge

The process of generating a vector or other low-dimensional representations of data in machine learning is known as *embedding*. An effective embedding ideally captures the semantics of the input, positioning semantically similar inputs in close proximity within the embedding space [73, 119]. In this context, our objective is to utilize the filtered

domain-dependent group-wise patterns to transform the information contained within an observed function into a compact representative vector in the embedding space.

4.1.1 The Concept of Agreement Scores

To achieve an effective embedding, we propose to map the observed functions to a representation space, constructed on the basis of the degrees of similarity between a sample and the identified group-wise patterns. We define the resulting representative vector as the set of *group agreement scores* that express the similarity between the observed function and the group patterns. The group patterns provide insights into the typical group-wise behavior or shape of the functional data. Therefore, using the group-wise patterns as anchors, we can measure the similarity degrees of observed functions (function-to-group distances) across all groups. The agreement scores serve as semantically interpretable products resulting from the embedding technique. By positioning inputs that share semantic similarities closer together in the embedding space, the projection of functions (agreement scores) with the same group membership will be situated closer in the score vector space.

For a new sample, its "agreement scores" are determined by computing the L_2 inner product (that is the projection) of its set of curves with the group-wise patterns. In a specific dimension d, the function of one sample, after being projected on the set of G group-wise patterns, yields a G-dimensional vector of agreement scores. Consequently, a multivariate set of D curves generates a $G \times D$ matrix of scores, representing the sought-after low-dimensional representation space. These generated scores are subsequently utilized as inputs for the classification methods.

We choose this strategy to compute the scores since in this way we are giving higher scores to the functions that show a shape similar to the group-wise patterns exactly in the time intervals where the patterns are not null, that is in the time zones that we detected as more significant. Using other strategies, like replacing the L_2 inner product with the convolution, would make us loose the focus on the time component of the problem.

4.1.2 Generating Agreement Scores Based on Group Patterns for Classification Methods

The procedure for generating agreement scores, based on the estimated group patterns, comprises two primary stages: centralization and measurement. These stages are depicted in Figure 4.2.

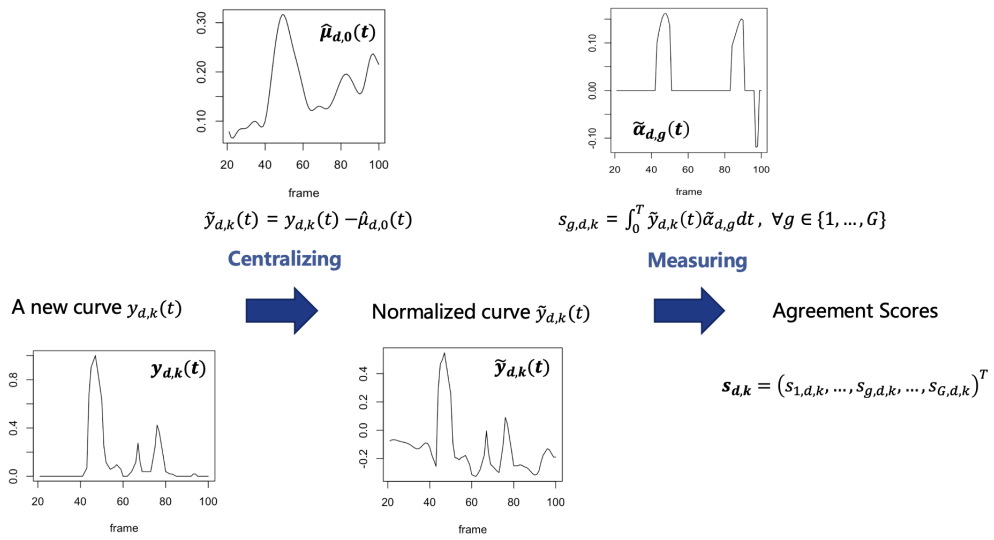


Figure 4.2: Scores building process based on the estimated group patterns.

In the centralization stage, the initial step involves centralizing the input function $y_{d,k}(t)$ by subtracting the estimated mean function $\hat{\mu}_{d,0}$ for each sample k and variate d. This process filters out information that is independent of the group to which the data belongs. The resulting normalized function $\tilde{y}_{d,k}(t)$ is given by:

$$\tilde{y}_{d,k}(t) = y_{d,k}(t) - \hat{\mu}_{d,0}(t). \tag{4.1}$$

After centralization, the measurement stage computes the agreement values between the normalized function and the set of G group-wise patterns using the L_2 inner product. This inner product gauges the agreement of the considered data $\tilde{y}_{d,k}(t)$ with each $\tilde{\alpha}_{d,g}$, quantifying their agreement degree for each group g. Practically, this is achieved by calculating the regular sums of element-wise products between every function and the

corresponding group-wise patterns, in the same dimension of the function, for all classes. Let $s_{g,d,k}$ denote the agreement score given to group g on variate d for sample k. Then we have:

$$s_{g,d,k} =< \tilde{y}_{d,k}(t), \tilde{\alpha}_{d,g}(t) >= \int_0^T \tilde{y}_{d,k}(t)\tilde{\alpha}_{g,d}(t)dt, \tag{4.2}$$

where $< \cdot >$ denotes the inner product between two functions in L_2.

The softmax function (4.3)

$$\tilde{s}_{g,d,k} = \frac{\exp(s_{g,d,k})}{\sum_{g=1}^{G} \exp(s_{g,d,k})} \in [0,1]. \tag{4.3}$$

is further used to normalize a vector of scores across all groups for each variate d and for each sample k. This normalization yields the relative percentages across groups, thus $\tilde{s}_{g,d,k},\ g = 1, \ldots, G$ represent the probabilities of assigning sample k to each of the G groups in dimension d. The softmax function's ability to handle negative or large values in the input vector ensures that the resulting probabilities are well-defined and satisfy the constraint of summing to one. The larger input values in the vector correspond to larger probabilities, thereby conveying useful information about the relative importance of each group for the given sample and variate.

We proceed to consider the normalized scores $\tilde{s}_{g,d,k}, k = 1, 2, ..., K$ as the k-th realization of the score variate $\tilde{s}_{g,d}$. Consequently, the score vector for a specific variate d encompassing G groups, is denoted as $\tilde{\mathbf{s}}_d = [\tilde{s}_{1,d}, \ldots, \tilde{s}_{g,d}, \ldots, \tilde{s}_{G,d}]^\intercal$. This vector representation of group membership distribution elucidates our first classification method, the consensus algorithm implemented through a voting scheme.

Moreover, we represent the vectorized version of matrix $\tilde{\mathbf{S}}$ as $\mathbf{s} = [s_1, \ldots, s_{GD}]^\intercal = [\tilde{s}_{1,1}, \ldots, \tilde{s}_{1,D}, \ldots, \tilde{s}_{g,1}, \ldots, \tilde{s}_{g,D}, \ldots, \tilde{s}_{G,1}, \ldots, \tilde{s}_{G,D}]^\intercal$. We will employ this representation of the feature set for sparse MLR and softmax regression methodology. Considering the K realizations of $\mathbf{s}$ as the sample set $\mathbf{s}_k,\ k = 1, ..., K$, and denoting by $c_k \in \{1, \ldots, G\},\ k =$

$1, ..., K$ the true group membership of the sample, we obtain a comprehensive representation of the data.

4.2 Consensus Algorithm via Voting Scheme

The consensus-based voting scheme, employed in the context of multivariate multi-classification tasks, utilizes a collective decision-making approach for classifying instances based on multiple input variables. Each individual classifier predicts the class of an instance using its respective input features. These predictions are then aggregated by a voting algorithm that considers the combined output from all classifiers to determine the final class of the instance. In the consensus algorithm, the independent votes from all variates are combined under a non-parametric setting. Various strategies can be employed to amalgamate the predictions of individual classifiers, such as the majority voting rule, where the class with the most votes is selected, or the weighted voting rule, which weighs each classifier's vote according to their performance or expertise.

This voting scheme is particularly advantageous when input features are intricate and challenging to interpret, like the expression evolutions in our case study, as it can bolster the precision and resilience of classification outcomes. Specifically, our approach begins by examining the column vector $\tilde{s}_d$ of the score matrix for one variate, comparing the scores of one specific variate across all groups. Each variate then votes for a class based on its maximum score in the column vector. The final decision is made by gathering votes from all variates and choosing the class with the majority of votes. The primary objective of the consensus-based voting scheme is to harness the collective knowledge and input of AU classifiers to enhance the accuracy and robustness of classification decisions. This process is illustrated in Algorithm 1.

We applied the method to our driving case study. The distribution of votes across different emotions for each AU is presented in Table 4.1. The last column of this table displays the correctness rates in multi-class classification, which are the ratios of correctly

classified samples to the total sample set. AU20, AU26, AU9, AU25, and AU5, in descending order, can be considered the top 5 classifiers with the highest correctness rates in the voting scheme. Notably, AU20 (Lip stretcher), AU25 (Lips part), and AU26 (Lips part) are associated with mouth movements, whereas AU5 (Upper Lid Raiser) relates to the eyes, and AU9 (Nose Wrinkler) pertains to the nose. These AUs, exhibiting higher classification rates, suggest that they encompass crucial local facial region information for expressing emotions.

Algorithm 1: Consensus Algorithm for multivariate multi-class classification

input: Dataset containing G labelled groups (real labels) with K samples in each group $\{\tilde{S}_k, y_k\}_{k=1,\dots,GK}$

output: Prediction of labels $\hat{y}_k$ given $\{\tilde{S}_k\}$ for $k \in \{1, \dots, GK\}$

initialization: $\#group_1 = \cdots = \#group_g = \cdots = \#group_G = 0$

for *sample* $k = 1$ *to* GK **do**

  for *variate* $d = 1$ *to* D **do**

    Compute the vote of each variate: $v_{k,d} = \arg_g \max \tilde{s}_{k,d,g}$

    for *group* $g = 1$ *to* G **do**

      count the votes for each group: **if** $v_{k,d} == g$ **then**
        $\#group_g = \#group_g + 1$

      end

    end

  end

  get the counts of the votes among G groups per sample:

  $\{\#group_1, \dots, \#group_g, \dots, \#group_G\}$

  Determine the class(es) with highest number of votes per sample:

  $samplevote_k = \arg_g \max\{\#group_1, \dots, \#group_g, \dots, \#group_G\}$

  If the classes with the highest counts are more than one, randomly sample one from the selected classes: **if** $length(samplevote_k) > 1$ **then**
    $samplevote_k = \text{sample}(samplevote_k)$

  end

  The predicted label per sample: $\hat{y}_k = samplevote_k$

end

Additionally, Table 4.2 showcases the distribution of votes for all AUs, compared with the actual labels of the corresponding samples from which AU curves are derived. Predominantly, a single AU targets only one or two groups, indicating that this voting scheme may not guarantee robust classification performance. For instance, AU45 allocates

Table 4.1: Distribution of votes on different emotions of each AU. The last column is correctness in multi-class classification, i.e., the ratio obtained as the size of correctly classified samples over the entire sample size.

	calm	happy	sad	angry	fearful	disgusted	surprised	total correctness rate
AU1	0.25	0.50	0.00	0.71	0.04	0.17	0.00	0.238
AU2	0.33	0.54	0.00	0.42	0.00	0.29	0.08	0.238
AU4	0.00	0.00	0.04	0.04	0.71	0.71	0.00	0.214
AU5	0.12	0.46	0.46	0.46	0.04	0.38	0.08	0.286
AU6	0.08	0.71	0.75	0.04	0.00	0.29	0.00	0.268
AU7	0.25	0.04	0.00	0.04	0.62	0.79	0.00	0.250
AU9	1.00	0.00	0.00	0.79	0.00	0.42	0.00	0.315
AU10	0.71	0.12	0.00	0.00	0.04	0.96	0.12	0.280
AU12	0.08	1.00	0.04	0.00	0.67	0.00	0.00	0.256
AU14	0.17	0.83	0.00	0.00	0.67	0.00	0.08	0.250
AU15	0.42	0.00	0.00	0.25	0.25	0.25	0.21	0.196
AU17	0.17	0.42	0.04	0.04	0.88	0.29	0.04	0.268
AU20	0.71	0.17	0.17	0.42	0.29	0.33	0.17	0.321
AU23	0.38	0.38	0.04	0.17	0.67	0.21	0.00	0.262
AU25	0.00	0.12	0.17	0.75	0.62	0.00	0.33	0.286
AU26	0.46	0.04	0.12	0.12	0.88	0.50	0.12	0.321
AU45	0.00	0.04	0.08	0.00	0.17	0.38	0.96	0.232

all its votes to the surprised emotion, irrespective of the real labels of the samples. While AU45 exhibits a high correctness rate for the surprised emotion in Table 4.1, its predictive capability as a classifier in this voting scheme may be limited.

Table 4.2: The distribution of votes from all AUs, given the real labels of the corresponding samples where AU curves are collected. The AUs marked in bold are those who vote correctly in the multi-class classification task.

real label	predicted label						
	calm	happy	sad	angry	fearful	disgusted	surprised
calm	**AU2,9,10,15,20,26**	AU1,5,12	AU6	AU25	AU4,7,14,17,23		AU45
happy	AU9,15,20	**AU1,2,5,6,12,14,23**			AU4,17,25	AU7,10,26	AU45
sad	AU9,10,15,20		**AU5,6**	AU1,2	AU7,12,14,17,23,25,26	AU4	AU45
angry		AU23	AU6	**AU1,2,5,9,20,25**	AU4,7,12,14,17, 26	AU10,15	AU45
fearful	AU9,10,15		AU6	AU1,2,5,20	**AU4,7,12,14,17,23,25,26**		AU45
disgust		AU6,14,23		AU1,2,5	AU12,17,25	**AU4,7,9,10,15,20,26**	AU45
surprised	AU9,10,15	AU2,17	AU6	AU1,5,20	AU4,7,12,14,23,25,26		**AU45**

4.3 Sparse Multinomial Logistic Regression for Multi-class Classification

For multi-class problems where the response variable is nominal, discrete, and devoid of a natural order in its values, Multinomial Logistic Regression (MLR) serves as an ap-

propriate method based on the multinomial distribution [18]. MLR extends the binary logistic regression framework, enabling the estimation of probabilities for various possible outcomes of the nominal response variable, through linear combinations of observed features and problem-specific parameters. Although the MLR method is not novel for multi-class classification, it is noteworthy that MLR eliminates the need for additional efforts to resolve multiple binary classification problems using the "one-versus-all" strategy and similar heuristics frequently employed in other techniques [96].

4.3.1 The Theoretical Setting of Multinomial Logistic Regression

Recall that we denoted in Section 4.1.2 the k-th realization of $\mathbf{s}$ as $\mathbf{s}_k$, with $c_k \in \{1, \ldots, G\}$ representing the true group membership of the k-th sample. Here we redefine $\mathbf{s} = [1, s_1, \ldots, s_{GD}]^\intercal$, adding a first component equal to 1, in order to simplify the notation used in the regression. For a multiclass classification problem with G classes, we have then a set of samples $\{(\mathbf{s}_k, c_k)\}_{k=1}^K$, where $\mathbf{s}_k$ is the feature vector for the k-th sample and $c_k \in \{1, 2, \ldots, G\}$ represents the true class label for the k-th sample.

Following the standard MLR settings, we define $\pi_g(\mathbf{s}_k) = \mathbb{P}(c_k = g|\mathbf{s}_k)$, the conditional probability of sample k belonging to the class g given its feature set $\mathbf{s}_k$. For the classical logistic function, all of its outputs are between between 0 and 1 which makes it a natural candidate to model the probability distribution in a multi-class problem. In the asymmetric multinomial regression model parametrization, it is typical to set in advance a reference class denoted by G and listed as the last one for notational simplicity. The aim of setting the reference class is to model the $G-1$ logits, the probabilities of assigning sample label c_k to the other $G-1$ classes relative to the reference class, using the linear combination of the feature set $\mathbf{s}_k$ and the regression parameters $\{\mathbf{b_g}\}_{g=1}^G, \mathbf{b_g} = [b_{g,0}, b_{g,1}, \ldots, b_{g,GD}]^\intercal \in \mathbb{R}^{(G+1)D}$. That is,

$$\log \frac{\pi_g(\mathbf{s}_k)}{\pi_G(\mathbf{s}_k)} = \mathbf{b_g}^\intercal \mathbf{s}_k, g = 1, \ldots, G-1. \tag{4.4}$$

We will further arrange $\mathbf{b_g}$ into the parameter matrix $\mathbf{B} = [\mathbf{b_1}, \ldots, \mathbf{b_G}]$ having dimension $(G+1)D \times G$. Equation 4.4 can also be written in the following format, via proper transformations:

$$\pi_g(\mathbf{s}_k) = P(c_k = g|\mathbf{s}_k) = \frac{\exp(\mathbf{b_g}^\intercal \mathbf{s}_k)}{1 + \sum_{h=1}^{G-1} \exp(\mathbf{b_h}^\intercal \mathbf{s}_k)}, \text{ with } \mathbf{b_G} = \mathbf{0}. \tag{4.5}$$

As mentioned in the previous chapter (Section 3.3 specifically), as a first attempt we used the asymmetric MLR without any regularization term to produce the multi-class classification correctness index, in order to measure the pattern detection's efficiency in the simulated data. The purpose at that moment was not to choose the optimal classification method, but to identify whether the detected patterns were able to capture the key group-wise differences. In the following, we will explore if other classification methods may bring improved results, in particular for our driving case study of emotion recognition.

As an alternative extension of binary logistic regression, softmax regression solves multi-class problems by employing the softmax function (introduced also in Section 4.1.2). The softmax function yields the relative percentages across groups, therefore taking properly the role of the link function to construct logistic regression for the conditional probability. Therefore we also have an alternative model:

$$\pi_g(\mathbf{s}_k) = \mathbb{P}(c_k = g|\mathbf{s}_k) = \frac{\exp(\mathbf{b_g}^\intercal \mathbf{s}_k)}{\sum_{h=1}^{G} \exp(\mathbf{b_h}^\intercal \mathbf{s}_k)}. \tag{4.6}$$

The normalization involved in the softmax function ensures that the sum of the output of all classes is equal to one $\sum_{g=1}^{G} \pi_g(\mathbf{s}_k) = 1$. We call this format of MLR with the softmax function the "symmetric" version of parametrization, as it is invariant to permutations of the classes, while the previous version, with a pre-fixed reference class, is called "asymmetric" version. Additionally, since the resulting estimator of the MLR method is invariant to permutations of the classes, the symmetric parametrization without a pre-

fixed reference class is preferred [91]. However, the symmetric parametrization based on the softmax function is not estimable without constraints, because the optimization problem that must be solved to estimate the parameters in (4.6) has not a unique solution, because of the collinearity and the high number of covariates. Nevertheless, with the LASSO penalization considered in parameter estimation, this ambiguity is solved in a natural way [70].

4.3.2 Sparse MLR classifier under regularized regression

Considering the vectorized version $\mathbf{s}$ of the score matrix $\tilde{\mathbf{S}}$ as the set of input variates under study, we encounter potentially correlated variates since each column of the matrix $\tilde{\mathbf{S}}$ sums up to 1. Moreover, the number of variates GD, in our driving case study, exceeds the sample size K.

To address the ill-conditioning problems arising from the limited sample size and correlated variates, we incorporate sparse feature selection using the LASSO regularization for multinomial logistic regression for multi-classification [49]. The Least Absolute Shrinkage and Selection Operator (LASSO) regression [147] promotes sparsity in MLR classifiers by regularizing the likelihood of the training data with a penalty on weights [23]. Large penalization weights in the loss function lead to fewer selected variates with nonzero weights, automatically obtaining meaningful features in the high-dimensional space [96]. Furthermore, LASSO enhances classification accuracy by being less sensitive to multicollinearity issues, as it tends to select one of the highly correlated features while shrinking the others to zero weights [114, 177]. Therefore, the regularization approach effectively tackles all the issues mentioned above, ultimately enhancing the performance and interpretability of the multi-class classification model.

Our goal is then to estimate the parameters in matrix $\mathbf{B}$ such that the modeled probability function is as close as possible to the true one. Given the training set $\{(\mathbf{s}_k, c_k)\}_{k=1}^{K}$, we consider one-hot transformation which maps the class labels c_k to a one-hot vector

representation $\mathbf{c}_k \in \{0,1\}^G$. The one-hot transformation can be defined as follows:

$$c_{k,g} = \begin{cases} 1 & \text{if } c_k = g, \\ 0 & \text{otherwise,} \end{cases} \tag{4.7}$$

where $c_{k,g}$ is the g-th element of the one-hot vector $\mathbf{c}_k$, and $g \in 1, \ldots, G$. This transformation is applied to each class label c_k of the samples, resulting in a set of one-hot vectors $\{\mathbf{c}_k\}_{k=1}^K$. The categorical cross-entropy loss function can be expressed as:

$$L(\mathbf{B}) = -\frac{1}{K} \sum_{k=1}^{K} \sum_{g=1}^{G} c_{k,g} \log(\pi_g(\mathbf{s}_k)), \tag{4.8}$$

where K is the number of samples, G is the number of classes, $\pi_g(\mathbf{s}_k)$ is the predicted probability for sample k and class g, and $\log(\pi_g(\mathbf{s}_k))$ is the log-likelihood function of the multinomial logistic regression. $\mathbf{B}$ is the matrix of parameters with dimensions $G \times (G+1)D$. To learn a sparse MLR classifier using LASSO regression, we introduce the penalty term into the loss function of the multinomial logistic regression. The regularized loss function becomes:

$$L_{\text{LASSO}}(\mathbf{B}) = -\frac{1}{K} \sum_{k=1}^{K} \sum_{g=1}^{G} c_{k,g} \log(\pi_g(\mathbf{s}_k)) - \lambda \sum_{g=1}^{G} \sum_{j=1}^{GD} |b_{g,j}|, \tag{4.9}$$

where GD is the number of features, λ is the regularization parameter controlling the trade-off between fitting the data well and keeping the model sparse. A larger value of λ would result in more sparsity and fewer selected features, while a smaller value would lead to less sparsity and potentially overfitting. $|b_{g,j}|$ denotes the absolute value of the coefficient for feature j in class g. The LASSO penalty term $\lambda \sum_{g=1}^{G} \sum_{j=1}^{GD} |b_{g,j}|$ enforces sparsity by shrinking the less relevant coefficients towards zero. The objective function can be written as:

$$\hat{\mathbf{B}} = \arg\max_{\mathbf{B}} \left[L_{\text{LASSO}}(\mathbf{B}) \right]. \tag{4.10}$$

To minimize the regularized loss function and update the weights, we employ a "proximal Newton" algorithm for optimizing this criterion, also referred to as iteratively reweighted penalized least squares [71]. This method is implemented in the R package "glmnet" [69] and involves iteratively using a quadratic approximation to the log-likelihood, followed by performing weighted coordinate descent on the corresponding penalized weighted least-squares problem for each class. Subsequently, the algorithm iterates through all classes. The proximal Newton approach is particularly effective for LASSO regularization, as it can directly estimate non-zero weights while setting the remaining weights to zero.

4.3.3 Model Training and Performance Evaluation

In this section, we present the results of our analysis, focusing on the performance of the regular MLR model under asymmetric parametrization and the sparse MLR model under symmetric parametrization. Our main objectives are to evaluate the multi-class classification correctness and to identify the relevant features that contribute to the correct identification of different classes.

Model Performance

The 168 samples that we used from the RAVDESS dataset are split into 75% for training (126 samples) and 25% for validation (42 samples). We train the regular MLR model using the R package "multinom" [131] to stay coherent with the previous chapter and serve as the baseline. The regular MLR model achieved 100% correctness on the training set and 42.86% on the validation set.

The symmetric sparse MLR is trained using the R package "glmnet" [69]. To estimate the coefficient parameters and the regularization parameter λ, we used a 10-fold cross-validation. This method involves dividing the data into 10 subsets, training the model on 9 subsets, and validating it on the remaining subset, iterating 9 times. Figure 4.3 displays the validation set misclassification error versus the log-transformed regularization parameter $\log(\lambda)$ assessed through cross-validation. The average performance is used to

evaluate each regularized model, shown together with the upper and lower standard deviation. The optimal regularization parameter is chosen as 0.0202, under which the model's classification correctness for training and testing sets are 88% and 62%, respectively.

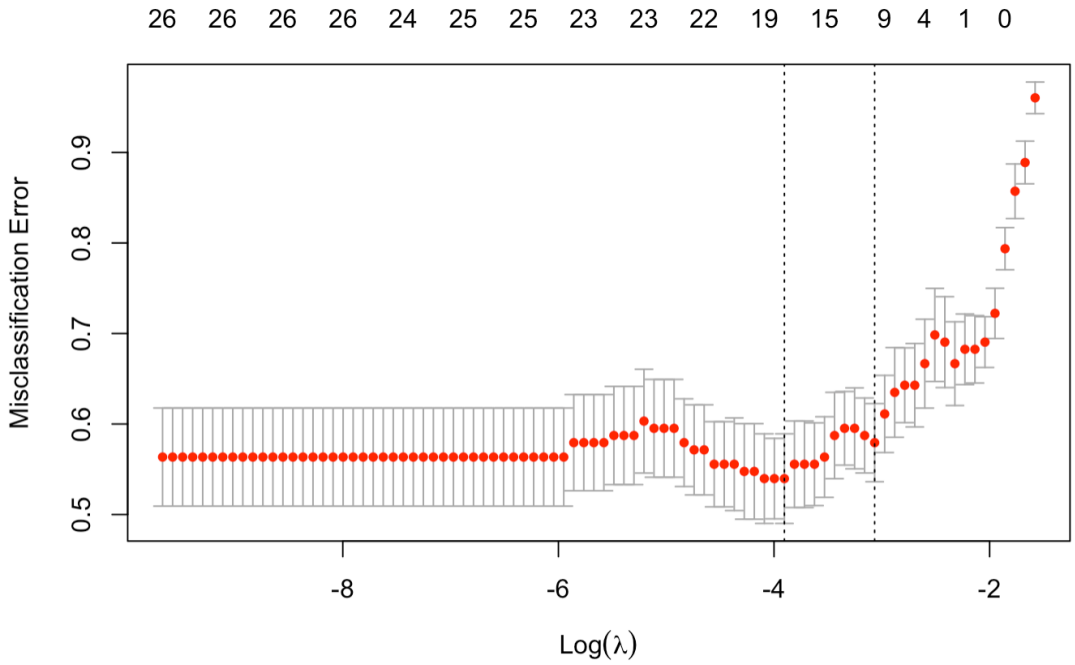


Figure 4.3: Training Set Misclassification Error vs. Log-Transformed Regularization Parameter log(λ) in Sparse Symmetric MLR Models

Table 4.3: Comparative Evaluation of MLR Models with Various Metrics

Table 4.4: Regular Asymmetric

Class	Precision	Recall	F1 Score	Accuracy
Calm	0.00	0.00	0.00	0.81
Happy	0.40	0.67	0.50	0.81
Sad	0.40	0.33	0.36	0.83
Angry	0.62	0.83	0.71	0.90
Fearful	0.38	0.50	0.43	0.81
Disgusted	0.75	0.50	0.60	0.90
Surprised	0.20	0.17	0.18	0.79

Table 4.5: Sparse Symmetric

Class	Precision	Recall	F1 Score	Accuracy
Calm	0.62	0.83	0.71	0.90
Happy	0.62	0.83	0.71	0.90
Sad	0.67	0.33	0.44	0.88
Angry	0.80	0.67	0.73	0.93
Fearful	0.55	1.00	0.71	0.88
Disgusted	0.75	0.50	0.60	0.90
Surprised	0.33	0.17	0.22	0.83

Table 4.3 presents a comparative evaluation of regular asymmetric MLR and sparse symmetric MLR, considering various metrics (precision, recall, F1 score, and accuracy) for multi-class classification. For all emotions, except for disgust, both precision and F1 score exhibit an improvement with the sparse symmetric MLR. In particular for the

calm emotion, all metrics show significant enhancement using the sparse symmetric MLR, whereas the metrics for disgust stay coherent. This phenomenon may arise because, in contrast to the calm emotion, disgust is an intense emotion characterized by long-term engaged patterns (as depicted in Figure 3.5), involving larger scores for numerous features, making it easier for the regular MLR model to accurately classify it, given all the features. In the case of sparse MLR, feature selection allows the model to concentrate only on the most crucial features for each class, thereby enhancing classification performance for all classes in which a few features are relevant for their recognition.

Feature Selection and Interpretation

Table 4.6 lists 89 variates selected using the optimal regularization parameter, together with their coefficients for the considered 7 classes. The g-th column in Table 4.6 indicates the features that affect the g-th classification. The result indicates that not only do the patterns in the correct class impact the prediction of one specific class, but the patterns of other classes also contribute to improving the model's prediction ability. From Table 4.6, we identify the top 3 AUs for each emotion, ranked by their weights in descending order, and display them in Table 4.7. Table 4.8 presents this information in a matrix format, offering a clearer visualization of the relationship between selected AUs and emotions. Notably, patterns of AU05 (Upper Lid Raiser) and AU20 (Lip stretcher) are crucial for classifying emotions, followed by AU17 (Chin Raiser), AU45 (Blink), AU12 (Lip Corner Puller), AU15 (Lip Corner Depressor), and AU26 (Jaw Drop), which have large impacts on at least two emotions. AU09 (Nose Wrinkler) is particularly important for the emotion calm, AU01 (Inner Brow Raiser) for the emotion disgusted, AU06 (Cheek Raiser) for the emotion happy, and AU04 (Brow Lowerer) for the emotion sad.

Table 4.6: In sparse MLR, 89 variates are selected under the best λ (0.0202), together with their coefficients for 7 classes.

	calm	happy	sad	angry	fearful	disgusted	surprised
(Intercept)	0.630	34.591	-24.874	1.480	7.526	8.478	-27.832
AU01 under patterns of emotion calm	.	.	.	-1.190	.	.	.
AU02 under patterns of emotion calm	0.588	.	.	.	.	.	.
AU04 under patterns of emotion calm	.	.	-0.724	.	2.112	.	-2.267
AU07 under patterns of emotion calm	0.980	.	.	.	.	.	.
AU09 under patterns of emotion calm	9.145	.	.	-10.853	.	.	.
AU10 under patterns of emotion calm	.	.	.	.	.	.	0.034
AU12 under patterns of emotion calm	5.394	.	.	.	.	.	.
AU14 under patterns of emotion calm	.	.	0.134	.	.	.	.
AU17 under patterns of emotion calm	2.321	-1.007	.	.	-0.452	.	.
AU20 under patterns of emotion calm	41.941	.	.	.	.	.	.
AU23 under patterns of emotion calm	1.067	-1.238	.	.	.	.	.
AU25 under patterns of emotion calm	.	.	6.443	2.859	-6.626	.	.
AU26 under patterns of emotion calm	1.693	.	.	.	.	.	.
AU45 under patterns of emotion calm	.	.	.	.	-12.341	.	.
AU04 under patterns of emotion happy	.	.	-0.087	.	0.086	.	-0.023
AU05 under patterns of emotion happy	.	.	.	.	.	-36.148	.
AU06 under patterns of emotion happy	.	1.614	.	.	.	.	.
AU07 under patterns of emotion happy	2.877	.	.	.	.	.	.
AU10 under patterns of emotion happy	.	0.294	.	.	.	.	.
AU12 under patterns of emotion happy	.	2.478	.	.	.	.	.
AU15 under patterns of emotion happy	.	.	.	.	-2.547	.	.
AU20 under patterns of emotion happy	.	76.504	.	.	.	.	.
AU23 under patterns of emotion happy	.	0.796	.	.	.	.	.
AU25 under patterns of emotion happy	.	.	-0.219	.	.	.	.
AU04 under patterns of emotion sad	.	.	2.889	.	.	.	.
AU05 under patterns of emotion sad	.	.	223.471	.	-11.439	.	.
AU06 under patterns of emotion sad	.	.	.	.	.	-0.235	.
AU07 under patterns of emotion sad	.	.	.	2.078	.	.	.
AU09 under patterns of emotion sad	.	.	.	.	.	.	-1.384
AU12 under patterns of emotion sad	.	.	.	.	3.606	.	.
AU14 under patterns of emotion sad	.	.	.	.	0.767	.	.
AU15 under patterns of emotion sad	.	.	7.131	.	-4.710	.	.
AU17 under patterns of emotion sad	.	-2.381	1.088	.	.	.	.
AU20 under patterns of emotion sad	.	.	.	-15.462	.	.	.
AU23 under patterns of emotion sad	.	-4.972	.	.	.	.	.
AU25 under patterns of emotion sad	2.028	.	.	.	-0.130	1.013	.
AU26 under patterns of emotion sad	.	.	0.154	.	.	.	-2.789
AU01 under patterns of emotion angry	.	.	.	0.150	.	.	.
AU05 under patterns of emotion angry	.	-66.193	.	44.086	.	.	.
AU06 under patterns of emotion angry	.	.	.	.	4.523	.	.
AU07 under patterns of emotion angry	.	.	.	.	.	1.334	.
AU09 under patterns of emotion angry	.	.	.	2.999	.	.	.
AU10 under patterns of emotion angry	.	0.165	.	.	.	.	-0.722
AU12 under patterns of emotion angry	-1.676	.	.	.	.	.	.
AU14 under patterns of emotion angry	.	.	.	.	.	.	3.371
AU15 under patterns of emotion angry	.	.	-5.566	3.408	.	.	.
AU17 under patterns of emotion angry	.	.	.	5.509	.	.	.
AU20 under patterns of emotion angry	.	.	.	1.499	.	.	.
AU23 under patterns of emotion angry	.	.	-7.476	9.042	.	.	.
AU25 under patterns of emotion angry	.	.	.	0.024	.	.	.
AU26 under patterns of emotion angry	.	.	-1.172	1.589	.	.	.
AU45 under patterns of emotion angry	.	.	.	-0.619	.	.	.
AU02 under patterns of emotion fearful	.	.	.	.	.	.	-2.544
AU04 under patterns of emotion fearful	.	.	.	.	.	-0.254	.
AU07 under patterns of emotion fearful	.	.	.	0.369	0.489	.	.
AU12 under patterns of emotion fearful	.	.	2.859	.	.	.	.
AU17 under patterns of emotion fearful	.	.	0.386	.	2.386	.	.
AU25 under patterns of emotion fearful	.	.	.	-0.284	.	.	.
AU26 under patterns of emotion fearful	-0.389	.	.	.	0.614	.	.
AU45 under patterns of emotion fearful	.	.	.	.	12.295	.	.
AU01 under patterns of emotion disgusted	.	.	.	.	.	5.581	.
AU02 under patterns of emotion disgusted	.	-5.678	.	.	.	1.780	.
AU04 under patterns of emotion disgusted	.	-0.147	0.774	.	.	0.329	.
AU06 under patterns of emotion disgusted	.	.	.	.	-0.207	.	.
AU09 under patterns of emotion disgusted	.	.	.	.	.	1.779	2.606
AU10 under patterns of emotion disgusted	.	.	.	.	.	2.336	.
AU12 under patterns of emotion disgusted	.	.	.	.	1.112	.	.
AU14 under patterns of emotion disgusted	.	.	.	3.791	.	.	4.389
AU17 under patterns of emotion disgusted	.	.	.	.	.	.	0.370
AU20 under patterns of emotion disgusted	.	.	.	.	.	9.115	-1.461
AU23 under patterns of emotion disgusted	-3.149	.	.	.	.	4.813	.
AU25 under patterns of emotion disgusted	.	.	5.553	.	.	.	-2.746
AU26 under patterns of emotion disgusted	.	.	.	.	.	3.028	.
AU02 under patterns of emotion surprised	.	.	.	9.367	.	.	-0.492
AU04 under patterns of emotion surprised	0.843	.	.	.	.	.	-0.103
AU05 under patterns of emotion surprised	.	-194.651	.	.	.	.	243.320
AU06 under patterns of emotion surprised	.	.	5.418	.	2.986	-2.527	.
AU07 under patterns of emotion surprised	.	.	.	.	4.497	.	.
AU09 under patterns of emotion surprised	.	.	.	.	.	.	-4.615
AU10 under patterns of emotion surprised	0.554	.	.	.	.	.	.
AU12 under patterns of emotion surprised	.	.	.	.	.	2.408	.
AU14 under patterns of emotion surprised	-1.429	.	.	.	.	3.024	.
AU17 under patterns of emotion surprised	.	.	.	.	.	.	0.072
AU20 under patterns of emotion surprised	.	.	.	.	0.248	.	8.257
AU23 under patterns of emotion surprised	.	.	.	.	10.179	.	.
AU25 under patterns of emotion surprised	.	.	.	-2.181	.	.	1.039
AU26 under patterns of emotion surprised	.	.	.	-0.809	.	.	1.905
AU45 under patterns of emotion surprised	.	.	.	.	.	-0.167	7.170

Table 4.7: Top3 AUs per emotion

Emotion	#1	#2	#3
Calm	AU20	AU09	AU12
Happy	AU20	AU12	AU06
Sad	AU05	AU15	AU04
Angry	AU05	AU17	AU15
Fearful	AU45	AU17	AU26
Disgusted	AU20	AU01	AU26
Surprised	AU05	AU20	AU45

Table 4.8: The relationship between selected AUs and emotions

	Calm	Happy	Sad	Angry	Fearful	Disgusted	Surprised
AU01						#2	
AU04			#3				
AU05			#1	#1			#1
AU06		#3					
AU09	#2						
AU12	#3	#2					
AU15			#2	#3			
AU17				#2	#2		
AU20	#1	#1				#1	#2
AU26					#3	#3	
AU45					#1		#3

Chapter 5

Entire pipeline for Emotion Recognition: Results and Interpretation

In the previous chapters, we have illustrated a real-world challenge under study related to emotion recognition and expression pattern detection from facial videos. We addressed different research problems of functional statistical learning methods through three distinct stages. Starting from the background knowledge of the research questions and the developed methodologies, this chapter aims to integrate various stages to address two core questions: (1) how human facial expressions convey emotions and (2) how to accurately classify newly observed facial videos into specific emotional categories. The proposed model pipeline for training and testing the prediction accuracy of the model is evaluated by leave-one-out cross-validation through feasible training, testing, and validation sets. Moreover, the integrated approach here proposed guarantees the applicability of the model when fed by different datasets. In this way, a fully comprehensive framework for pattern detection and/or multi-class classification, built from the ground up, has paved the way for our research into emotion analysis based on facial expressions, as well as other multivariate multi-class function analysis-related inquiries, across different applications.

5.1 Three Stages of Curve Registration, Emotional Pattern Detection, and Multi-Classification

As stated in Chapter 1, the proposed methodology aims to explore a comprehensive and efficient solution for handling general time series data with multi-class comparison properties. Our methodology contains three stages: starting from curve registration and interpolation (Chapter 2), going through group-wise pattern detection (Chapter 3), to end to multi-classification (Chapter 4). The flowchart in Figure 5.1 illustrates these three stages of the model training process.

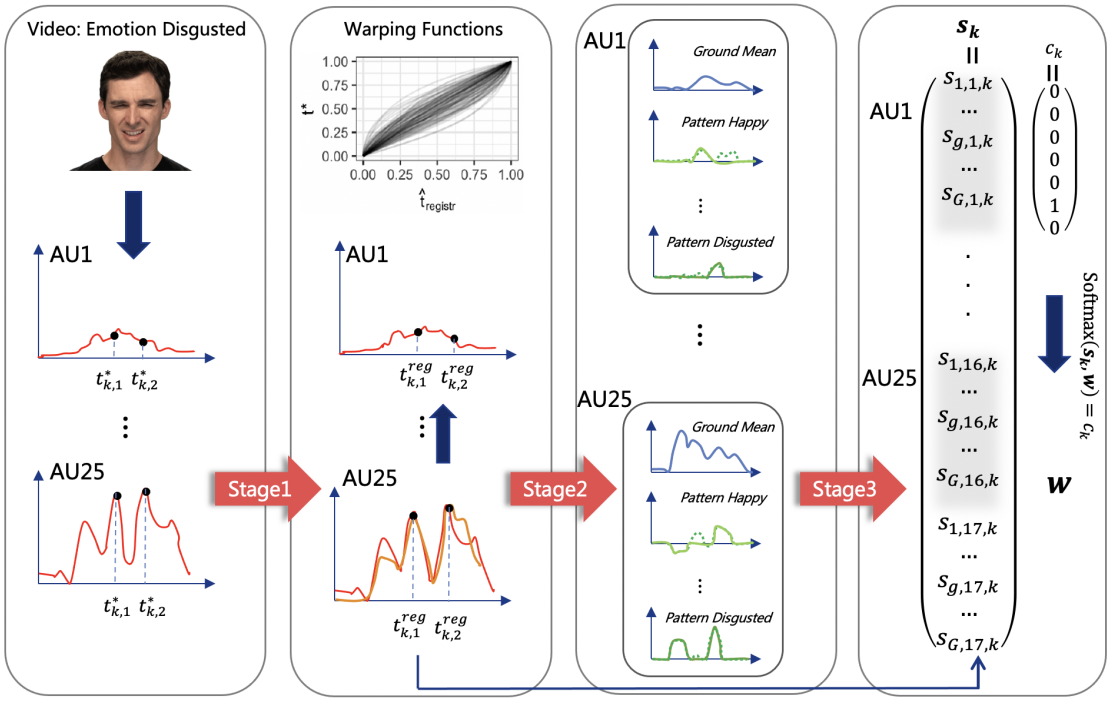


Figure 5.1: Flowchart Illustrating Three Stages of Model Training Process for Multivariate Multi-Class Functional Data Analysis.

Before delving into the three stages, it is important to consider the mathematical space in which the search target lives. The object under study is a set of multi-dimensional random functions that capture the evolution of engagement degrees extracted from facial videos over time, as illustrated in Section 1.2. The shape feature of these functions consists

of the sample mean curve and individual variability, which allows for a focus on the entire evolution process without the need to consider timely high-frequency correlations as discussed in Section 2.1.

Unlike ordinary multi-dimensional random variables with amplitude variability, which represents the magnitude of engagement values, these functions include an evolution process in the time domain, therefore involving phase variability. Even for videos performed by the same actor, there exists deviation in the time domain, i.e., the differences in the speeds of two evolution processes with similar trends, due to the independence of samples. Therefore, models and algorithms must have sufficient tolerance to handle unaligned evolution processes, which may require a large number of training samples and limited phase deviations between samples. Alternatively, as explained in Section 2.2, phase variability needs to be controlled and removed from the total variability. If traditional models designed for multi-dimensional variables are used to analyze a set of unaligned functions where time points at different samples are processing the same evolution process but at different stages, phase variability is ignored and mixed into the analysis of amplitude variability. As a result, the sample mean curve cannot accurately represent the original shape of the evolution of the samples.

To analyze the properties of a sample set, especially in detecting typical behaviors or patterns, it is necessary to align the samples by eliminating phase variability while preserving amplitude variability. However, this task is challenging, as there is no available template for alignment, and real-world data is often noisy and irregular. To address this problem, in Section 2.2.2, the Expectation-Maximization (EM) algorithm is applied and modified to alternatively optimize alignment and estimation processes and obtain gradual improvement. In each iteration of the EM algorithm, the sample mean function in each dimension is extracted as the template function in the expectation step. Then, in the maximization step, the samples begin to align with the current template function, resulting in a set of warping functions that minimize the alignment loss or maximize the likelihood. In the next round, a new template function is estimated from the aligned

samples in the previous iteration. This process continues until the error is less than a predetermined threshold, indicating that the alignment is precise enough under certain measures, and the aligned samples shown in Section 2.2.3 are sufficiently concentrated to support further analysis. In a summary, the EM algorithm's alternating optimization scheme enables gradual improvement of alignment and estimation, providing a powerful tool for detecting typical behaviors or patterns in a sample set.

After aligning the samples in the time domain by removing the phase variability using warping functions, we turn our attention to amplitude variability, which describes how much each action unit is activated during the aligned period. Note that the engagement degrees of the action units are heavily influenced not only by emotions but also by the pronunciation of sentences. To explore systematic emotional evolution patterns in natural communication scenes, we need to control the influence of pronunciation and movements triggered by it. This requires performers to speak the same thing under different emotions, which is rare in available datasets except for RAVDESS. RAVDESS, introduced in Section 1.3, contains performances of the same statements under neutral and seven other emotional modes, allowing us to separate the pronunciation-triggered engagement values from emotion-triggered ones. Assuming a fixed spoken speed, gender, and intensity level, we estimate the ground mean curve from all eight emotional sample sets as the pronunciation-triggered engagement values in Section 3.1, leaving the inner-group mean as the emotion-triggered patterns. We then test in Section 3.2 the significant parts of the emotional patterns under the hypothesis of the equality of pairwise inner group means between neutral and one emotional mode, to detect significant periods and durations of the entire time zones. By considering the properties of AU patterns in the selected time zones, we can summarize how facial expressions (AUs) are activated in terms of timing and engagement degrees, shown at the end of Section 3.3.

We have successfully tackled the first fundamental question concerning the identification of emotional patterns of action units. In contrast to the first question, where a single emotion and its corresponding samples were selected for pattern detection, the second

question aims to classify newly observed samples into one of several emotion classes using the detected emotional patterns as prior domain knowledge. Rather than relying on direct functional classification strategies, we aim to leverage the detected patterns to eliminate confounding factors, such as spoken speed, communication statements, and individual pronunciation deviation. To accomplish this, we first estimate in Section 4.1 the similarity between the newly observed sample and the set of patterns for each emotion class. Since the functions are processed in the functional space, we use the L_1 norm, i.e. the inner products of the sample curves and patterns, to construct the sample's "agreement scores" of each emotion. Next, the agreement scores are normalized and transformed using softmax functions, which have a probabilistic interpretation. After comparing with Consensus Algorithm via Voting Scheme (Section 4.2) , multinomial logistic regression with Cross Entropy Loss (Section 4.3.1) is chosen to classify the agreement scores generated from the new sample curves. This approach offers an effective way to classify emotions while controlling for confounding spoken-related factors.

5.2 Actor Performance Evaluation via Leave-One-Out Cross-Validation

After discussing the three stages of the model training process, we need to revisit the practical aspect of the case of emotion analysis - the limited availability of datasets supporting human facial emotional pattern detection due to concerns related to personal privacy and motion capture techniques development. The RAVDESS dataset, which suits the best for this study, also suffers from limited sample sizes. Additionally, the dataset's reliance on professional actors and actresses introduces the possibility of outliers due to the exaggerated understanding of emotions or personalized expressing habits. Mingling in the size-limited samples, potential outliers could impact largely on detecting the accurate and sociable emotional patterns.

Given the sample size issue, we adopt the leave-one-out (LOO) approach [87, 163] and

design a framework of training and testing double pipelines to evaluate the model's results and the quality of actors' performances. This involves iteratively holding out one actor's two performances as test samples while training the model with the remaining samples. Figure 5.2 provides a visual representation of training and testing pipelines outlined in the proposed framework.

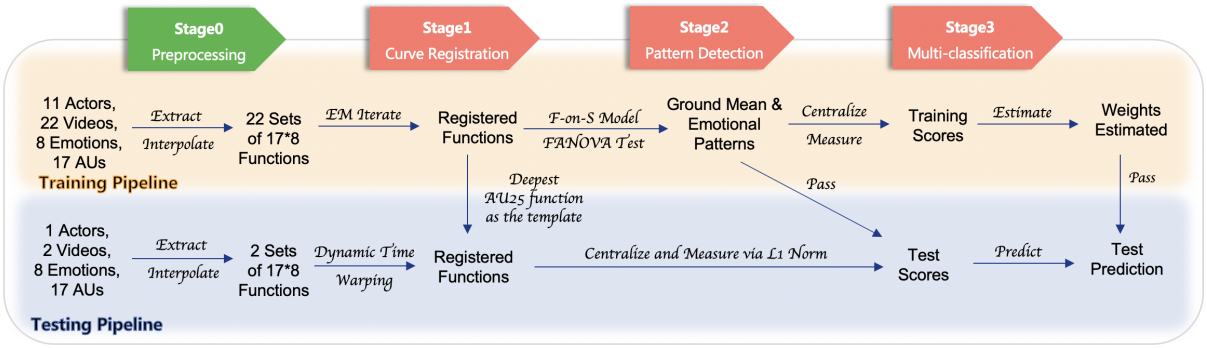


Figure 5.2: Training and testing pipelines in the proposed framework.

We have illustrated and summarized the training pipeline consisting of the three stages of the methodology in the previous section. We consider those operations composing the training pipeline. As one actor's performances are used as test samples, the remained samples are randomly divided into training samples with the performances of seven actors and validation samples with the performances of four actors.

For the test process, the goal shifts to validate the trained models. Test samples are first aligned to the deepest curve of AU25 in the registered training samples to catch the same registered timeline and remove pronunciation effects coherently. The depth for registered functional data is estimated by Modified Band Depth through the R package 'DepthTools' [149], while the alignment is achieved through dynamic time warping [113] (more detail in Appendix A.1). The registered test samples are then measured by the emotional patterns learned from the training process, resulting in test scores used for classification. The weights obtained from the trained Softmax model are applied to test scores to predict the class of test samples. The predicted class is then compared with the true test labels to evaluate the model's accuracy on unseen samples and assess the

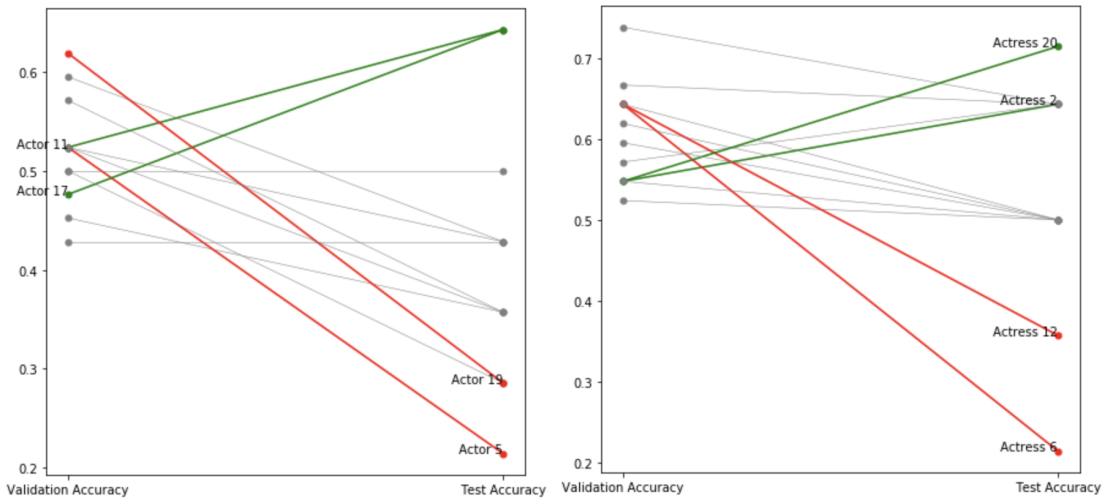
framework's performance.

The proposed framework with leave-one-out cross-validation provides both the validation accuracy in the training pipeline to evaluate the model's performance, and the test accuracy in the test pipelines to compare the similarity between the leaved-out samples and the remained ones used in the training pipeline. Thus, we can first verify whether the model's accuracy is heavily impacted by removing an actor's performance, and then confirm if the removed performances are outliers. This leads to further post hoc analysis of outlier detection and also completes the research on the limited-sample-size case study. Secondly, for the well-qualified samples without outliers, we can extract more precise emotional patterns, providing a better understanding of human emotions.

Table 5.1: Training accuracy, validation accuracy, and test accuracy achieved by the proposed framework's training and testing pipelines following the exclusion of each actor or actress's performances.

Male Actors				Female Actors			
ID	Training	Validation	Test	ID	Training	Validation	Test
1	0.973	0.452	0.357	2	0.991	0.548	0.643
3	0.964	0.524	0.429	4	0.973	0.571	0.643
5	0.955	0.524	0.214	6	0.991	0.643	0.214
7	0.946	0.500	0.500	8	0.964	0.524	0.500
9	0.938	0.595	0.429	10	0.973	0.738	0.643
11	0.973	0.524	0.643	12	0.982	0.643	0.357
13	0.973	0.500	0.286	14	0.982	0.667	0.643
15	0.973	0.429	0.429	16	0.982	0.619	0.500
17	0.911	0.476	0.643	18	0.964	0.643	0.500
19	0.946	0.619	0.286	20	0.973	0.548	0.714
21	0.884	0.524	0.357	22	0.955	0.548	0.500
23	0.946	0.571	0.357	24	0.973	0.595	0.500

Table 5.1 illustrates the training accuracy, validation accuracy, and test accuracy achieved by the proposed framework's training and testing pipelines following the exclusion of each actor or actress's performances. Further, we highlight the validation accuracy and test accuracy of four actors and four actresses in Figure 5.3 to better visualize the performer's influence. The table and the figure depict two situations worth discussing: high validation accuracy with low test accuracy (colored in red in Figure 5.3), and low validation accuracy with high test accuracy (colored in green).



(a) Males. **(b)** Females.

Figure 5.3: Validation accuracy VS Test accuracy for "good" and "bad" actors.

In the first case, high validation accuracy with low test accuracy colored in red, some actors (actor 5 and actor 19, actress 6 and actress 12) may exhibit performances that do not conform to the patterns of the other actors, leading to outliers. By removing the performances of these "bad actors", we eliminate outliers in the sample set of limited size and the model estimation is improved, as illustrated in Table 5.2. The second case, low validation accuracy with high test accuracy (colored in green), shows that despite low accuracy in the training process, the emotions expressed by the actor or actress are rather correctly classified in the test process. This indicates that the model's pattern detection ability is heavily influenced by these individuals (actor 11 and actor 17, actress 2 and actress 20), who are then leveraging the model's robustness. Therefore, we refer to actors with this trend as "good actors".

We summarize the "good actors" and "bad actors" for males and females in Table 5.2. The action unit evolution curves for these eight actors shown in Appendix A.5 can support further our analysis with a closer look at the individual's performances.

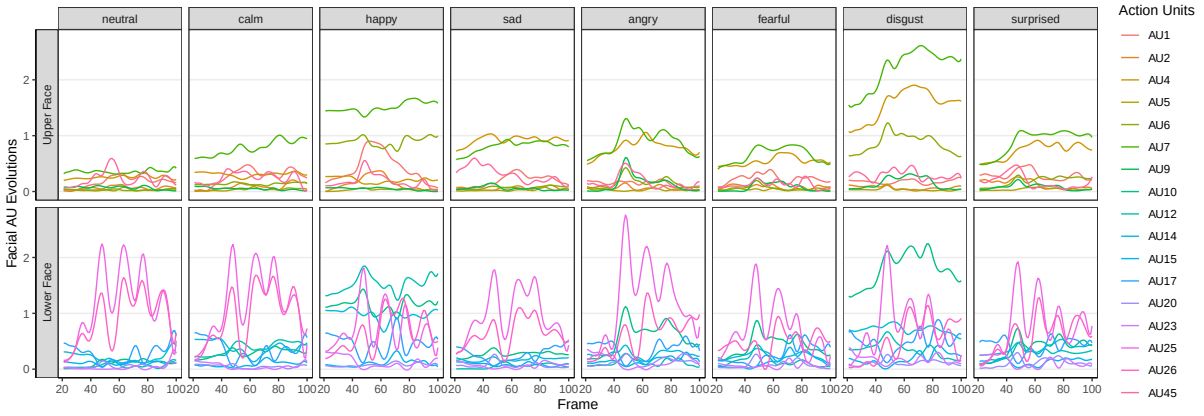
Table 5.2: Summary of "good actors" and "bad actors" for males and females, along with the model's accuracy following the exclusion of performances by respective actor pairs.

	Male		Accuracy	Female		Accuracy
"Good actors"	Actor 11	Actor 17	45.7%	Actor 2	Actor 20	51.4%
"Bad actors"	Actor 5	Actor 19	54.3%	Actor 6	Actor 12	68.6%

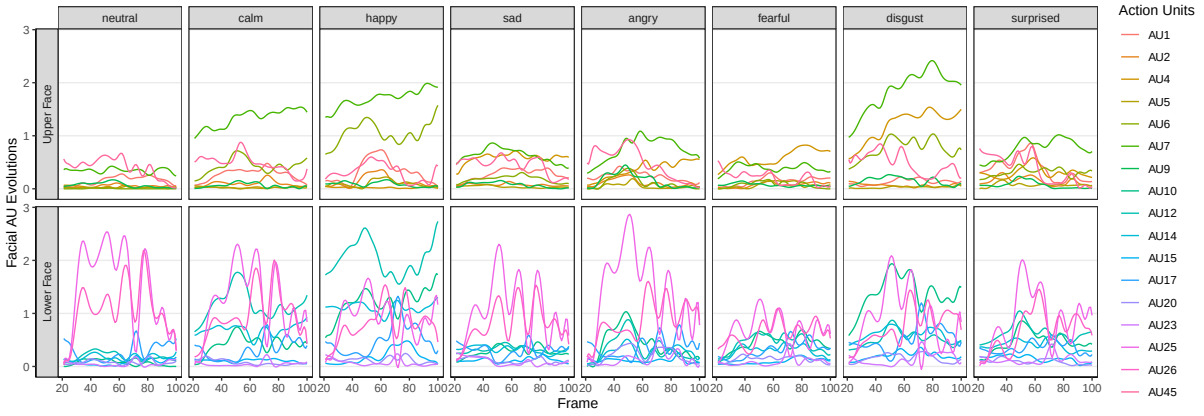
5.3 The Analysis of Gender Effects on Detected Emotional Patterns

In the previous chapters, only male actors' performances were considered because the aims were to evaluate the most proper methods to be used for each stage in the training pipeline, excluding most of the confounding variables. Now, with the unified framework of emotion pattern detection and classification proposed, we are able to analyze and compare the gender effect on emotional patterns, as males and females have been revealed to express emotions differently in the literature [34, 72, 138, 139]. Our investigation utilizes separately the performances of males and females in the group-wise functional linear model to scrutinize gender disparities in facial expressions. After the exclusion of outliers, in this section, we analyze how emotions are expressed and detected by their related Action Units (AUs) under gender effects. The group sample means of 17 AUs under eight emotions, after removing the performances of "bad" actors, are shown in Figure 5.4. Figure 5.13 provides a summary of ground means and emotional patterns for males and females, using a heatmap representation, while the graphs for each emotion are further displayed and zoomed in figures 5.5-5.12. Our findings on the detected patterns provide insight into the complex and nuanced nature of emotions and underscore the importance of considering gender differences in facial expressions. In the following, we motivate this statement by analyzing the results obtained on every single emotion.

Let us start from the heatmap of the ground mean curves for two genders, extracted independently from two datasets and shown in Figure 5.5. The homogeneity of the ground means attests to the model's effectiveness in capturing corresponding facial features while



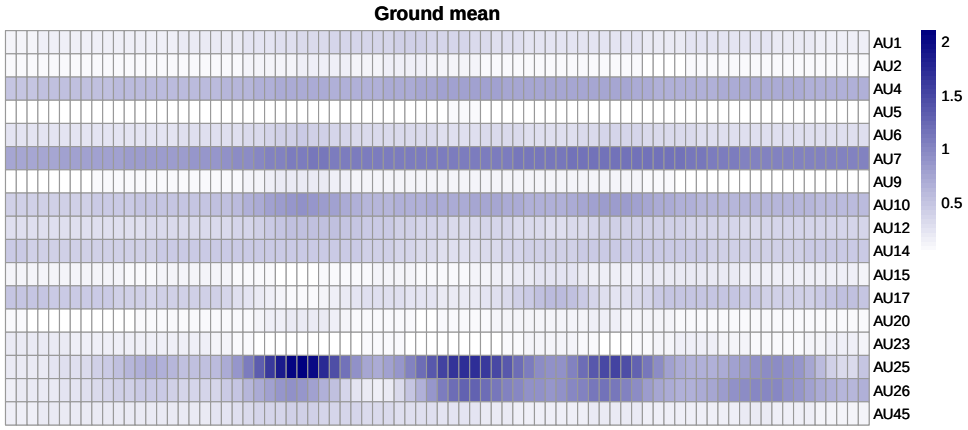
(a) Males except actor 5 and actor 19.



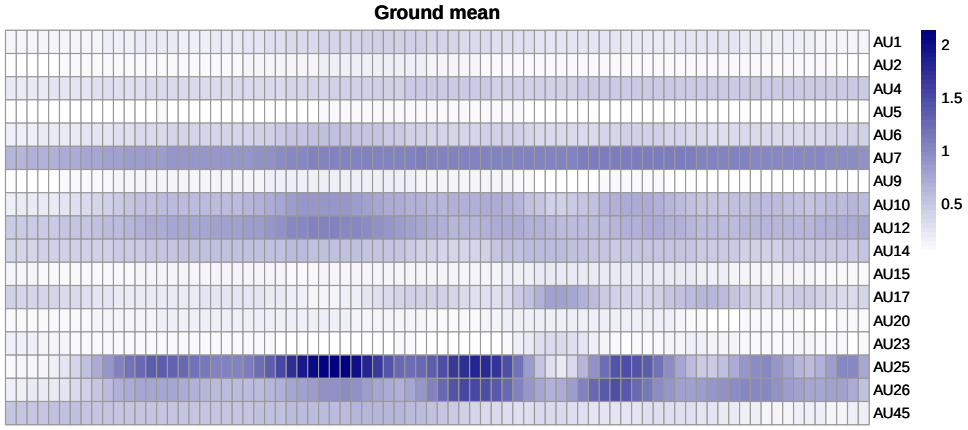
(b) Females except actress 6 and actress 12.

Figure 5.4: The group sample means of 17 AUs under eight emotions, after removing the performances of "bad" actors. The top row reports the patterns related to the upper part of the face, while the bottom row reports the patterns related to the lower part of the face.

performing the same utterance. Mouth-related action units, like Lips part (AU25), Jaw Drop (AU26), and Upper Lip Raiser (AU10), along with the eye-related action unit Lid tightener (AU7), provide insights into the fundamental muscular activities involved in normal speech production, irrespective of gender. Nevertheless, our analysis reveals notable gender-based differences in various action units, including the more substantial activations of AU4 among males and AU12, AU26, and AU45 among females. These disparities suggest that males often display more pronounced brow-lowering and stern facial expressions, while females tend to exhibit more distinct lip corner pulling, jaw-dropping, and



(a) Males.

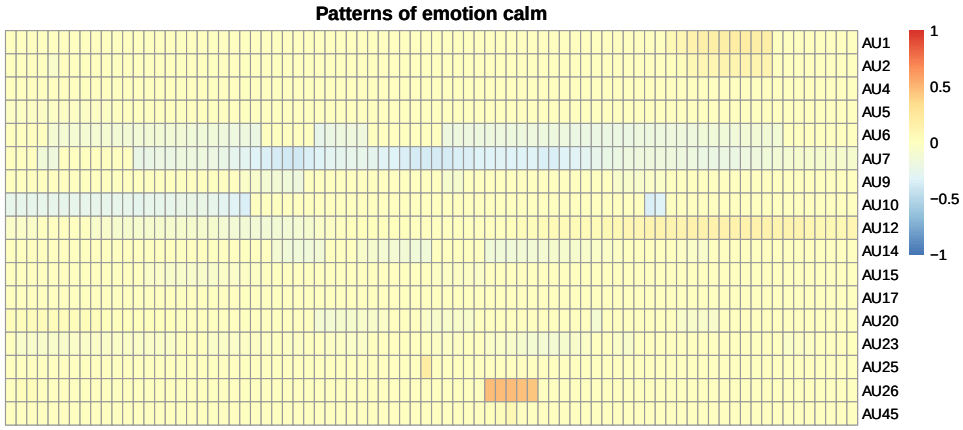


(b) Females.

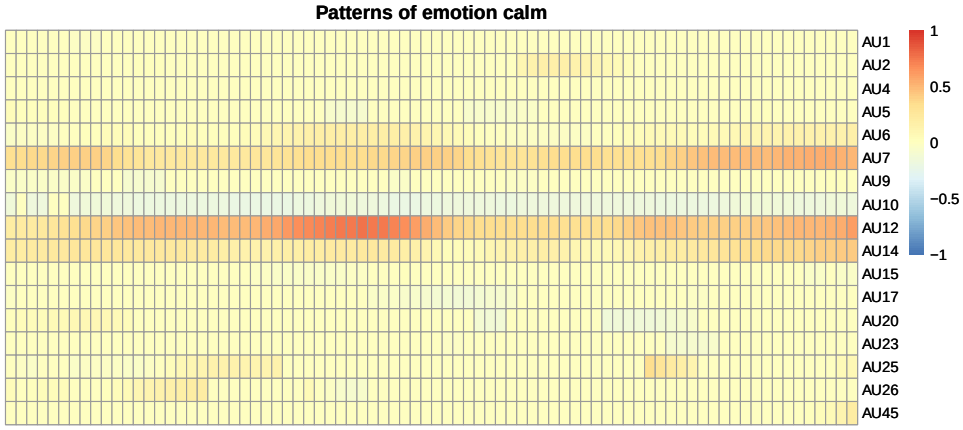
Figure 5.5: Heatmap of group-wise ground mean for males and females.

winking, pointing to a proclivity towards milder expressions.

To provide further clarity, our study utilizes facial expression analysis to detect deviations from the normal spoken mode. In examining patterns of emotion calm (shown in Figure 5.6), we find that most action units were activated to a lesser degree compared to the normal mode. However, we consistently observe the activation of Lid Tightener (AU7) in both male and female participants, with males exhibiting negative activation and females showing positive activation. This indicates that males tend to relax their lid tighteners, while females tend to increase their strength and tighten their lids more during calm speech. Notably, we also observe that females tend to exhibit more Lip Corner Pulling, as evidenced by the positive activation of AU12, while males seem to display the



(a) Males.

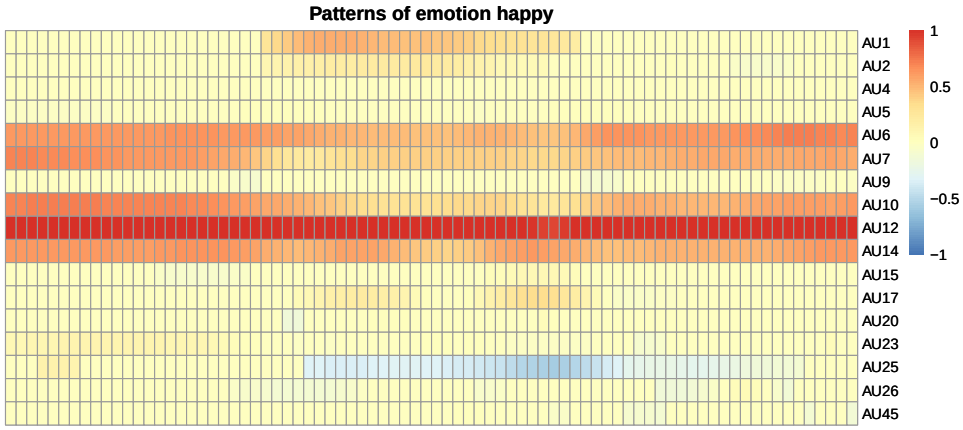


(b) Females.

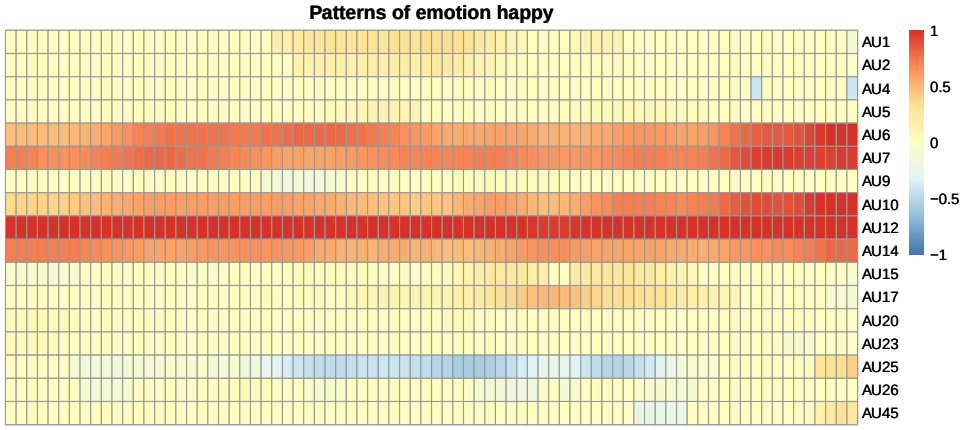
Figure 5.6: Heatmap of group-wise patterns detected for emotion calm.

same behavior for AU12 no matter for calm or normal speech.

The emotion of happiness is characterized by distinct facial expressions, as illustrated in Figure 5.7. Specifically, we find that the Lip Corner Puller (AU12) is highly activated in both male and female participants, followed by AU10 and AU14, which respectively denote the expressions of Dimpler and Upper Lip Raiser. This is consistent with the natural tendency to smile when feeling happy. Additionally, we observe that AU6 and AU7, which represent Cheek Raiser and Lid Tightener in the upper face, respectively, are activated, in line with existing literature suggesting that genuine smiles involve both the mouth and the eyes. Notably, we also notice the reduced activation of AU25 compared to the normal spoken mode, suggesting that when we communicate while experiencing



(a) Males.

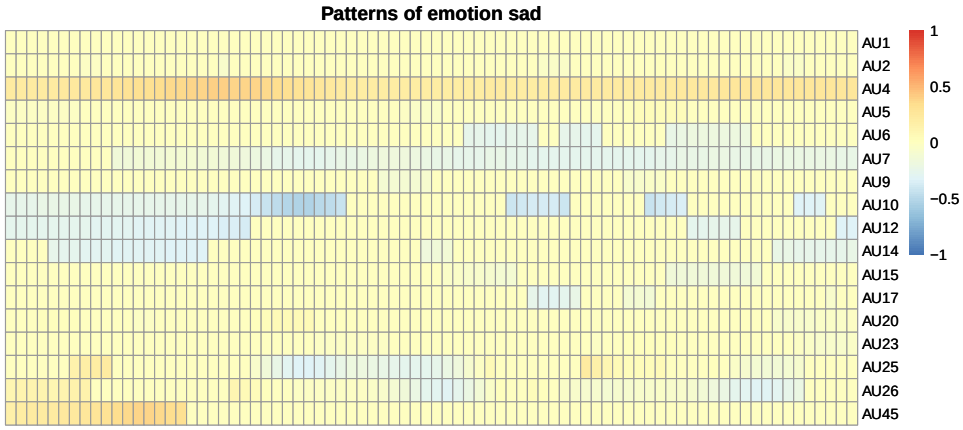


(b) Females.

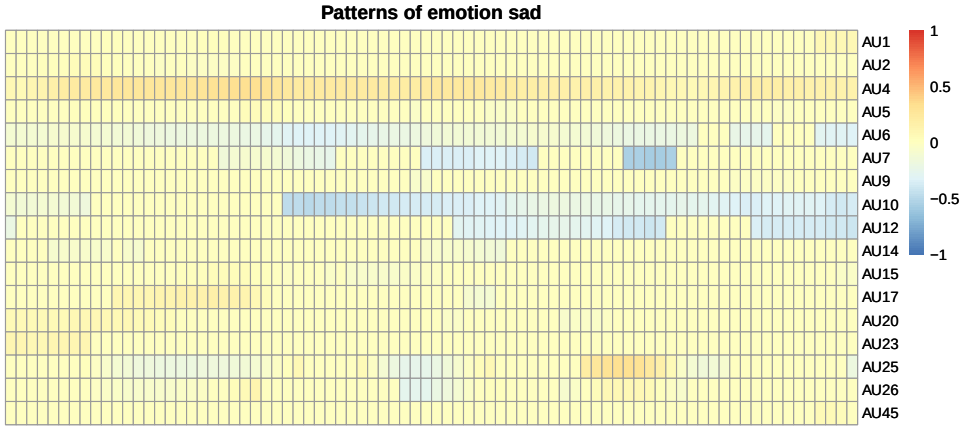
Figure 5.7: Heatmap of group-wise patterns detected for emotion happiness.

happiness, we rely on a wider range of mouth-related expressions. In exploring the gender effects, we find that males exhibit heightened activation of the Inner Brow Raiser (AU1), whereas females exhibit enhanced activations of the Lid Tightener (AU7) and Chin Raiser (AU17).

The first negative emotion under analysis is sadness, which exhibits multiple Action Units (AUs) with lower intensities in Figure 5.8, making it challenging to detect and summarize patterns. This is also reflected in the low correctness rate for this emotion in classification. For both males and females, AU4 shows long-term enhancement during sadness. However, there are differences in the inhibition of AUs between males and females. Specifically, for males, AU10, AU12, and AU14 are influenced at the beginning of



(a) Males.

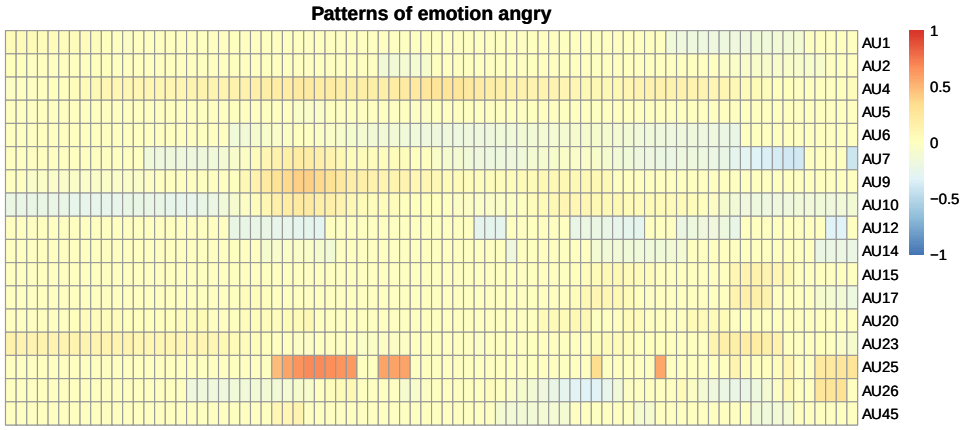


(b) Females.

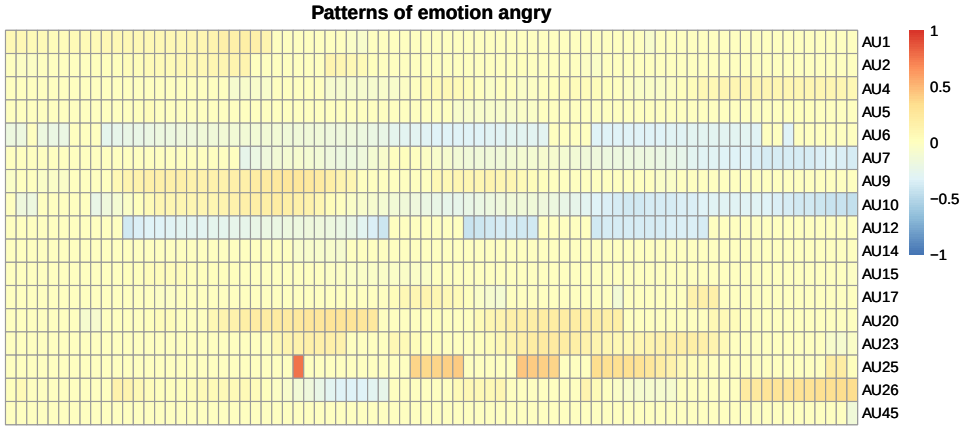
Figure 5.8: Heatmap of group-wise patterns detected for emotion sadness.

the performance, whereas for females, they are influenced in the ending part. Additionally, males show intermittent inhibition of AU6 and continuous inhibition of AU7, while the opposite is true for females. These results suggest that a more in-depth analysis with larger datasets is necessary to elucidate the patterns of AUs under the emotion of sadness.

Figure 5.9 shows the detected patterns of the emotional expression of anger. We observe significant increases in short periods in the activation of AU25, which suggests that the pronunciation of stressed words is intensified under the influence of anger. Furthermore, we find a simultaneous increase in the activation of AU9 (Nose Wrinkler) and AU10 (Upper Lip Raiser) during the peak activation of AU25 in the ground mean (pronunciation mode) in previous figures, indicating that individuals expressing anger tend to



(a) Males.

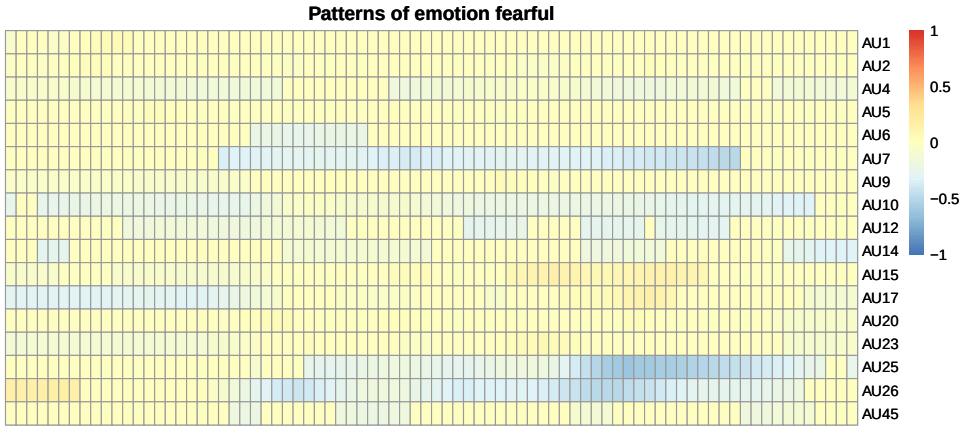


(b) Females.

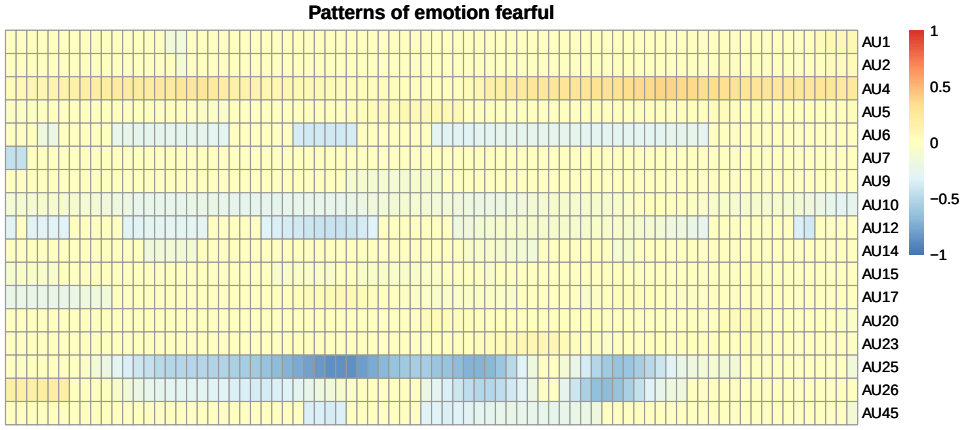
Figure 5.9: Heatmap of group-wise patterns detected for emotion angry.

accentuate their facial expression towards the central point of the nose, especially when speaking verbs. Intriguingly, AU10 (Upper Lip Raiser) is activated solely during this moment, while it is mostly inhibited in other moments compared to the normal pronunciation mode, together with AU6 (Cheek Raiser), AU7 (Lid Tightener), and AU12 (Lip Corner Puller). This phenomenon is found to be more prominent and persistent among females than males. Additionally, we observe that males tend to use more of their upper face muscles, whereas females tend to use more of their lower face muscles, as evident from the slight enhancement of AU4 (Brow Lowerer) in males, and AU20 (Lip stretcher) and AU23 (Lip Funneler) in females.

Figure 5.10 associated with fear shows significant negative inhibition of certain Action



(a) Males.

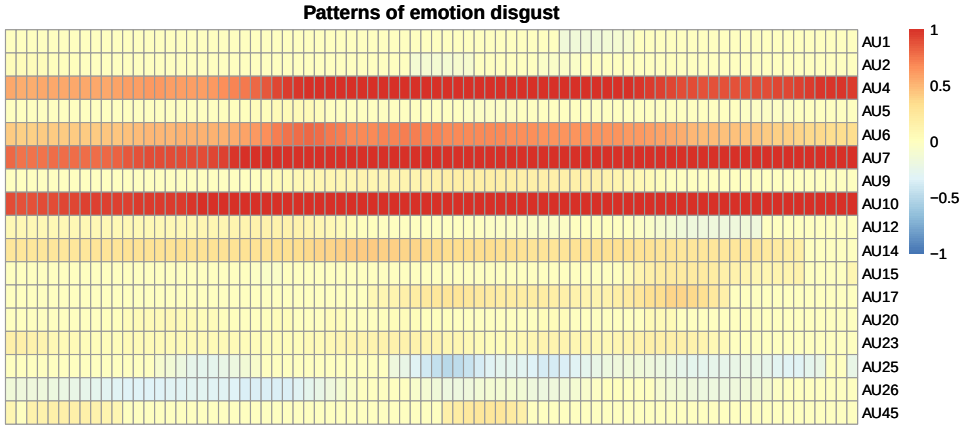


(b) Females.

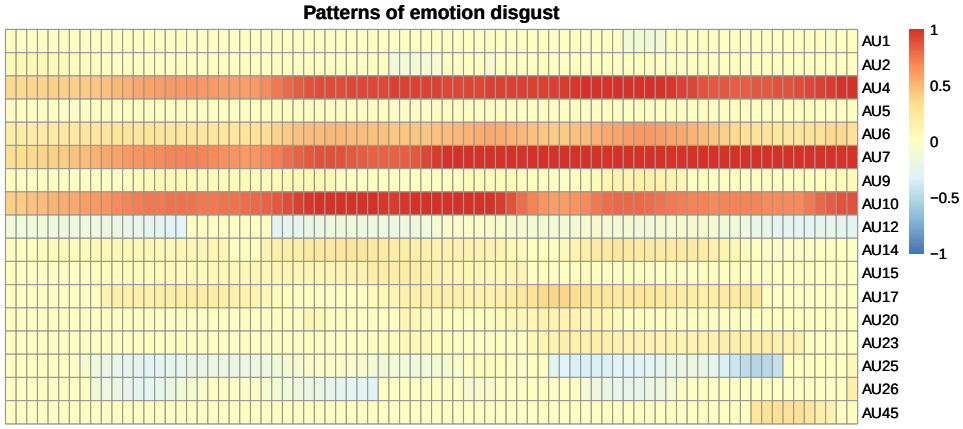
Figure 5.10: Heatmap of group-wise patterns detected for emotion fearful.

Units (AUs) when compared to the pronunciation mode. It suggests that fear prompts individuals to control their facial expressions and avoid revealing their emotional state. Specifically, we observe the strongest negative inhibition in AU25, followed by AU26 and AU10 in both males and females. However, we also find gender-dependent differences in the patterns of inhibited AUs. Males exhibit further negative inhibitions in AU7 (Lid Tightener) and AU17 (Chin Raiser), in contrast, females in AU6 (Cheek Raiser) and AU12 (Lip Corner Puller). Furthermore, it seems that the activation of AU5 is not inhibited in females when expressing fear. This may indicate that the activation of this AU is a more reliable indicator of fear in females. It is also possible that females use AU5 to enhance their expressions of fear, perhaps as a way to communicate their emotional state to others

more effectively.



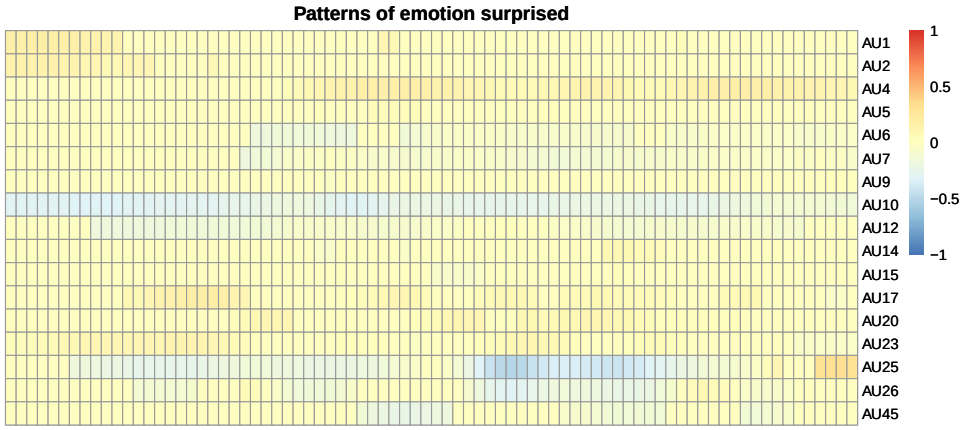
(a) Males.



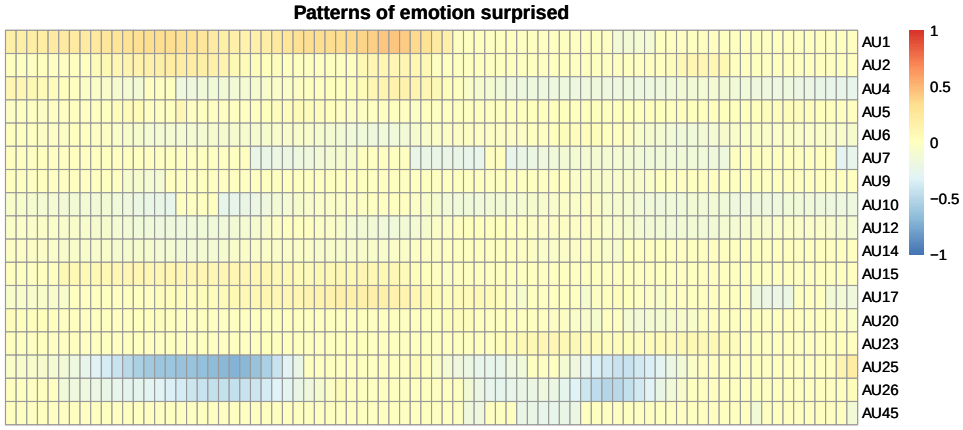
(b) Females.

Figure 5.11: Heatmap of group-wise patterns detected for emotion disgust.

In comparison to other negative emotions with complicated and variant patterns, the expression of disgust exhibits distinct patterns that are shared by both males and females, as shown in Figure 5.11. In particular, there is a strong and prolonged enhancement of AU4 (Brow Lowerer), AU6 (Cheek Raiser), AU7 (Lid Tightener), and AU10 (Upper Lip Raiser), along with some periodical inhibitions of AU25 (Lips part) and AU26 (Jaw Drop) in both genders. However, there are some gender-specific differences as well. For males, there is a slight enhancement of AU14 (Dimpler), whereas, in females, there is a more pronounced inhibition of AU12 (Lip Corner Puller) when experiencing disgust compared to the pronunciation mode.



(a) Males.



(b) Females.

Figure 5.12: Heatmap of group-wise patterns detected for emotion surprised.

The final emotion analyzed is emotion surprised, which is a complex emotion with considerable variation in its expression among individuals. In general, both males and females exhibit inhibition of AU25 and AU26, although females show greater inhibition during the first half of the pronunciation. Males experience continuous inhibition of AU10, while females show inhibition in a more fragmented manner. AU4 is slightly influenced during the surprise, with a positive effect on males and a negative effect on females. Notably, females display a significantly larger and more prolonged enhancement of AU1 than males when experiencing surprise. These findings suggest that the expression of surprise is influenced by a combination of factors, including gender and variations in the timing and intensity of specific Action Units performed by different individuals. It may

be worthwhile to use a larger sample size to detect more significant patterns.

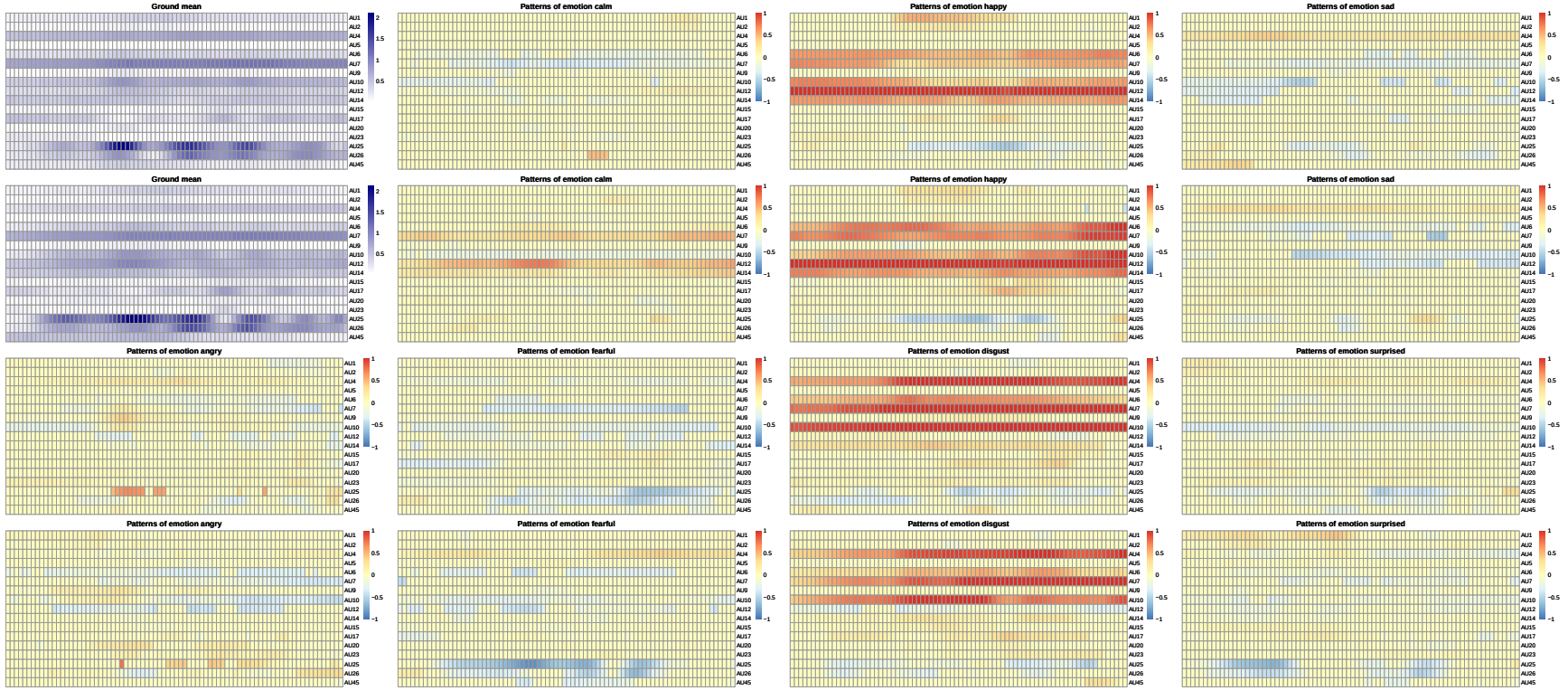


Figure 5.13: A summary of figures for ground mean and emotional patterns for males (the first and third lines) and females (the second and fourth lines) after removing the outliers.

5.4 Conclusion and Insights

In this thesis, we have addressed the real-world challenges associated with emotion recognition and expression pattern detection from facial videos. The problem at hand consisted of various consecutive stages, such as smooth function construction, registration across multivariates and multiple classes, significant and distinct pattern detection, and final multi-class classification. The primary challenges arose from realistic data acquisition constraints, including the limited quality of most video-type data, small sample sets with only 24 actors in the RAVDESS dataset, the validity and variability of actors' performances, and the heterogeneous understanding and expression of emotions influenced by potential confounding variables like gender, race, and culture. We explored and implemented several relevant statistical learning methods for each sub-task, comparing the model's efficiency, stability, and explainability to identify the optimal combination for solving the problem comprehensively. In our research, we put forward a dynamic integration of model training and prediction testing processes, which has been thoroughly evaluated using leave-one-out cross-validation implemented via practical training and testing pipelines. Our meticulously crafted integrated approaches are tailor-made to facilitate optimal performance of the models, even when subjected to variable parameter settings and disparate input data, rendering our framework both robust and versatile.

In a summary, we developed a methodology that employed multivariate function-on-scalar regression models and functional analysis of variance (FANOVA) to detect latent emotion-specific patterns in an interpretable manner and achieve automatic emotion recognition. Our functional regression model disentangled common information from group-specific influences and individual noise through paired group comparisons, and we analyzed the results using permutation-based FANOVA tests. This approach identified time zones with significantly different group pattern means and filtered the patterns accordingly. The filtered group patterns, which reflected notable mean characteristics in grouped units, were used as prior knowledge for multi-classification in a reduced feature

space. We tested the efficiency of this methodology on simulated data and applied it to the RAVDESS dataset, which features professional actors performing various emotions. We examine the transformation of observed functions into a shared low-dimensional space based on group-wise patterns across all classes. We then use the generated scores to classify new functional data, employing both non-parametric consensus voting and parametric multinomial logistic regression methods for multi-class classification, evaluating their predictive and explanatory capabilities.

Our primary contribution is the development of a comprehensive, explainable, and efficient pipeline that addresses the complex real-world challenge of automatic multi-class classification of multivariate functional data. Our secondary goal is to understand and transfer facial expression patterns to virtual humans, which requires explaining which, when, and how much different action units are activated under various emotions. This comprehensive framework offers a solid foundation for pattern detection, multi-classification, and potential applications in various fields, including emotion analysis based on expressions and multivariate multi-class function analysis-related inquiries. Our resulting framework has unveiled intriguing insights and comparisons through the detailed analysis of detected patterns of specific expressions under seven emotions. These discoveries further highlight the efficiency and explainability of our approach. While we may not have proposed entirely novel models, we have adapted suitable existing methods to meet our requirements and managed to connect well the various stages. Based on the current solid framework, we would be able to delve deeper into each subtask in the future.

In future work, we can explore several avenues for extension. While we applied standard classification methods in this thesis, ensemble learning classifiers based on agreement scores could yield improved classification results. As the proposed methodology and the related software can be considered under a multi-class format with a potential common mean and pair comparisons between control and treatment groups, furthermore, our methodology's application could be expanded to other functional data scenarios across different fields, such as studying stock performances in various industries alongside baseline in-

dexes in the financial market, or analyzing the sound of performances after pre-processing analysis instead of visual facial movements.

Acknowledgment

I would like to wholeheartedly acknowledge the financial support provided by the BIGMATH project. The BIGMATH project has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No 812912. I am incredibly grateful for the invaluable opportunity and experience that the BIGMATH project has offered me in both academic research and industrial internship. This unique experience has enriched my knowledge and skills, and played a significant role in my professional growth.

I would like to take this opportunity to express my profound gratitude to my advisors, Professor Alessandra Micheletti and co-supervisor Professor Natasa Krklec Jerinkic, for their invaluable guidance, unwavering support, and mentorship throughout my Ph.D. journey. Their invaluable guidance has been instrumental in shaping my research direction and providing me with the necessary tools and knowledge to tackle complex problems. They have always been patient in explaining intricate concepts, addressing my questions, and helping me develop a strong foundation in the field. I am grateful for their dedication to my growth as a researcher. I cannot thank them enough for their unwavering support throughout my journey, as they have always believed in my potential, even in moments when I doubted myself. Their encouragement and reassurance have been a great source of motivation for me and have inspired me to persevere through the challenges I have faced. Their mentorship has played a crucial role in my personal and professional development. They have not only shared their wealth of knowledge and experience but have also consistently provided me with guidance on how to navigate the academic world and

make the most of my Ph.D. journey. Their wisdom and advice have been invaluable, and I feel fortunate to have had the opportunity to learn from such distinguished mentors. I would also like to extend my gratitude to my industrial co-supervisor Zoranka Desnica from company 3Lateral for her insightful feedback on my thesis and for providing me with a unique perspective on the practical applications of my research.

I am deeply thankful for the opportunity to work alongside my exceptional colleagues in the BIGMATH program. I have not only gained invaluable knowledge from them but also learned important life lessons that will stay with me forever.

Lastly, I would like to express my heartfelt gratitude to my family for their constant support, encouragement, and love. Their faith in me has been my pillar of strength, and I dedicate my accomplishments to them.

Bibliography

[1] Stream iclone multi-device mocap with unreal live link. https://www.pinterest.com/pin/stream-iclone-multidevice-mocap-with-unreal-live-link--400961173076099459/. Accessed: 2022-08-01.

[2] IMOTIONS facial action coding system (facs) – a visual guidebook. https://imotions.com/blog/facial-action-coding-system/. Accessed: 2022-08-01.

[3] Selected facs action units. https://www.i3b.org/sites/default/files/u6/i3B%20Caf%C3%A9_VicarVision-2013-05-16.pdf. Accessed: 2022-08-01.

[4] B. J. Abbaschian, D. Sierra-Sosa, and A. Elmaghraby. Deep learning techniques for speech emotion recognition, from databases to models. *Sensors*, 21(4):1249, 2021.

[5] B. Amos, B. Ludwiczuk, M. Satyanarayanan, et al. Openface: A general-purpose face recognition library with mobile applications. *CMU School of Computer Science*, 6(2):20, 2016.

[6] T. W. Anderson. An introduction to multivariate statistical analysis. Technical report, 1958.

[7] G. Aneiros, R. Cao, R. Fraiman, C. Genest, and P. Vieu. Recent advances in functional data analysis and high-dimensional statistics. *Journal of Multivariate Analysis*, 170:3–9, 2019.

[8] D. Ayata, Y. Yaslan, and M. Kamaşak. Emotion recognition via random forest and galvanic skin response: Comparison of time based feature sets, window sizes and

wavelet approaches. In *2016 Medical Technologies National Congress (TIPTEKNO)*, pages 1–4. IEEE, 2016.

[9] E. Bagheri, A. Bagheri, P. G. Esteban, and B. Vanderborgth. A novel model for emotion detection from facial muscles activity. In *Iberian Robotics conference*, pages 237–249. Springer, 2019.

[10] T. Baltrušaitis, P. Robinson, and L.-P. Morency. Openface: an open source facial behavior analysis toolkit. In *2016 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1–10. IEEE, 2016.

[11] R. F. Barber, M. Reimherr, and T. Schill. The function-on-scalar lasso with applications to longitudinal gwas. *Electronic Journal of Statistics*, 11(1):1351–1389, 2017.

[12] K. Benhenni, F. Ferraty, M. Rachdi, and P. Vieu. Local smoothing regression with functional data. *Computational Statistics*, 22(3):353–369, 2007.

[13] J. R. Berrendero, A. Justel, and M. Svarc. Principal components for multivariate functional data. *Computational Statistics & Data Analysis*, 55(9):2619–2634, 2011.

[14] S. Blain, A. Mihailidis, and T. Chau. Assessing the potential of electrodermal activity as an alternative access pathway. *Medical engineering & physics*, 30(4): 498–505, 2008.

[15] P. Blomsma, G. Skantze, and M. Swerts. Backchannel behavior influences the perceived personality of human and artificial communication partners. *Frontiers in Artificial Intelligence*, 5, 2022.

[16] G. Boente and R. Fraiman. Kernel-based functional principal components. *Statistics & probability letters*, 48(4):335–345, 2000.

[17] G. Boente, M. S. Barrera, and D. E. Tyler. A characterization of elliptical distributions and some optimality properties of principal components for functional data. *Journal of Multivariate Analysis*, 131:254–264, 2014.

[18] D. Böhning. Multinomial logistic regression algorithm. *Annals of the institute of Statistical Mathematics*, 44(1):197–200, 1992.

[19] C. Busso, M. Bulut, C.-C. Lee, A. Kazemzadeh, E. Mower, S. Kim, J. N. Chang, S. Lee, and S. S. Narayanan. Iemocap: Interactive emotional dyadic motion capture database. *Language resources and evaluation*, 42(4):335–359, 2008.

[20] T. T. Cai and P. Hall. Prediction in functional linear regression. *The Annals of Statistics*, 34(5):2159–2179, 2006.

[21] F. Canento, A. Fred, H. Silva, H. Gamboa, and A. Lourenço. Multimodal biosignal sensor data handling for emotion recognition. In *SENSORS, 2011 IEEE*, pages 647–650. IEEE, 2011.

[22] G. Casella and R. Berger. *Statistical Inference*. Duxbury Resource Center, June 2001. ISBN 0534243126.

[23] G. Cawley, N. Talbot, and M. Girolami. Sparse multinomial logistic regression via bayesian l1 regularisation. *Advances in neural information processing systems*, 19, 2006.

[24] L. Chen, Z. Wu, J. Ling, R. Li, X. Tan, and S. Zhao. Transformer-s2a: Robust and efficient speech-to-animation. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7247–7251. IEEE, 2022.

[25] J.-M. Chiou and H.-G. Müller. Diagnostics for functional regression via residual processes. *Computational Statistics & Data Analysis*, 51(10):4849–4863, 2007.

[26] J.-M. Chiou, Y.-T. Chen, and Y.-F. Yang. Multivariate functional principal component analysis: A normalization approach. *Statistica Sinica*, pages 1571–1596, 2014.

[27] J.-M. Chiou, Y.-F. Yang, and Y.-T. Chen. Multivariate functional linear regression and prediction. *Journal of Multivariate Analysis*, 146:301–312, 2016.

[28] E. Chung and J. P. Romano. Exact and asymptotically robust permutation tests. *The Annals of Statistics*, 41(2):484–507, 2013.

[29] G. Claeskens, B. W. Silverman, and L. Slaets. A multiresolution approach to time warping achieved by a bayesian prior–posterior transfer fitting strategy. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 72(5):673–694, 2010.

[30] R. Cowie, E. Douglas-Cowie, N. Tsapatsoulis, G. Votsis, S. Kollias, W. Fellenz, and J. G. Taylor. Emotion recognition in human-computer interaction. *IEEE Signal processing magazine*, 18(1):32–80, 2001.

[31] A. Cuevas. A partial overview of the theory of statistics with functional data. *Journal of Statistical Planning and Inference*, 147:1–23, 2014.

[32] A. Cuevas, M. Febrero, and R. Fraiman. An anova test for functional data. *Computational statistics & data analysis*, 47(1):111–122, 2004.

[33] W. Dai and M. G. Genton. Directional outlyingness for multivariate functional data. *Computational Statistics & Data Analysis*, 131:50–65, 2019.

[34] B. C. Davis, B. J. Warnick, A. H. Anglin, and T. H. Allison. Gender and counterstereotypical facial expressions of emotion in crowdfunded microlending. *Entrepreneurship Theory and Practice*, 45(6):1339–1365, 2021.

[35] J. Di, A. Spira, J. Bai, J. Urbanek, A. Leroux, M. Wu, S. Resnick, E. Simonsick, L. Ferrucci, J. Schrack, et al. Joint and individual representation of domains of

physical activity, sleep, and circadian rhythmicity. *Statistics in biosciences*, 11(2): 371–402, 2019.

[36] J. Ehret, A. Bönsch, L. Aspöck, C. T. Röhr, S. Baumann, M. Grice, J. Fels, and T. W. Kuhlen. Do prosody and embodiment influence the perceived naturalness of conversational agents' speech? *ACM Transactions on Applied Perception (TAP)*, 18(4):1–15, 2021.

[37] P. Ekman. Basic emotions. *Handbook of cognition and emotion*, 98(45-60):16, 1999.

[38] P. Ekman, W. V. Friesen, and J. C. Hager. *Facial Action Coding System. The manual.* Research Nexus division of Network Information Research Corporation, 2002.

[39] R. Ekman. *What the face reveals: Basic and applied studies of spontaneous expression using the Facial Action Coding System (FACS)*. Oxford University Press, USA, 1997.

[40] H. A. Elfenbein and N. Ambady. On the universality and cultural specificity of emotion recognition: a meta-analysis. *Psychological bulletin*, 128(2):203, 2002.

[41] G. Enderlein. Scheffé, h.: The analysis of variance. wiley, new york 1959, 477 seiten, 1961.

[42] Z. Fan and M. Reimherr. High-dimensional adaptive function-on-scalar regression. *Econometrics and statistics*, 1:167–183, 2017.

[43] M. Febrero-Bande and W. González-Manteiga. Generalized additive models for functional data. *Test*, 22(2):278–292, 2013.

[44] H. Feng, H. M. Golshan, and M. H. Mahoor. A wavelet-based approach to emotion classification using eda signals. *Expert Systems with Applications*, 112:77–86, 2018.

[45] F. Ferraty and P. Vieu. *Nonparametric functional data analysis: theory and practice*, volume 76. Springer, 2006.

[46] R. A. Fisher. On the'probable error'of a coefficient of correlation deduced from a small sample. *Metron*, 1:1–32, 1921.

[47] S. Fremdt, L. Horváth, P. Kokoszka, and J. G. Steinebach. Functional data analysis with increasing number of projections. *Journal of Multivariate Analysis*, 124:313–332, 2014.

[48] A. Fridlund. Human facial expression: An evolutionary view. *Academic Press*, 2014.

[49] J. Friedman, T. Hastie, and R. Tibshirani. Regularization paths for generalized linear models via coordinate descent. *Journal of statistical software*, 33(1):1, 2010.

[50] E. Fu and N. Heckman. Model-based curve registration via stochastic approximation em algorithm. *Computational Statistics & Data Analysis*, 131:159–175, 2019.

[51] K. Fukunaga and W. L. Koontz. Application of the karhunen-loeve expansion to feature selection and ordering. *IEEE Transactions on computers*, 100(4):311–318, 1970.

[52] A. Fydanaki and Z. Geradts. Evaluating openface: an open-source automatic facial comparison algorithm for forensics. *Forensic sciences research*, 3(3):202–209, 2018.

[53] E. Games. Recording facial animation from an ios device, 2021. URL `https://docs.unrealengine.com/4.27/en-US/AnimatingObjects/SkeletalMeshAnimation/FacialRecordingiPhone`.

[54] T. Gasser and A. Kneip. Searching for structure in curve samples. *Journal of the american statistical association*, 90(432):1179–1188, 1995.

[55] D. Gervini. Warped functional regression. *Biometrika*, 102(1):1–14, 2015.

[56] D. Gervini and T. Gasser. Nonparametric maximum likelihood estimation of the structural mean of a sample of curves. *Biometrika*, 92(4):801–820, 2005.

[57] A. Goia, C. May, and G. Fusai. Functional clustering and linear regression for peak load forecasting. *International Journal of Forecasting*, 26(4):700–711, 2010.

[58] J. Goldsmith, V. Zipunnikov, and J. Schrack. Generalized multilevel function-on-scalar regression and principal component analysis. *Biometrics*, 71(2):344–353, 2015.

[59] T. Górecki and Ł. Smaga. A comparison of tests for the one-way anova problem for functional data. *Computational Statistics*, 30(4):987–1010, 2015.

[60] T. Górecki and Ł. Smaga. Multivariate analysis of variance for functional data. *Journal of Applied Statistics*, 44(12):2172–2189, 2017.

[61] T. Górecki, M. Krzyśko, Ł. Waszak, and W. Wołyński. Selected statistical methods of data analysis for multivariate functional data. *Statistical Papers*, 59(1):153–182, 2018.

[62] T. Górecki, M. Krzyśko, and W. Wołyński. Independence test and canonical correlation analysis based on the alignment between kernel matrices for multivariate functional data. *Artificial Intelligence Review*, 53(1):475–499, 2020.

[63] B. Gregorutti, B. Michel, and P. Saint-Pierre. Grouped variable importance with random forests and application to multiple functional data analysis. *Computational Statistics & Data Analysis*, 90:15–35, 2015.

[64] M. Grimm, K. Kroschel, and S. Narayanan. The vera am mittag german audio-visual emotional speech database. In *2008 IEEE international conference on multimedia and expo*, pages 865–868. IEEE, 2008.

[65] H. Gunes and M. Piccardi. Bi-modal emotion recognition from expressive face and body gestures. *Journal of Network and Computer Applications*, 30(4):1334–1345, 2007.

[66] J. Guo, Z. Lei, J. Wan, E. Avots, N. Hajarolasvadi, B. Knyazev, A. Kuharenko, J. C. S. J. Junior, X. Baró, H. Demirel, et al. Dominant and complementary emotion recognition from still images of faces. *IEEE Access*, 6:26391–26403, 2018.

[67] P. Z. Hadjipantelis, J. A. Aston, H.-G. Müller, and J. P. Evans. Unifying amplitude and phase analysis: A compositional data approach to functional multivariate mixed-effects modeling of mandarin chinese. *Journal of the American Statistical Association*, 110(510):545–559, 2015.

[68] C. Happ and S. Greven. Multivariate functional principal component analysis for data observed on different (dimensional) domains. *Journal of the American Statistical Association*, 113(522):649–659, 2018.

[69] T. Hastie and J. Qian. Glmnet vignette. *Retrieved June*, 9(2016):1–30, 2014.

[70] T. Hastie, R. Tibshirani, J. H. Friedman, and J. H. Friedman. *The elements of statistical learning: data mining, inference, and prediction*, volume 2. Springer, 2009.

[71] T. Hastie, J. Qian, and K. Tay. An introduction to glmnet. *CRAN R Repositary*, 2021.

[72] U. Hess, R. B. Adams Jr, and R. E. Kleck. Facial appearance, gender, and emotion expression. *Emotion*, 4(4):378, 2004.

[73] G. R. Hjaltason and H. Samet. Properties of embedding methods for similarity searching in metric spaces. *IEEE Transactions on Pattern Analysis and machine intelligence*, 25(5):530–549, 2003.

[74] L. Horváth and P. Kokoszka. *Inference for functional data with applications*, volume 200. Springer Science & Business Media, 2012.

[75] L. Horváth, P. Kokoszka, and M. Reimherr. Two sample inference in functional linear models. *Canadian Journal of Statistics*, 37(4):571–591, 2009.

[76] L. Horváth, M. Hušková, and G. Rice. Test of independence for functional data. *Journal of Multivariate Analysis*, 117:100–119, 2013.

[77] W. Huang, K. A. Gallivan, A. Srivastava, P.-A. Absil, et al. Riemannian optimization for elastic shape analysis. In *Mathematical theory of Networks and Systems*, 2014.

[78] M. Hubert, P. J. Rousseeuw, and P. Segaert. Multivariate functional outlier detection. *Statistical Methods & Applications*, 24(2):177–202, 2015.

[79] F. Ieva, A. M. Paganoni, D. Pigoli, and V. Vitelli. Multivariate functional clustering for the morphological analysis of electrocardiograph curves. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 62(3):401–418, 2013.

[80] S. A. Jackson, S. Jackson, and D. E. Brashers. *Random factors in ANOVA*, volume 98. Sage, 1994.

[81] J. Jacques and C. Preda. Clustering multivariate functional data. In *COMPSTAT 2012*, pages 353–366, 2012.

[82] J. Jacques and C. Preda. Model-based clustering for multivariate functional data. *Computational Statistics & Data Analysis*, 71:92–106, 2014.

[83] G. M. James and T. J. Hastie. Functional linear discriminant analysis for irregularly sampled curves. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 63(3):533–550, 2001.

[84] R. Jenke, A. Peer, and M. Buss. Feature extraction and selection for emotion recognition from eeg. *IEEE Transactions on Affective computing*, 5(3):327–339, 2014.

[85] S. Jerritta, M. Murugappan, R. Nagarajan, and K. Wan. Physiological signals based human emotion recognition: a review. In *2011 IEEE 7th international colloquium on signal processing and its applications*, pages 410–415. IEEE, 2011.

[86] T. Kanade, J. F. Cohn, and Y. Tian. Comprehensive database for facial expression analysis. In *Proceedings fourth IEEE international conference on automatic face and gesture recognition (cat. No. PR00580)*, pages 46–53. IEEE, 2000.

[87] M. Kearns and D. Ron. Algorithmic stability and sanity-check bounds for leave-one-out cross-validation. In *Proceedings of the tenth annual conference on Computational learning theory*, pages 152–162, 1997.

[88] R. A. Khalil, E. Jones, M. I. Babar, T. Jan, M. H. Zafar, and T. Alhussain. Speech emotion recognition using deep learning techniques: A review. *IEEE Access*, 7: 117327–117345, 2019.

[89] R. Khan and R. Debnath. Human distraction detection from video stream using artificial emotional intelligence. *International Journal of Image, Graphics and Signal Processing*, 10(2):19, 2020.

[90] I. Kim, S. Balakrishnan, and L. Wasserman. Minimax optimality of permutation tests. *The Annals of Statistics*, 50(1):225–251, 2022.

[91] Y. Kim, S. Kwon, and S. H. Song. Multiclass sparse logistic regression for classification of multiple cancer types using gene expression data. *Computational Statistics & Data Analysis*, 51(3):1643–1655, 2006.

[92] A. Kneip and T. Gasser. Statistical tools to analyze data representing a sample of curves. *The Annals of Statistics*, pages 1266–1305, 1992.

[93] A. Kneip and J. O. Ramsay. Combining registration and fitting for functional models. *Journal of the American Statistical Association*, 103(483):1155–1165, 2008.

[94] B. C. Ko. A brief review of facial emotion recognition based on visual information. *sensors*, 18(2):401, 2018.

[95] D. R. Kowal and D. C. Bourgeois. Bayesian function-on-scalars regression for high-dimensional data. *Journal of Computational and Graphical Statistics*, 29(3):629–638, 2020.

[96] B. Krishnapuram, L. Carin, M. A. Figueiredo, and A. J. Hartemink. Sparse multinomial logistic regression: Fast algorithms and generalization bounds. *IEEE transactions on pattern analysis and machine intelligence*, 27(6):957–968, 2005.

[97] M. Krzyśko and Ł. Waszak. Canonical correlation analysis for functional data. *Biometrical Letters*, 50(2):95–105, 2013.

[98] H.-J. Lee and K.-S. Hong. A study on emotion recognition method and its application using face image. In *2017 International Conference on Information and Communication Technology Convergence (ICTC)*, pages 370–372. IEEE, 2017.

[99] E. L. Lehmann, J. P. Romano, and G. Casella. *Testing statistical hypotheses*, volume 3. Springer, 2005.

[100] K. Li and S. Luo. Bayesian functional joint models for multivariate longitudinal and time-to-event data. *Computational statistics & data analysis*, 129:14–29, 2019.

[101] J. Z. Lim, J. Mountstephens, and J. Teo. Emotion recognition using eye-tracking: taxonomy, review and current challenges. *Sensors*, 20(8):2384, 2020.

[102] Y.-L. Lin and G. Wei. Speech emotion recognition based on hmm and svm. In *2005 international conference on machine learning and cybernetics*, volume 8, pages 4898–4901. IEEE, 2005.

[103] S. Livingstone and F. Russo. The ryerson audio-visual database of emotional speech and song (ravdess): A dynamic, multimodal set of facial and vocal expressions in north american english. *PloS one*, 13(5):e0196391, 2018. https://smartlaboratory.org/ravdess/.

[104] S. López-Pintado and J. Romo. On the concept of depth for functional data. *Journal of the American statistical Association*, 104(486):718–734, 2009.

[105] S. López-Pintado, J. Romo, and A. Torrente. Robust depth-based tools for the analysis of gene expression data. *Biostatistics*, 11(2):254–264, 2010.

[106] C. Luna-Jiménez, R. Kleinlein, D. Griol, Z. Callejas, J. M. Montero, and F. Fernández-Martínez. A proposal for multimodal emotion recognition using aural transformers and action units on ravdess dataset. *Applied Sciences*, 12(1):327, 2021.

[107] J. S. Marron, J. O. Ramsay, L. M. Sangalli, and A. Srivastava. Functional data analysis of amplitude and phase variation. *Statistical Science*, pages 468–484, 2015.

[108] J. Martínez Torres, J. Pastor Pérez, J. Sancho Val, A. McNabola, M. Martínez Comesaña, and J. Gallagher. A functional data analysis approach for the detection of air pollution episodes and outliers: A case study in dublin, ireland. *Mathematics*, 8(2):225, 2020.

[109] J. Matuk, K. Bharath, O. Chkrebtii, and S. Kurtek. Bayesian framework for simultaneous registration and estimation of noisy, sparse, and fragmented functional data. *Journal of the American Statistical Association*, pages 1–17, 2021.

[110] A. Milton, S. S. Roy, and S. T. Selvi. Svm scheme for speech emotion recognition using mfcc feature. *International Journal of Computer Applications*, 69(9), 2013.

[111] L. Molina. *Celebrity Avatars: A Technical Approach to Creating Digital Avatars for Social Marketing Strategies*. PhD thesis, Florida Atlantic University, 2021.

[112] D. Montgomery. The regression approach to the analysis of variance. *Design and Analysis of Experiments*, pages 121–126, 1991.

[113] M. Müller. Dynamic time warping. *Information retrieval for music and motion*, pages 69–84, 2007.

[114] R. Muthukrishnan and R. Rohini. Lasso: A feature selection technique in predictive modeling for machine learning. In *2016 IEEE international conference on advances in computer applications (ICACA)*, pages 18–20. IEEE, 2016.

[115] H.-W. Ng, V. D. Nguyen, V. Vonikakis, and S. Winkler. Deep learning for emotion recognition on small datasets using transfer learning. In *Proceedings of the 2015 ACM on international conference on multimodal interaction*, pages 443–449, 2015.

[116] L. M. Oberman, P. Winkielman, and V. S. Ramachandran. Slow echo: facial emg evidence for the delay of spontaneous, but not voluntary, emotional mimicry in children with autism spectrum disorders. *Developmental science*, 12(4):510–520, 2009.

[117] G. Pala and C. E. Erdem. Performance comparison of deep learning based face identification methods for video under adverse conditions. In *2019 15th International Conference on Signal-Image Technology & Internet-Based Systems (SITIS)*, pages 90–97. IEEE, 2019.

[118] H. K. Palo, M. Chandra, and M. N. Mohanty. Emotion recognition using mlp and gmm for oriya language. *International Journal of Computational Vision and Robotics*, 7(4):426–442, 2017.

[119] M. T. Pilehvar and J. Camacho-Collados. Embeddings in natural language processing: Theory and advances in vector representations of meaning. *Synthesis Lectures on Human Language Technologies*, 13(4):1–175, 2020.

[120] A. Pini and S. Vantini. The interval testing procedure: a general framework for inference in functional data analysis. *Biometrics*, 72(3):835–845, 2016.

[121] A. Pini and S. Vantini. Interval-wise testing for functional data. *Journal of Nonparametric Statistics*, 29(2):407–424, 2017.

[122] A. Pini, S. Vantini, B. M. Colosimo, and M. Grasso. Domain-selective functional analysis of variance for supervised statistical profile monitoring of signal data. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 67(1):55–81, 2018.

[123] C. Preda, G. Saporta, and C. Lévéder. Pls classification of functional data. *Computational Statistics*, 22(2):223–235, 2007.

[124] Z. Qiu, J. Chen, and J.-T. Zhang. Two-sample tests for multivariate functional data with applications. *Computational Statistics & Data Analysis*, 157:107160, 2021.

[125] J. Ramsay and B. Silverman. *Functional data analysis*. Springer, 1997.

[126] J. Ramsay and B. Silverman. Principal components analysis for functional data. *Functional data analysis*, pages 147–172, 2005.

[127] J. O. Ramsay and X. Li. Curve registration. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 60(2):351–363, 1998.

[128] J. O. Ramsay and J. B. Ramsey. Functional data analysis of the dynamics of the monthly index of nondurable goods production. *Journal of Econometrics*, 107(1-2): 327–344, 2002.

[129] J. O. Ramsay and B. W. Silverman. Functional data analysis. *İnternet Adresi: http*, 2008.

[130] P. T. Reiss, L. Huang, and M. Mennes. Fast function-on-scalar regression with penalized basis expansions. *The international journal of biostatistics*, 6(1), 2010.

[131] B. Ripley, W. Venables, and M. B. Ripley. Package 'nnet'. *R package version*, 7 (3-12):700, 2016.

[132] B. B. Rønn. Nonparametric maximum likelihood estimation for shifted curves. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 63(2): 243–259, 2001.

[133] E. L. Rosenberg and P. Ekman. *What the face reveals: Basic and applied studies of spontaneous expression using the Facial Action Coding System (FACS)*. Oxford University Press, 2020.

[134] L. Salmaso and F. Pesarin. *Permutation tests for complex data: theory, applications and software*. John Wiley & Sons, 2010.

[135] A. Saxena, A. Khanna, and D. Gupta. Emotion recognition and detection methods: A comprehensive survey. *Journal of Artificial Intelligence and Systems*, 2(1):53–79, 2020.

[136] F. Scheipl, A.-M. Staicu, and S. Greven. Functional additive mixed models. *Journal of Computational and Graphical Statistics*, 24(2):477–501, 2015.

[137] L. Shu, Y. Yu, W. Chen, H. Hua, Q. Li, J. Jin, and X. Xu. Wearable emotion recognition using heart rate data from a smart bracelet. *Sensors*, 20(3):718, 2020.

[138] R. W. Simon. Sociological scholarship on gender differences in emotion and emotional well-being in the united states: A snapshot of the field. *Emotion Review*, 6 (3):196–201, 2014.

[139] R. W. Simon and L. E. Nath. Gender and emotion in the united states: Do men and women differ in self-reports of feelings and expressive behavior? *American journal of sociology*, 109(5):1137–1176, 2004.

[140] M. Spezialetti, G. Placidi, and S. Rossi. Emotion recognition for human-robot interaction: Recent advances and future perspectives. *Frontiers in Robotics and AI*, 7:532279, 2020.

[141] A. Srivastava and E. P. Klassen. *Functional and shape data analysis*, volume 1. Springer, 2016.

[142] A. Srivastava, W. Wu, S. Kurtek, E. Klassen, and J. S. Marron. Registration of functional data using fisher-rao metric. *arXiv preprint arXiv:1103.3817*, 2011.

[143] J. Strait, O. Chkrebtii, and S. Kurtek. Automatic detection and uncertainty quantification of landmarks on elastic curves. *Journal of the American Statistical Association*, 114(527):1002–1017, 2019.

[144] K. Sun and J. Yu. Video affective content representation and recognition using video affective tree and hidden markov models. In *International Conference on Affective Computing and Intelligent Interaction*, pages 594–605. Springer, 2007.

[145] R. Tang and H.-G. Müller. Pairwise curve synchronization for functional data. *Biometrika*, 95(4):875–889, 2008.

[146] J. Tarrío-Saavedra, S. Naya, M. Francisco-Fernández, J. López-Beceiro, and R. Artiaga. Functional nonparametric classification of wood species from thermal data. *Journal of thermal analysis and calorimetry*, 104(1):87–100, 2011.

[147] R. Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1):267–288, 1996.

[148] S. Tokushige, H. Yadohisa, and K. Inada. Crisp and fuzzy k-means clustering algorithms for multivariate functional data. *Computational Statistics*, 22(1):1–16, 2007.

[149] A. Torrente, S. López-Pintado, and J. Romo. Depthtools: an r package for a robust analysis of gene expression data. *BMC bioinformatics*, 14(1):1–11, 2013.

[150] J. D. Tucker, W. Wu, and A. Srivastava. Generative models for functional data using phase and amplitude separation. *Computational Statistics & Data Analysis*, 61:50–66, 2013.

[151] J. D. Tucker, J. R. Lewis, and A. Srivastava. Elastic functional principal component regression. *Statistical Analysis and Data Mining: The ASA Data Science Journal*, 12(2):101–115, 2019.

[152] J. D. Tucker, L. Shand, and K. Chowdhary. Multimodal bayesian registration of noisy functions using hamiltonian monte carlo. *Computational Statistics & Data Analysis*, 163:107298, 2021.

[153] S. Tulyakov, M.-Y. Liu, X. Yang, and J. Kautz. Mocogan: Decomposing motion and content for video generation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1526–1535, 2018.

[154] S. Ullah and C. F. Finch. Applications of functional data analysis: A systematic review. *BMC medical research methodology*, 13(1):1–12, 2013.

[155] M. J. Valderrama. An overview to modelling functional data, 2007.

[156] C. Vilchis and M. Gonzalez-Mendoza. Democratization of real-time facial tracking frameworks for digital humans.

[157] J.-L. Wang, J.-M. Chiou, and H.-G. Müller. Functional data analysis. *Annual Review of Statistics and Its Application*, 3:257–295, 2016.

[158] K. Wang and T. Gasser. Alignment of curves by dynamic time warping. *The annals of Statistics*, 25(3):1251–1276, 1997.

[159] K. Wang, X. Peng, J. Yang, S. Lu, and Y. Qiao. Suppressing uncertainties for large-scale facial expression recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6897–6906, 2020.

[160] S. Wang and Q. Ji. Video affective content analysis: a survey of state-of-the-art methods. *IEEE Transactions on Affective Computing*, 6(4):410–430, 2015.

[161] S. C. Watanapa, B. Thipakorn, and N. Charoenkitkarn. A sieving ann for emotion-based movie clip classification. *IEICE transactions on information and systems*, 91 (5):1562–1572, 2008.

[162] R. K. Wong, Y. Li, and Z. Zhu. Partially linear functional additive models for multivariate functional data. *Journal of the American Statistical Association*, 114 (525):406–418, 2019.

[163] T.-T. Wong. Performance evaluation of classification algorithms by k-fold and leave-one-out cross validation. *Pattern Recognition*, 48(9):2839–2846, 2015.

[164] J. Wrobel and A. Bauer. registr 2.0: Incomplete curve registration for exponential family functional data. *Journal of Open Source Software*, 6(61):2964, 2021.

[165] J. Wrobel, V. Zipunnikov, J. Schrack, and J. Goldsmith. Registration for exponential family functional data. *Biometrics*, 75(1):48–57, 2019.

[166] W. Wu and A. Srivastava. Analysis of spike train data: Alignment and comparisons using the extended fisher-rao metric. *Electronic Journal of Statistics*, 8(2):1776–1785, 2014.

[167] W. M. Wundt. *Grundriss der psychologie*. A. Kröner, 1913.

[168] Y. Xu, Y. Li, and D. Nettleton. Nested hierarchical functional data modeling and inference for the analysis of functional plant phenotypes. *Journal of the American Statistical Association*, 113(522):593–606, 2018.

[169] M. Yamamoto and Y. Terada. Functional factorial k-means analysis. *Computational statistics & data analysis*, 79:133–148, 2014.

[170] J. Yang, K. Wang, X. Peng, and Y. Qiao. Deep recurrent multi-instance learning with spatio-temporal features for engagement intensity prediction. In *Proceedings of the 20th ACM international conference on multimodal interaction*, pages 594–598, 2018.

[171] A. Yazdani, E. Skodras, N. Fakotakis, and T. Ebrahimi. Multimedia content analysis for emotional characterization of music video clips. *EURASIP Journal on Image and Video Processing*, 2013(1):1–10, 2013.

[172] J. Zhang. Analysis of variance for functional data. *Monographs on statistics and applied probability*, 127:127, 2014.

[173] J. Zhang, G. J. Siegle, T. Sun, W. D'andrea, and R. T. Krafty. Interpretable principal component analysis for multilevel multivariate functional data. *Biostatistics*, 2021.

[174] J.-T. Zhang and X. Liang. One-way anova for functional data via globalizing the pointwise f-test. *Scandinavian Journal of Statistics*, 41(1):51–71, 2014.

[175] Z. Zhang, E. Klassen, and A. Srivastava. Phase-amplitude separation and modeling of spherical trajectories. *Journal of Computational and Graphical Statistics*, 27(1): 85–97, 2018.

[176] Z. Zhang, X. Wang, L. Kong, and H. Zhu. High-dimensional spatial quantile function-on-scalar regression. *Journal of the American Statistical Association*, pages 1–16, 2021.

[177] P. Zhao and B. Yu. On model selection consistency of lasso. *The Journal of Machine Learning Research*, 7:2541–2563, 2006.

[178] H. Zhu, J. S. Morris, F. Wei, and D. D. Cox. Multivariate functional response regression, with application to fluorescence spectroscopy in a cervical pre-cancer study. *Computational statistics & data analysis*, 111:88–101, 2017.

Appendix A

Methods and Extended Results

A.1 Brief Introduction of Modified Band Depth on Functional Data

The notion of depth for functional data is based on the graphic representation of the functions. Given a collection of curves, the idea is to measure the centrality of a function, and the depth of curves provides a natural center-outward order for the sample functions [104]. In this way, we can find the most representative (or deepest) samples within collections of observations in different classes, generalizing thus the concept of "median" in a functional setting.

Modified Band Depth (MBD) is the proportion of time in which a function is inside the band formed by any other two functions [105]. More formally, for any of the functions x in a sample $x_1, \ldots, x_n$, let

$$A_j(x) := A(x; x_{i_1}, \ldots, x_{i_j}) := \{t \in I : \min_{r=i_1,\ldots,i_j} x_r(t) \leq x(t) \leq \max_{r=i_1,\ldots,i_j} x_r(t)\}, j \geq 2,$$

be the set of points in the interval I where the function x is in the band determined by the observations $x_{i_1}, \ldots, x_{i_j}$. If λ is the Lebesgue measure on I, the proportion of time when x is in the band is: $\lambda_r(A_j(x)) = \frac{\lambda(A_j(x))}{\lambda(I)}$.

Note that normally the increase of j will not disturb the selection of the median function, so we can set $j = 2$ for simplicity, and we have

$$MBD(x) = \binom{n}{2}^{-1} \frac{1}{\lambda(I)} \sum_{1 \le i_1 \le i_2 \le n} \lambda(A_2(x)).$$

A.2 Results related with Chapter 1: Emotion Recognition and Expression Detection from Facial Video Data

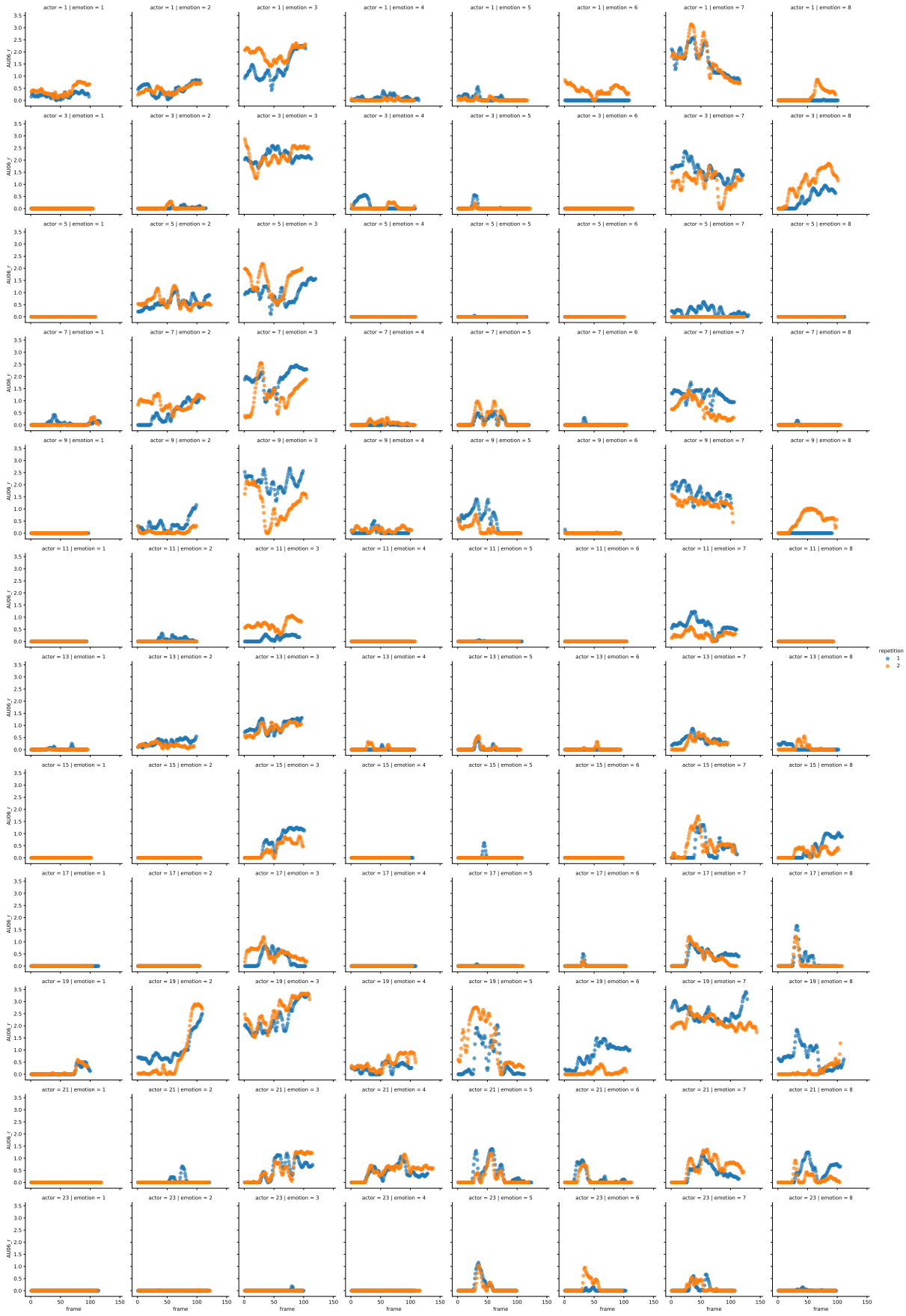


Figure A.1: The evolution curves of AU06 detected from the videos of all male actors under different emotions. The behavior of actor 1 in Figure 1.7a is involved in the first line.

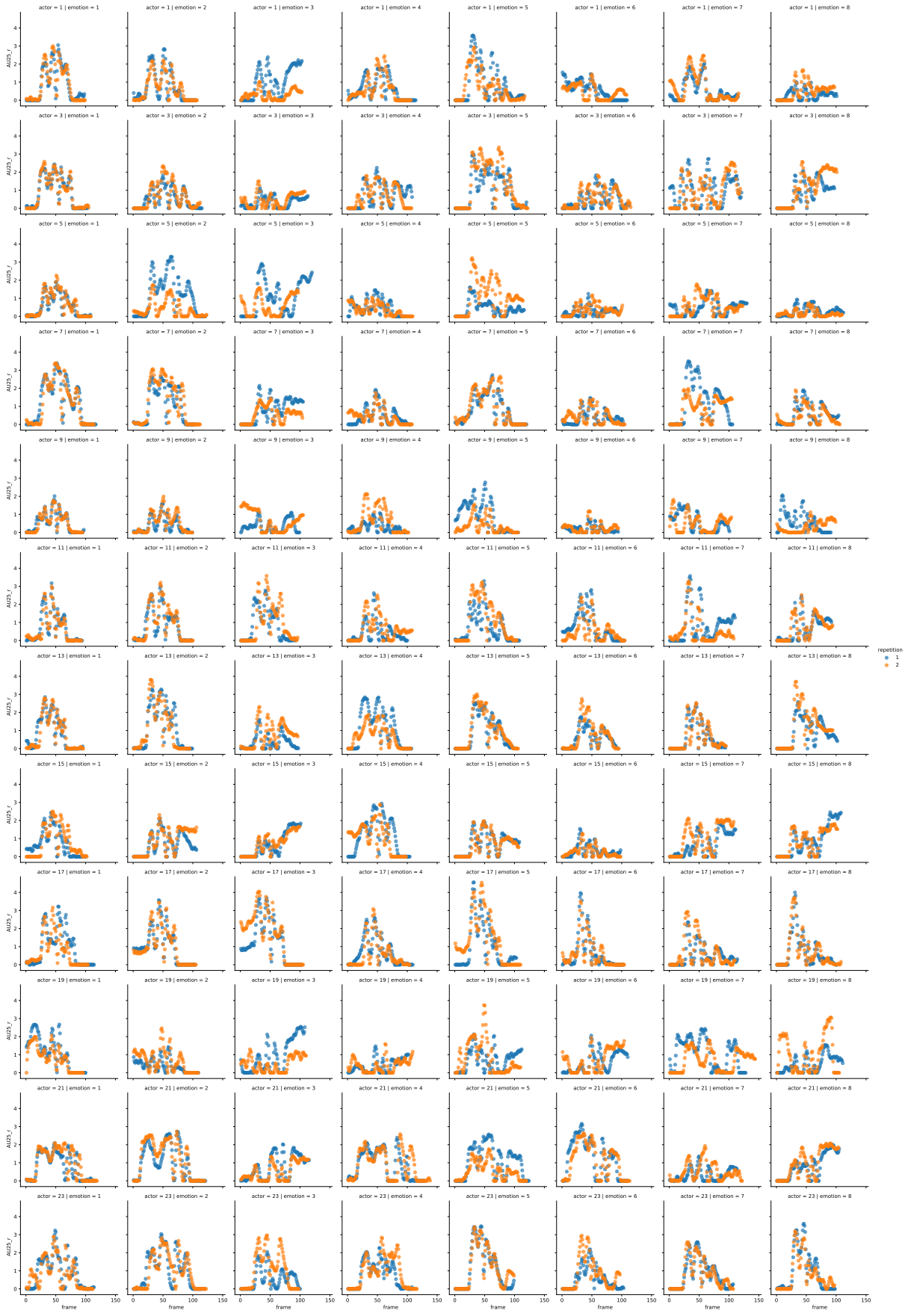


Figure A.2: The evolution curves of AU25 detected from the videos of all male actors under different emotions.The behavior of actor 1 in Figure 1.7b is involved in the first line.

A.3 Results related with Chapter 2: Data Preprocessing: Functional Data Format and Functional Curve Registration

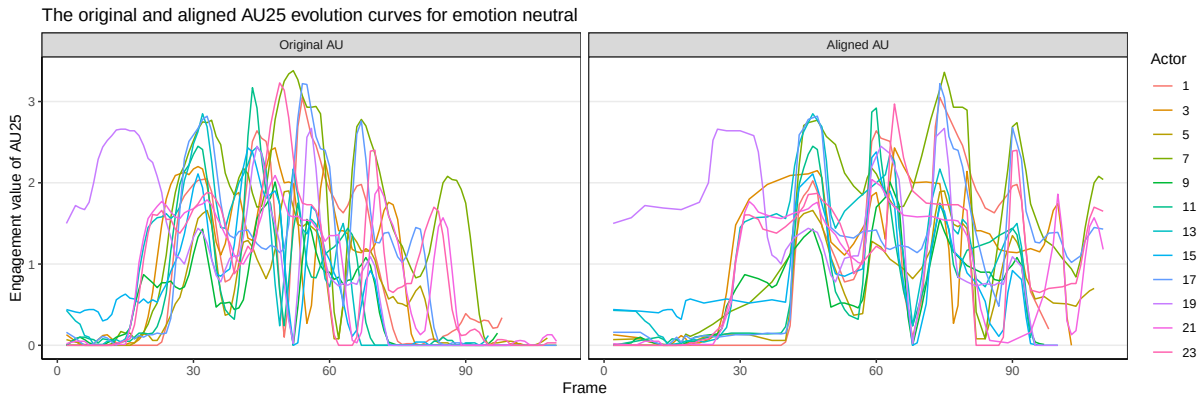


Figure A.3: Example of registration of the curves of AU25 for the male actors representing the emotion neutral in the corresponding video.

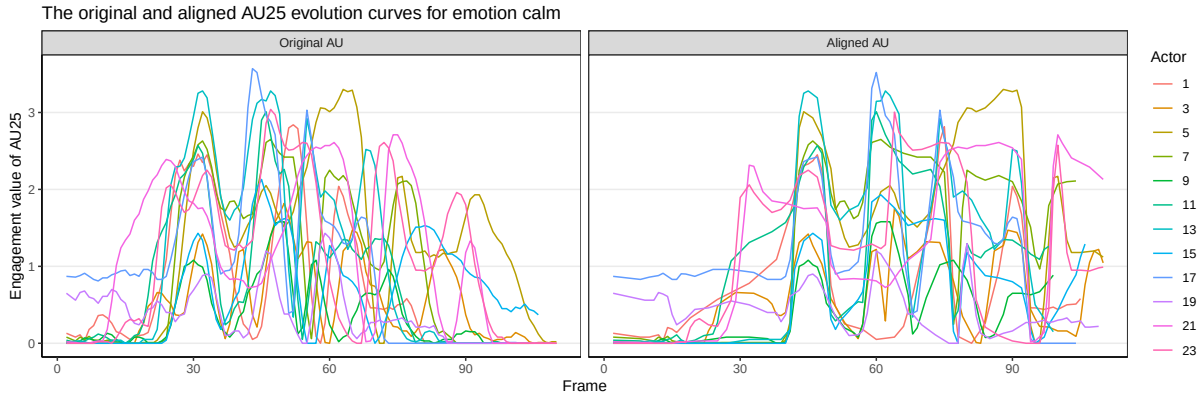


Figure A.4: Example of registration of the curves of AU25 for the male actors representing the emotion calm in the corresponding video.

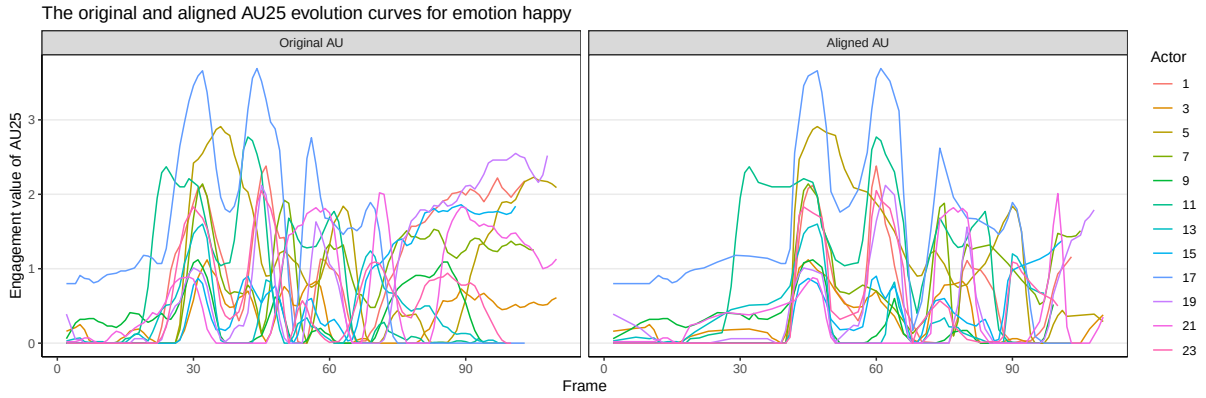


Figure A.5: Example of registration of the curves of AU25 for the male actors representing the emotion happy in the corresponding video.

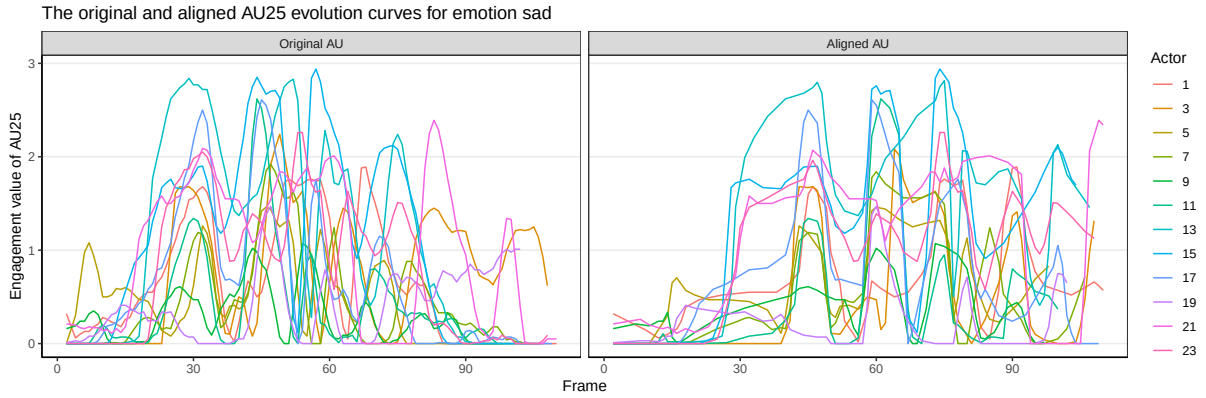


Figure A.6: Example of registration of the curves of AU25 for the male actors representing the emotion sad in the corresponding video.

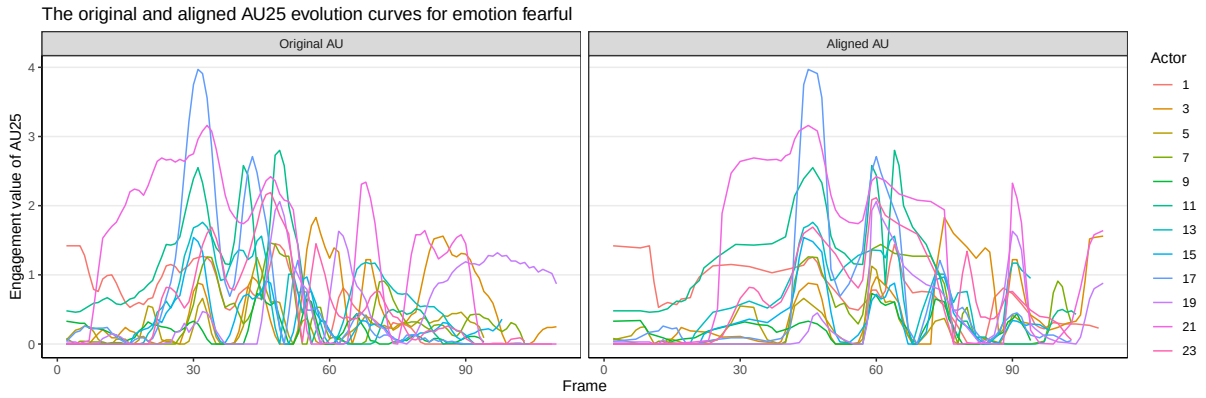


Figure A.7: Example of registration of the curves of AU25 for the male actors representing the emotion fearful in the corresponding video.

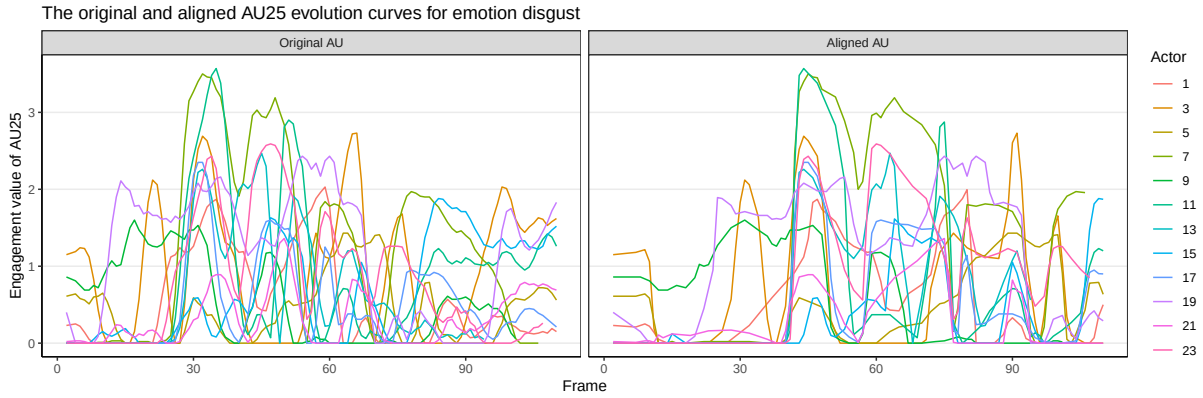


Figure A.8: Example of registration of the curves of AU25 for the male actors representing the emotion disgust in the corresponding video.

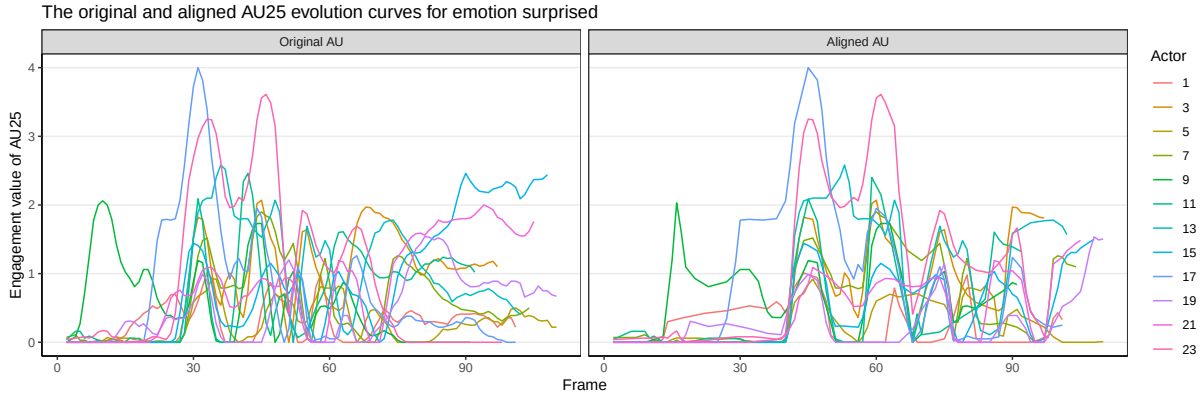
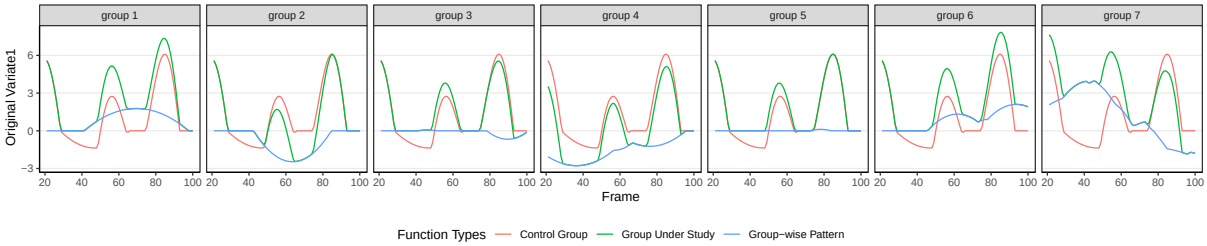
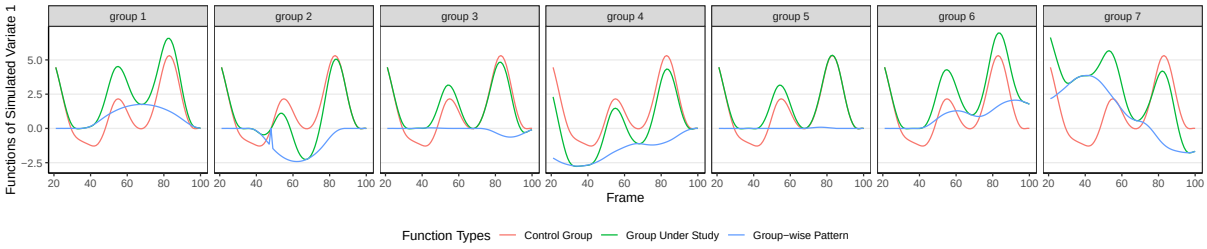


Figure A.9: Example of registration of the curves of AU25 for the male actors representing the emotion surprised in the corresponding video.

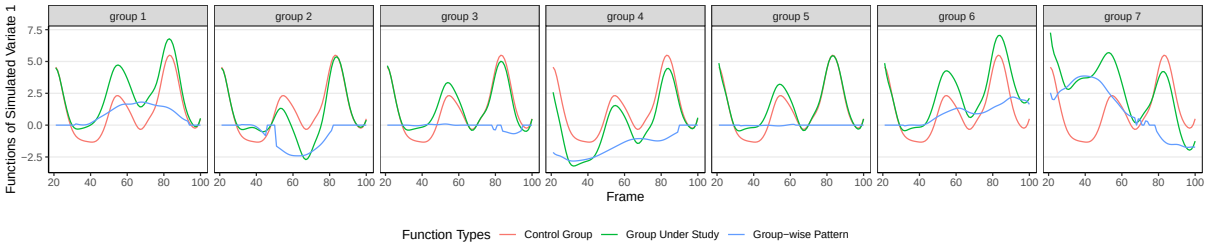
A.4 Results related with Chapter 3: Model Construction: Group-wise Effect Tests of Multiple Multivariate Function-on-scalar Regression



(a) The means of the original simulated data

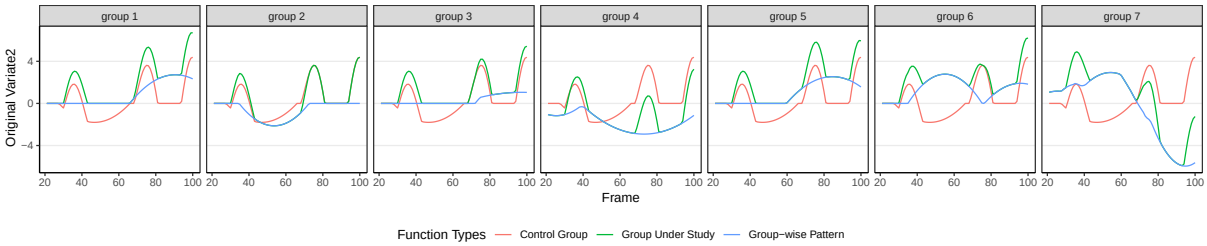


(b) Estimation results based on permutation test with the standard deviation of data set as 0.05.

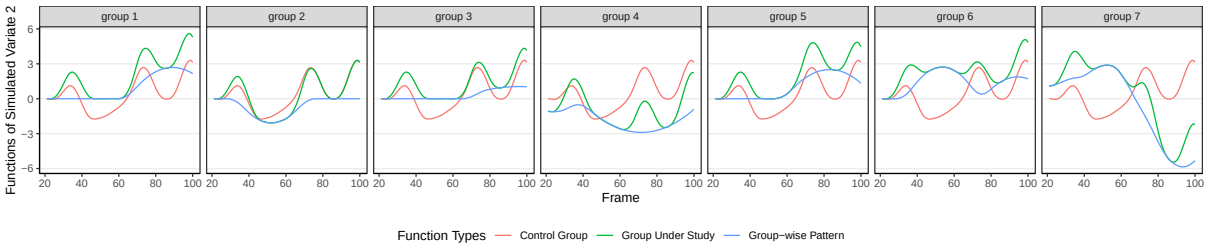


(c) Estimation results based on permutation test with the standard deviation of data set as 2.

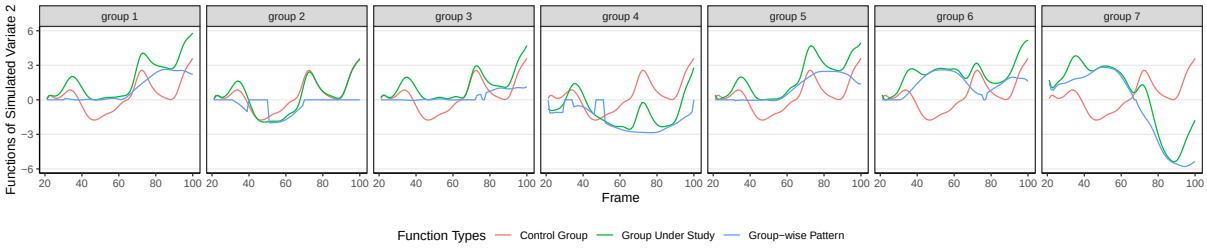
Figure A.10: The results of the estimation of the functional mean and of the group-wise mean effects through the permutation test, using a sample size of 24 and a quantile $q = 1 - \alpha = 90\%$. These are the results for variate 2 in the simulated case.



(a) The means of the original simulated data

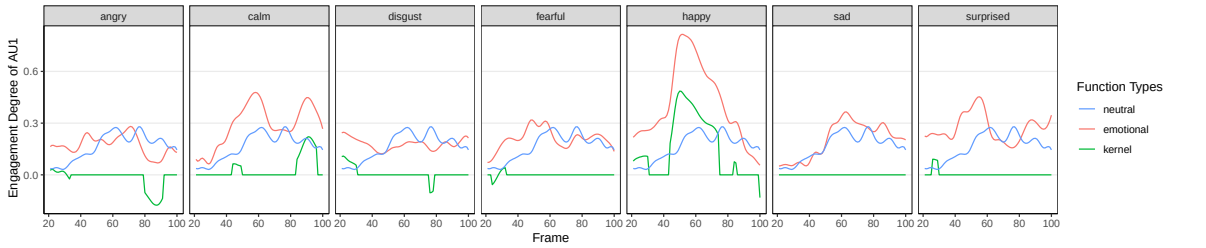


(b) Estimation results based on permutation test with the standard deviation of data set as 0.05.

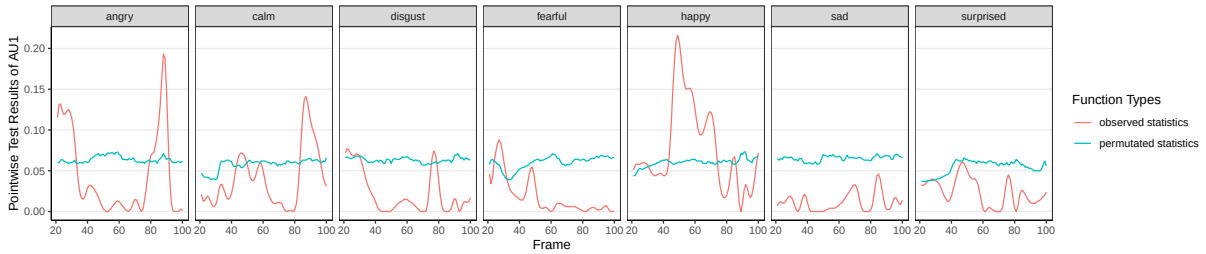


(c) Estimation results based on permutation test with the standard deviation of data set as 2.

Figure A.11: The results of the estimation of the functional mean and of the group-wise mean effects through the permutation test, using a sample size of 24 and a quantile $q = 1 - \alpha = 90\%$. These are the results for variate 3 in the simulated case.

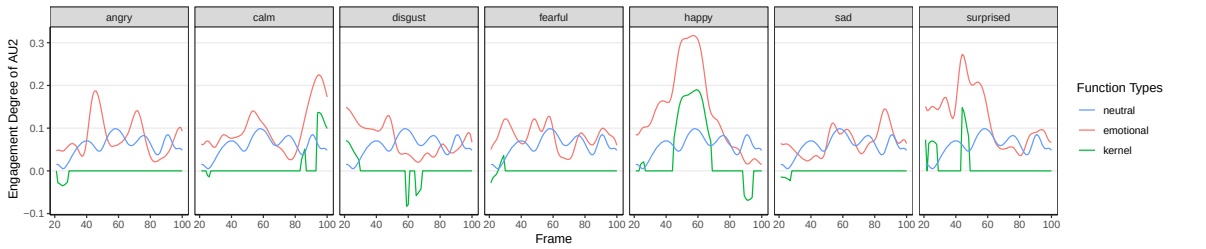


(a) Group-wise patterns estimated via group comparisons.

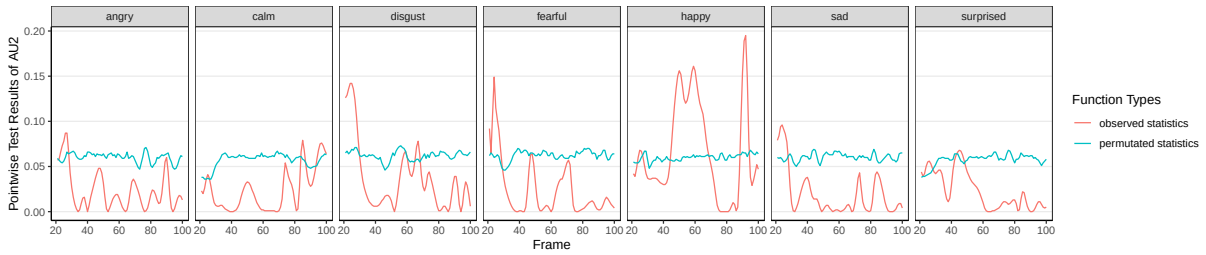


(b) FANOVA test results of observed and permutated statistics.

Figure A.12: The Group-wise patterns of AU01 under neutral and seven emotions, together with the corresponding FANOVA test statistics

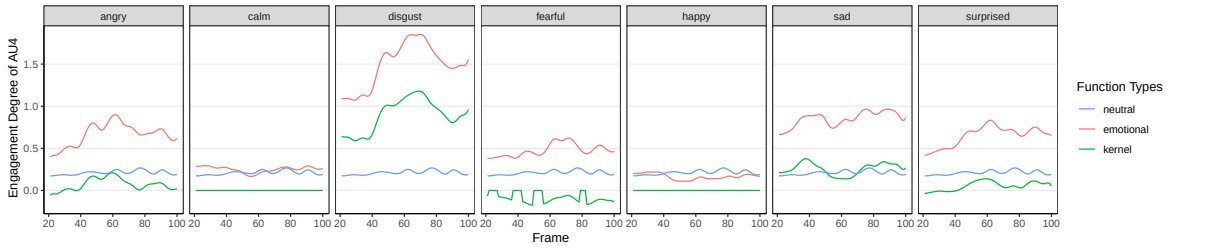


(a) Group-wise patterns estimated via group comparisons.

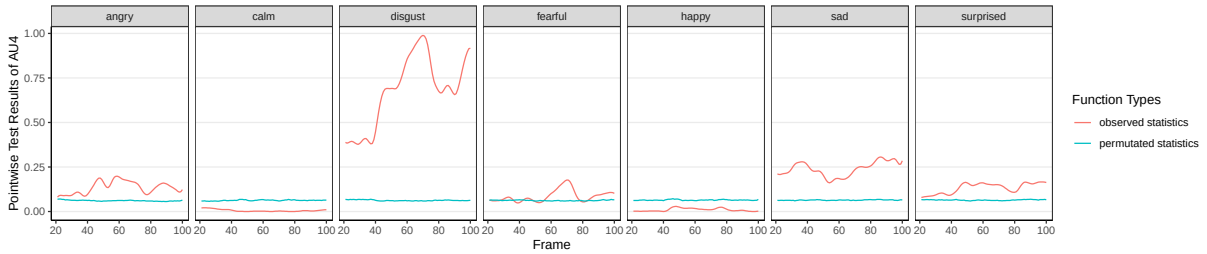


(b) FANOVA test results of observed and permutated statistics.

Figure A.13: The Group-wise patterns of AU02 under neutral and seven emotions, together with the corresponding FANOVA test statistics

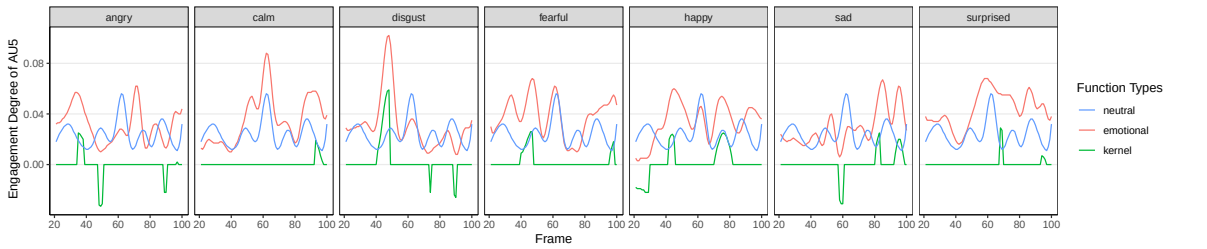


(a) Group-wise patterns estimated via group comparisons.

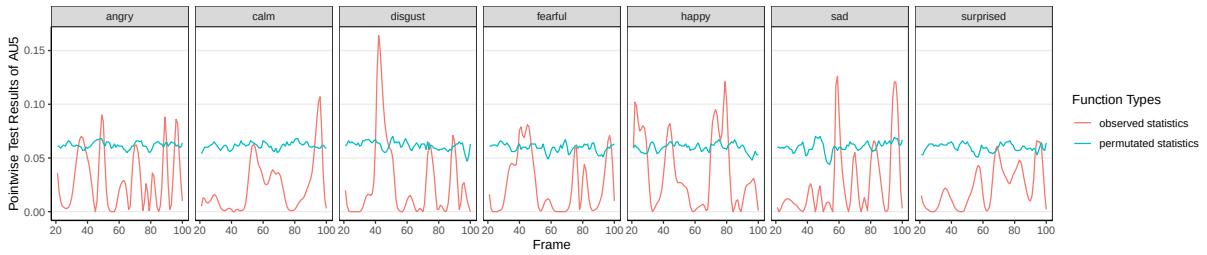


(b) FANOVA test results of observed and permutated statistics.

Figure A.14: The Group-wise patterns of AU04 under neutral and seven emotions, together with the corresponding FANOVA test statistics

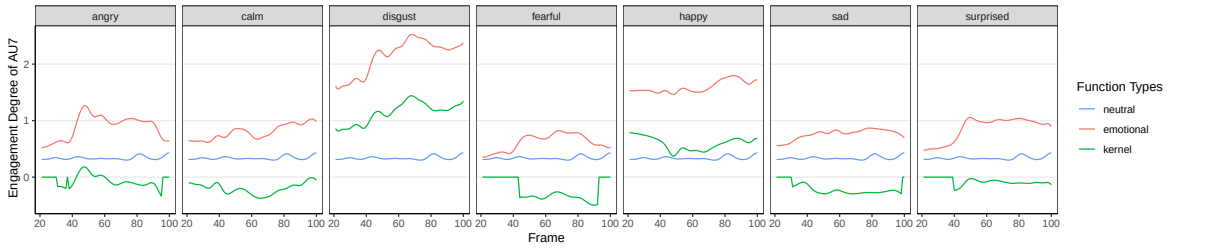


(a) Group-wise patterns estimated via group comparisons.

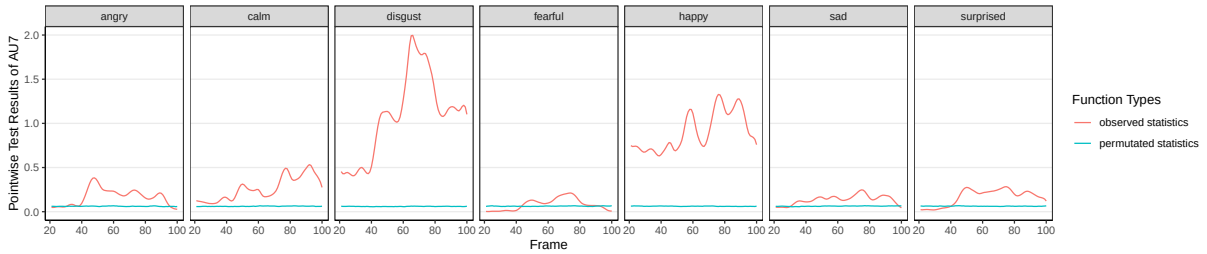


(b) FANOVA test results of observed and permutated statistics.

Figure A.15: The Group-wise patterns of AU05 under neutral and seven emotions, together with the corresponding FANOVA test statistics

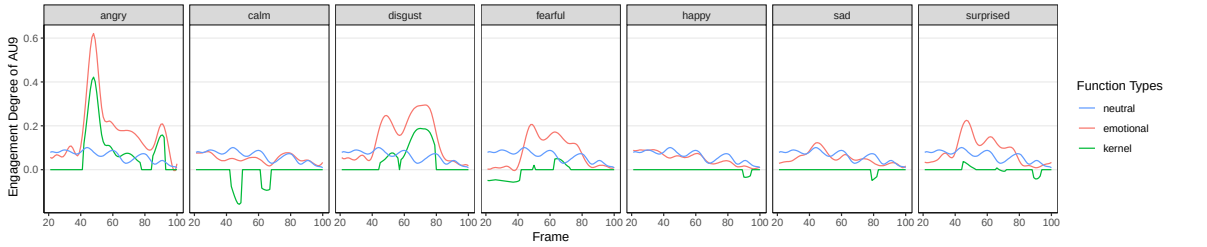


(a) Group-wise patterns estimated via group comparisons.

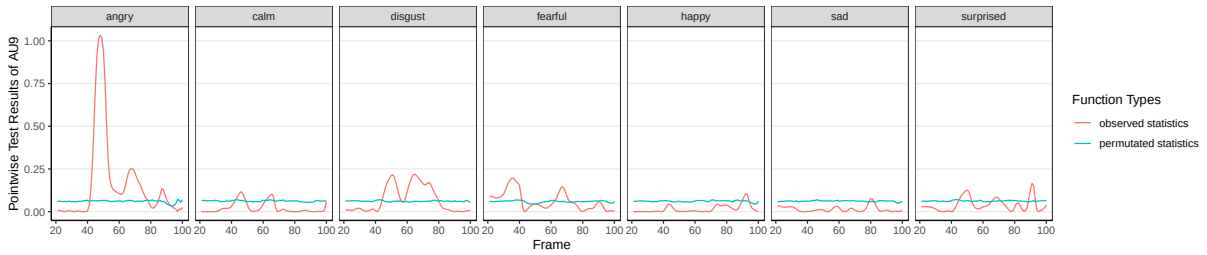


(b) FANOVA test results of observed and permutated statistics.

Figure A.16: The Group-wise patterns of AU07 under neutral and seven emotions, together with the corresponding FANOVA test statistics

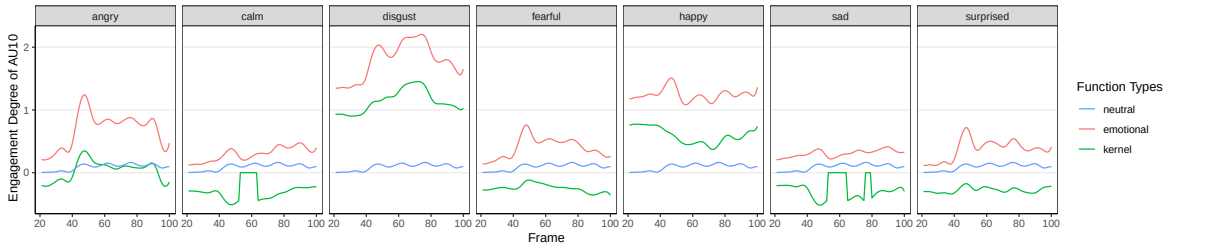


(a) Group-wise patterns estimated via group comparisons.

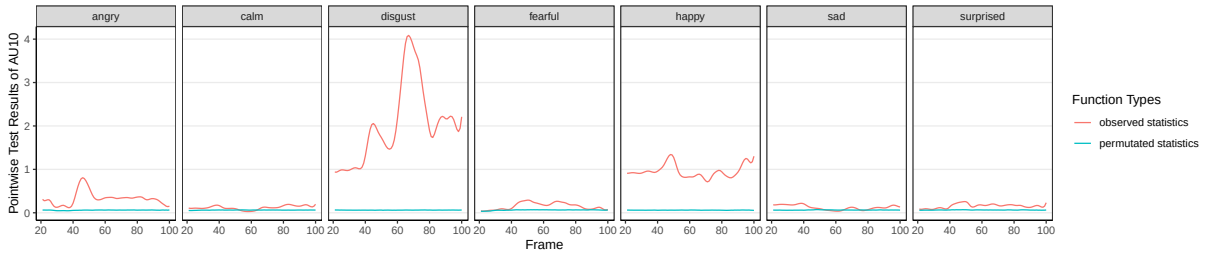


(b) FANOVA test results of observed and permutated statistics.

Figure A.17: The Group-wise patterns of AU09 under neutral and seven emotions, together with the corresponding FANOVA test statistics

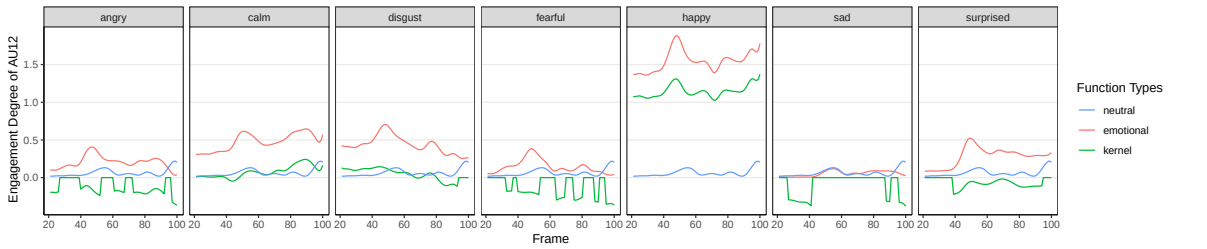


(a) Group-wise patterns estimated via group comparisons.

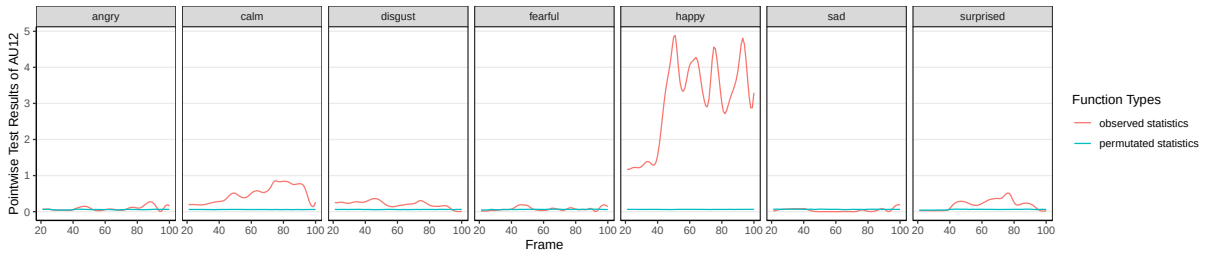


(b) FANOVA test results of observed and permutated statistics.

Figure A.18: The Group-wise patterns of AU10 under neutral and seven emotions, together with the corresponding FANOVA test statistics

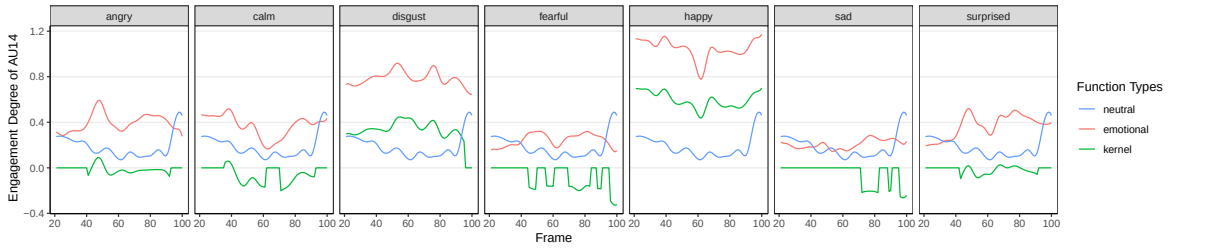


(a) Group-wise patterns estimated via group comparisons.

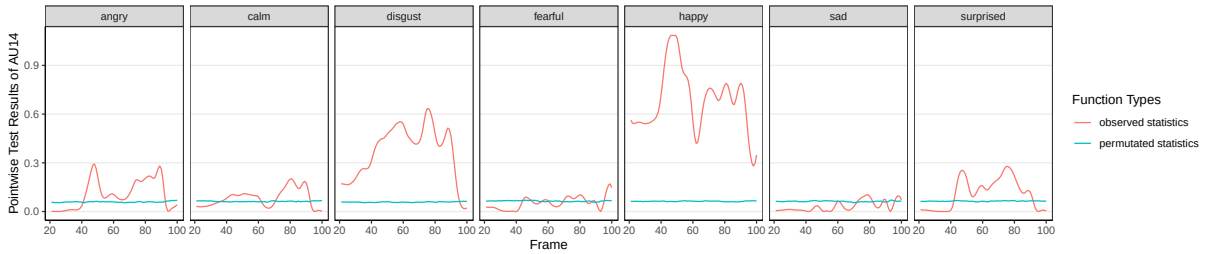


(b) FANOVA test results of observed and permutated statistics.

Figure A.19: The Group-wise patterns of AU12 under neutral and seven emotions, together with the corresponding FANOVA test statistics

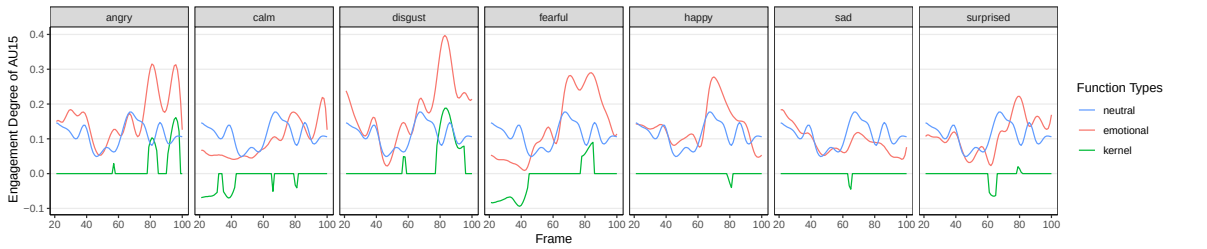


(a) Group-wise patterns estimated via group comparisons.

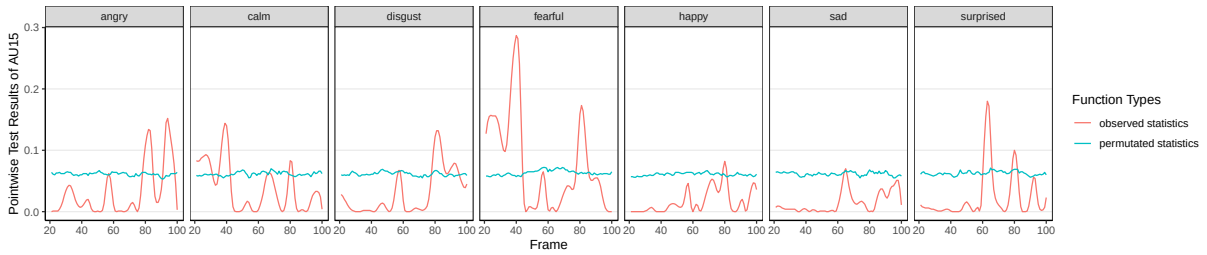


(b) FANOVA test results of observed and permutated statistics.

Figure A.20: The Group-wise patterns of AU14 under neutral and seven emotions, together with the corresponding FANOVA test statistics

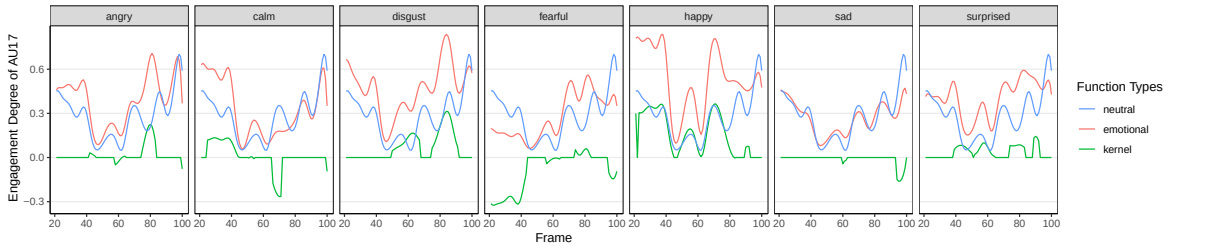


(a) Group-wise patterns estimated via group comparisons.

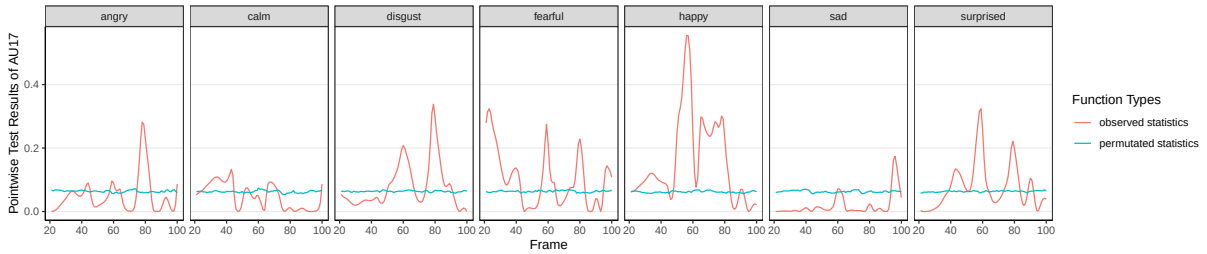


(b) FANOVA test results of observed and permutated statistics.

Figure A.21: The Group-wise patterns of AU15 under neutral and seven emotions, together with the corresponding FANOVA test statistics

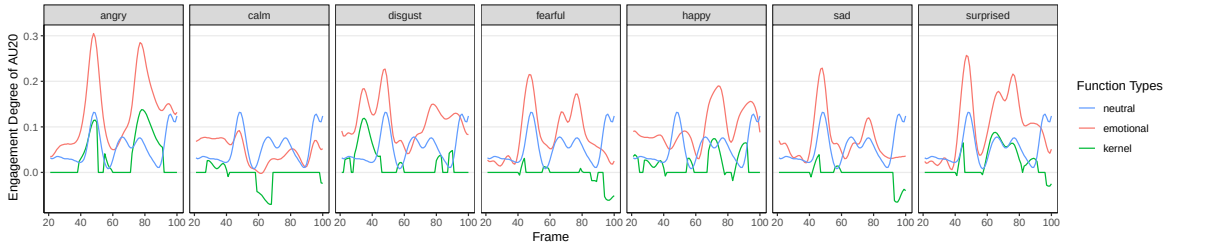


(a) Group-wise patterns estimated via group comparisons.

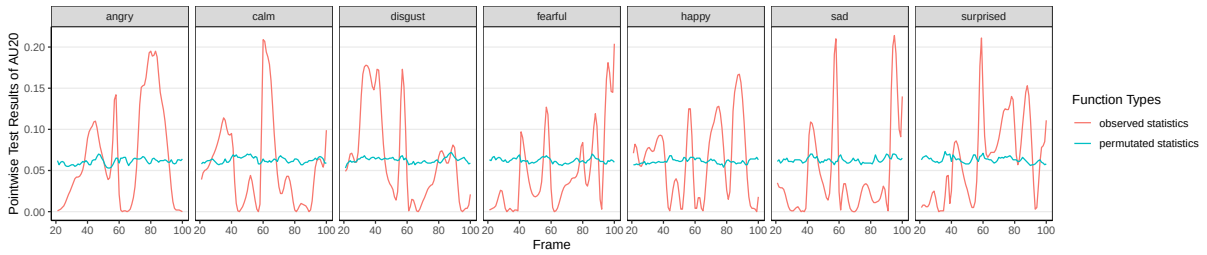


(b) FANOVA test results of observed and permutated statistics.

Figure A.22: The Group-wise patterns of AU17 under neutral and seven emotions, together with the corresponding FANOVA test statistics

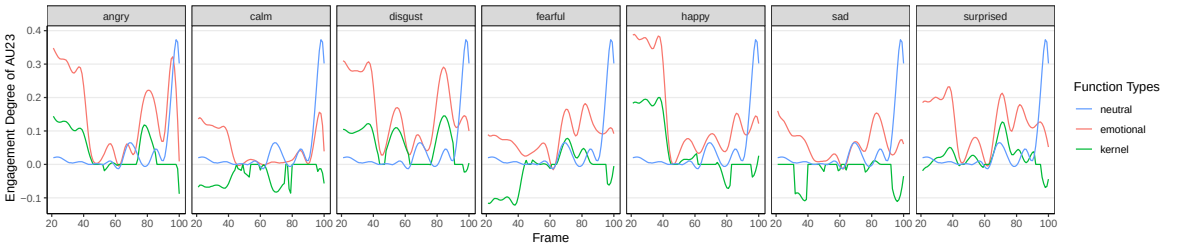


(a) Group-wise patterns estimated via group comparisons.

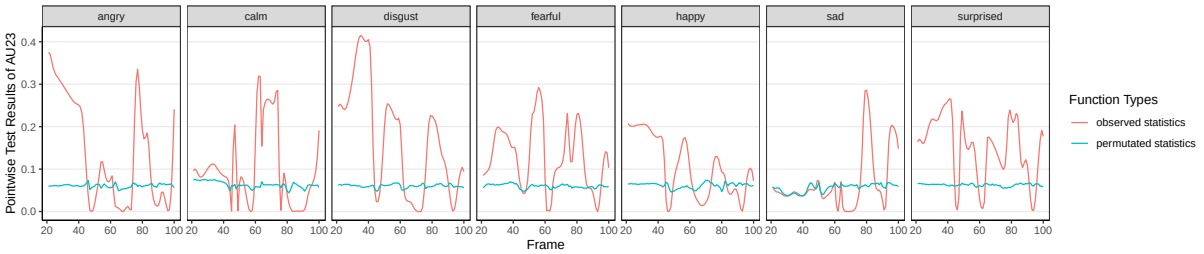


(b) FANOVA test results of observed and permutated statistics.

Figure A.23: The Group-wise patterns of AU20 under neutral and seven emotions, together with the corresponding FANOVA test statistics

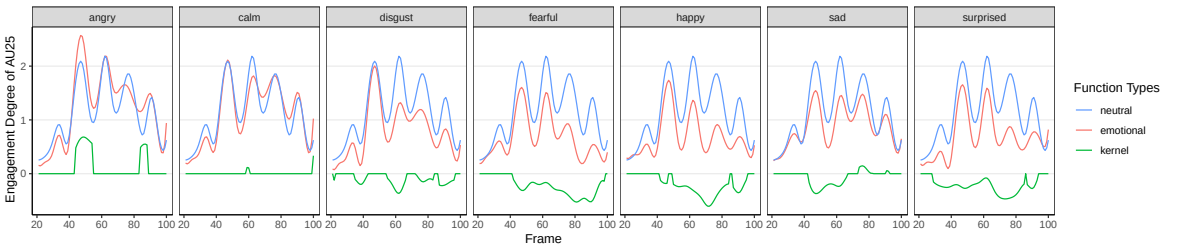


(a) Group-wise patterns estimated via group comparisons.

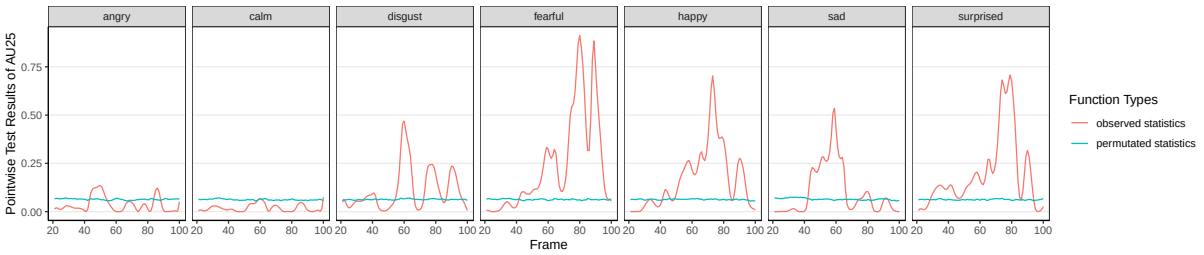


(b) FANOVA test results of observed and permutated statistics.

Figure A.24: The Group-wise patterns of AU23 under neutral and seven emotions, together with the corresponding FANOVA test statistics

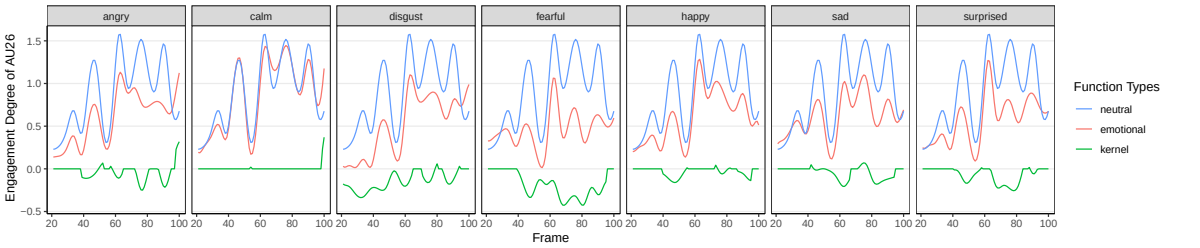


(a) Group-wise patterns estimated via group comparisons.

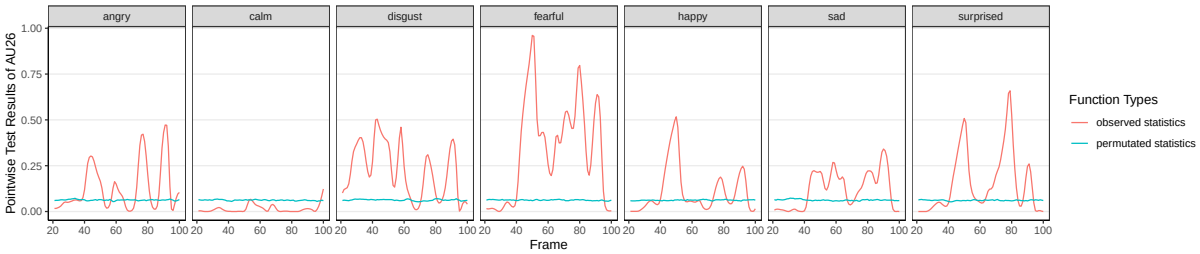


(b) FANOVA test results of observed and permutated statistics.

Figure A.25: The Group-wise patterns of AU25 under neutral and seven emotions, together with the corresponding FANOVA test statistics

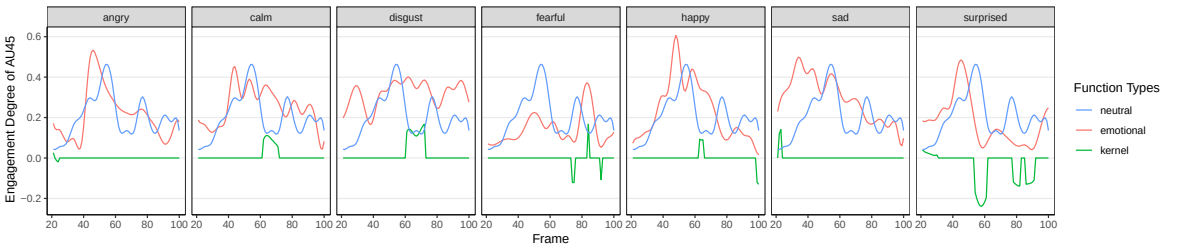


(a) Group-wise patterns estimated via group comparisons.

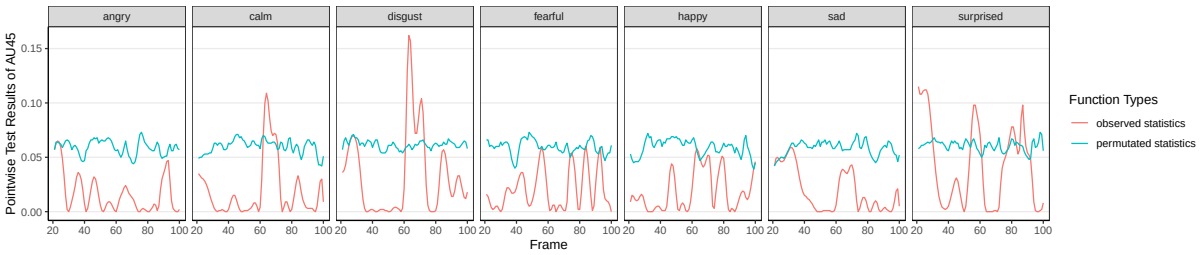


(b) FANOVA test results of observed and permutated statistics.

Figure A.26: The Group-wise patterns of AU26 under neutral and seven emotions, together with the corresponding FANOVA test statistics

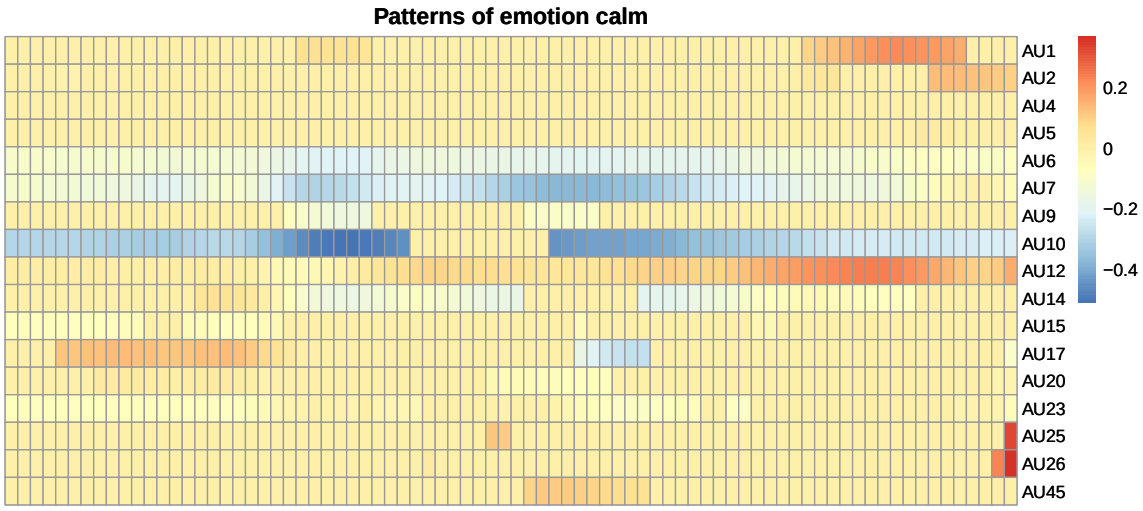


(a) Group-wise patterns estimated via group comparisons.

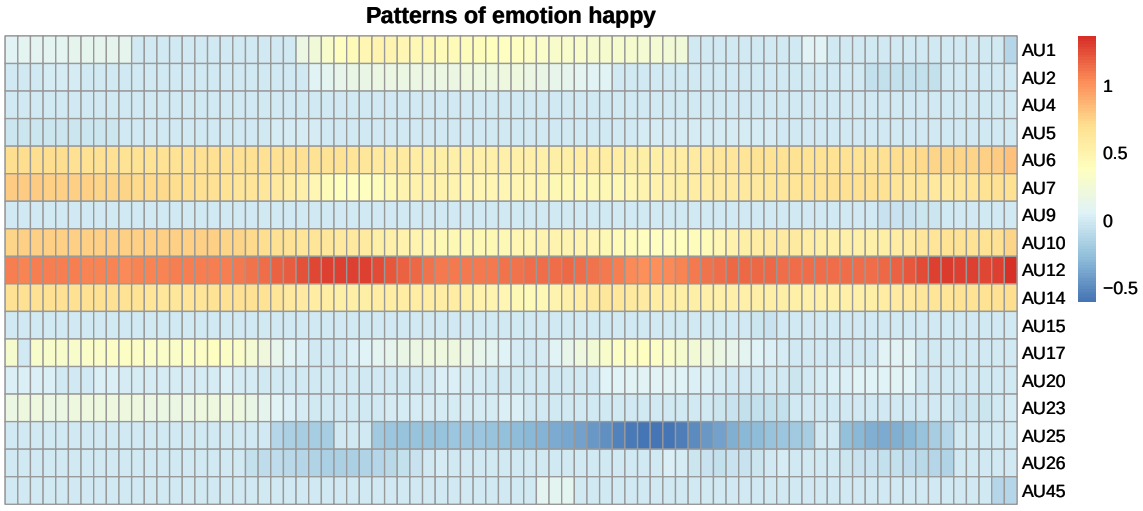


(b) FANOVA test results of observed and permutated statistics.

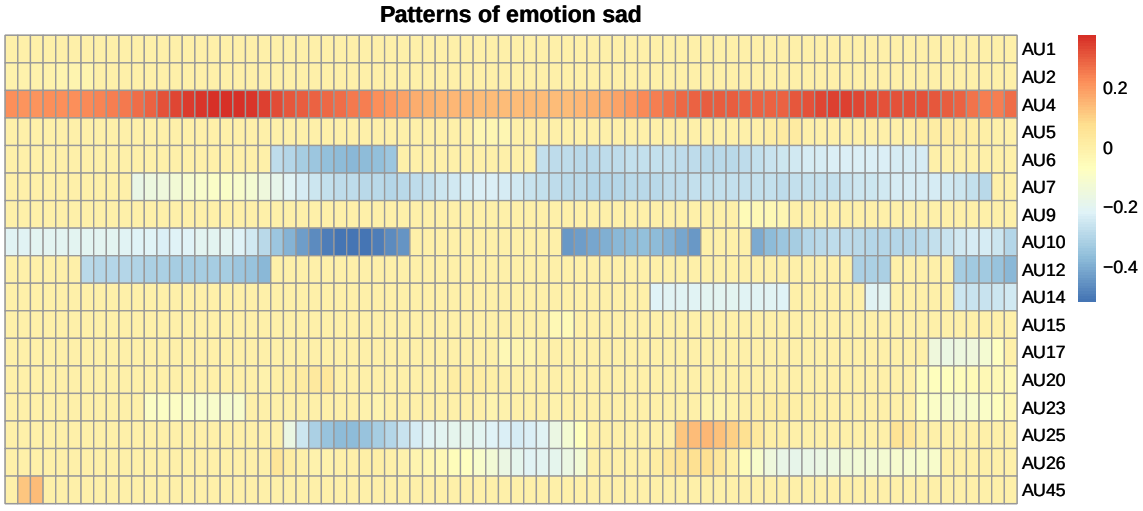
Figure A.27: The Group-wise patterns of AU45 under neutral and seven emotions, together with the corresponding FANOVA test statistics



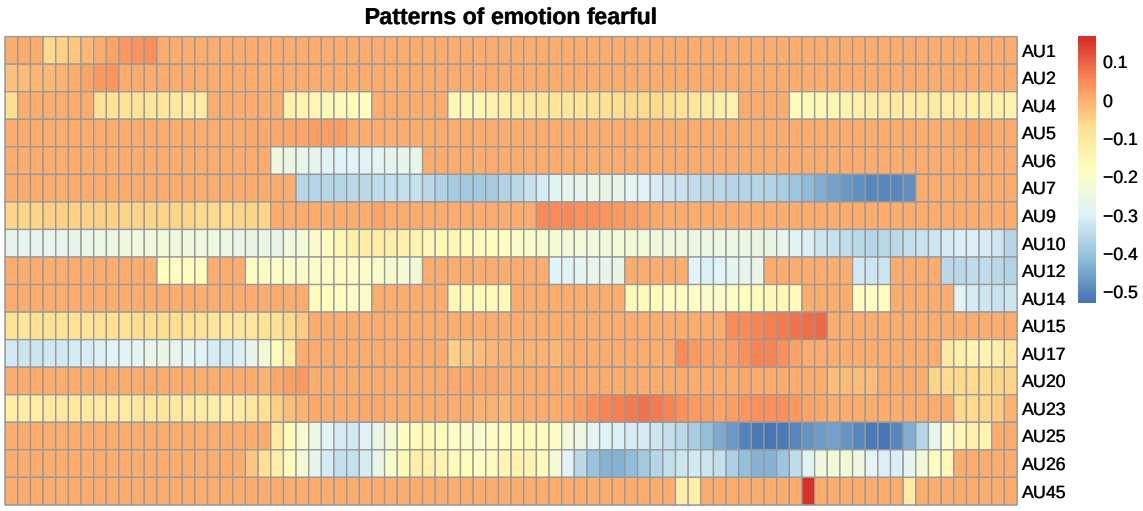
(a) Heatmap of group-wise patterns detected for emotion calm.



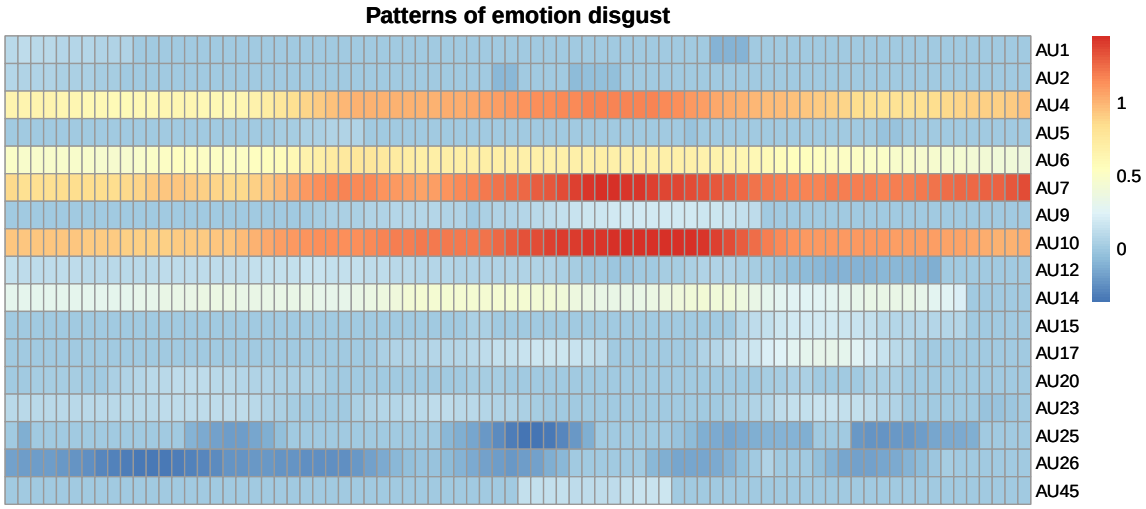
(b) Heatmap of group-wise patterns detected for emotion happy.



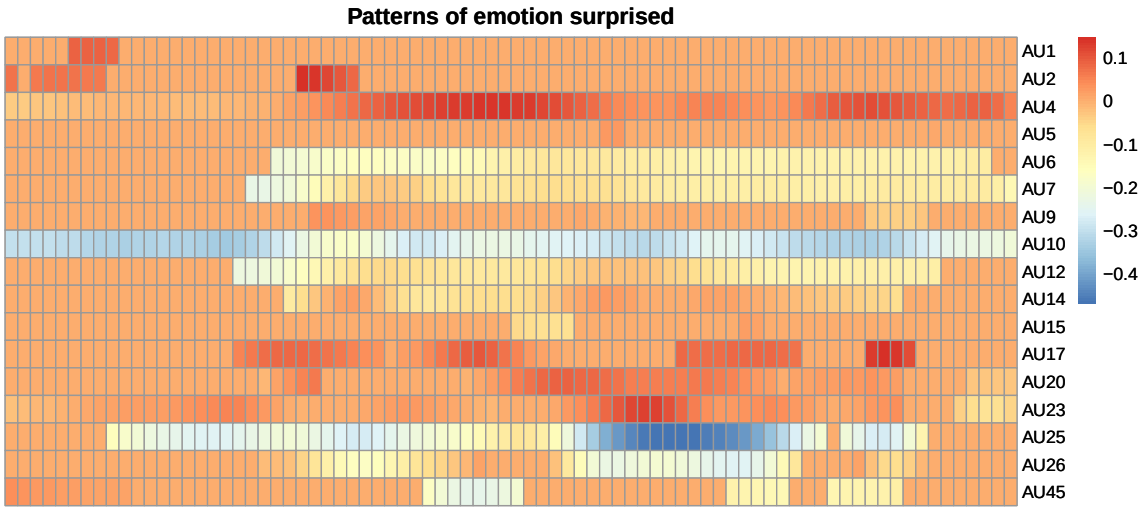
(c) Heatmap of group-wise patterns detected for emotion sad.



(a) Heatmap of group-wise patterns detected for emotion fearful.



(b) Heatmap of group-wise patterns detected for emotion disgusted.



(c) Heatmap of group-wise patterns detected for emotion surprised.

A.5 Results related with Chapter 5: Framework for Multivariate Multi-class Functions: Summary and Conclusions

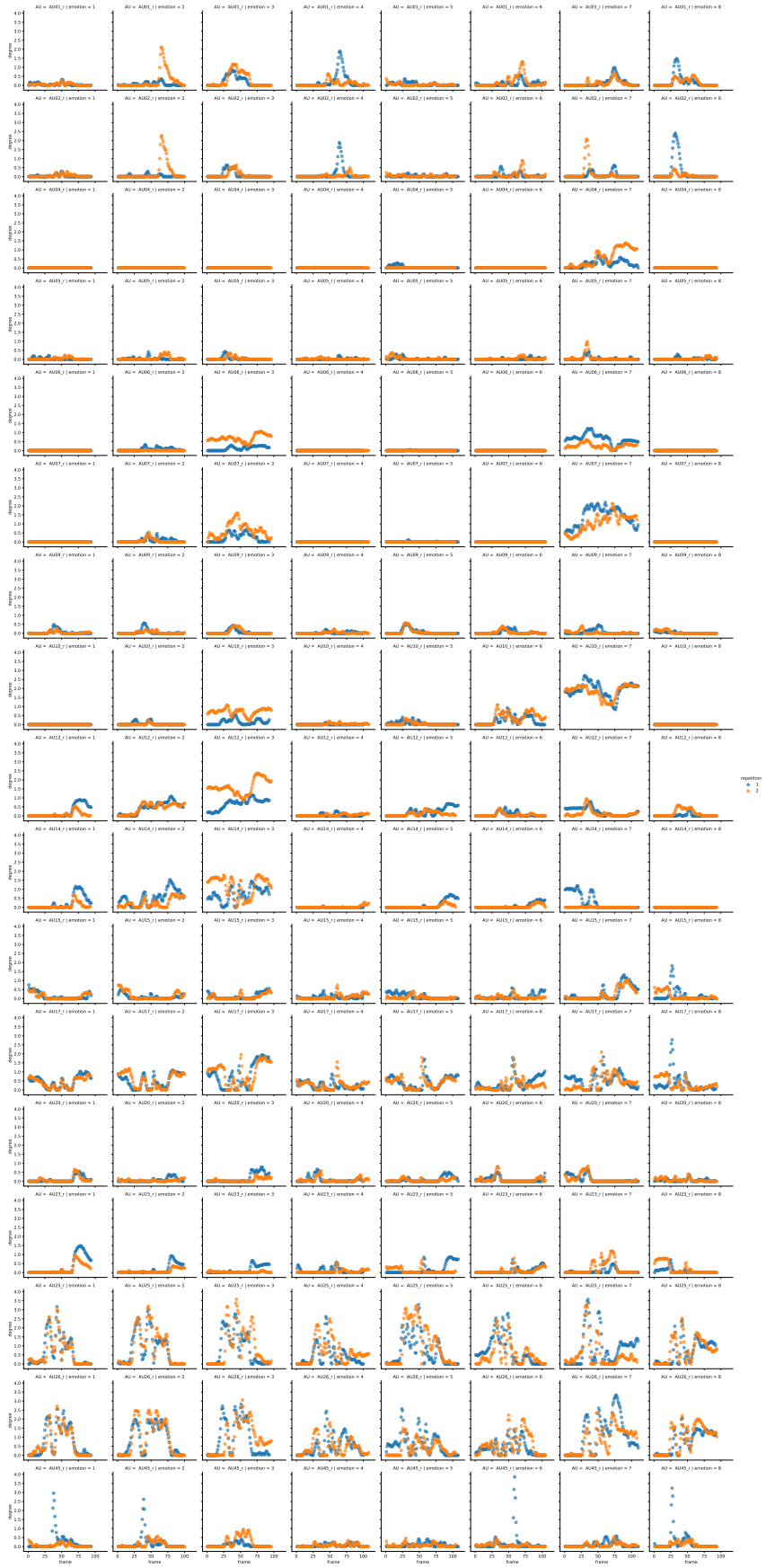


Figure A.30: The evolution curves of AUs detected from the videos of actor11 under different emotions. This actor is considered a good male actor.

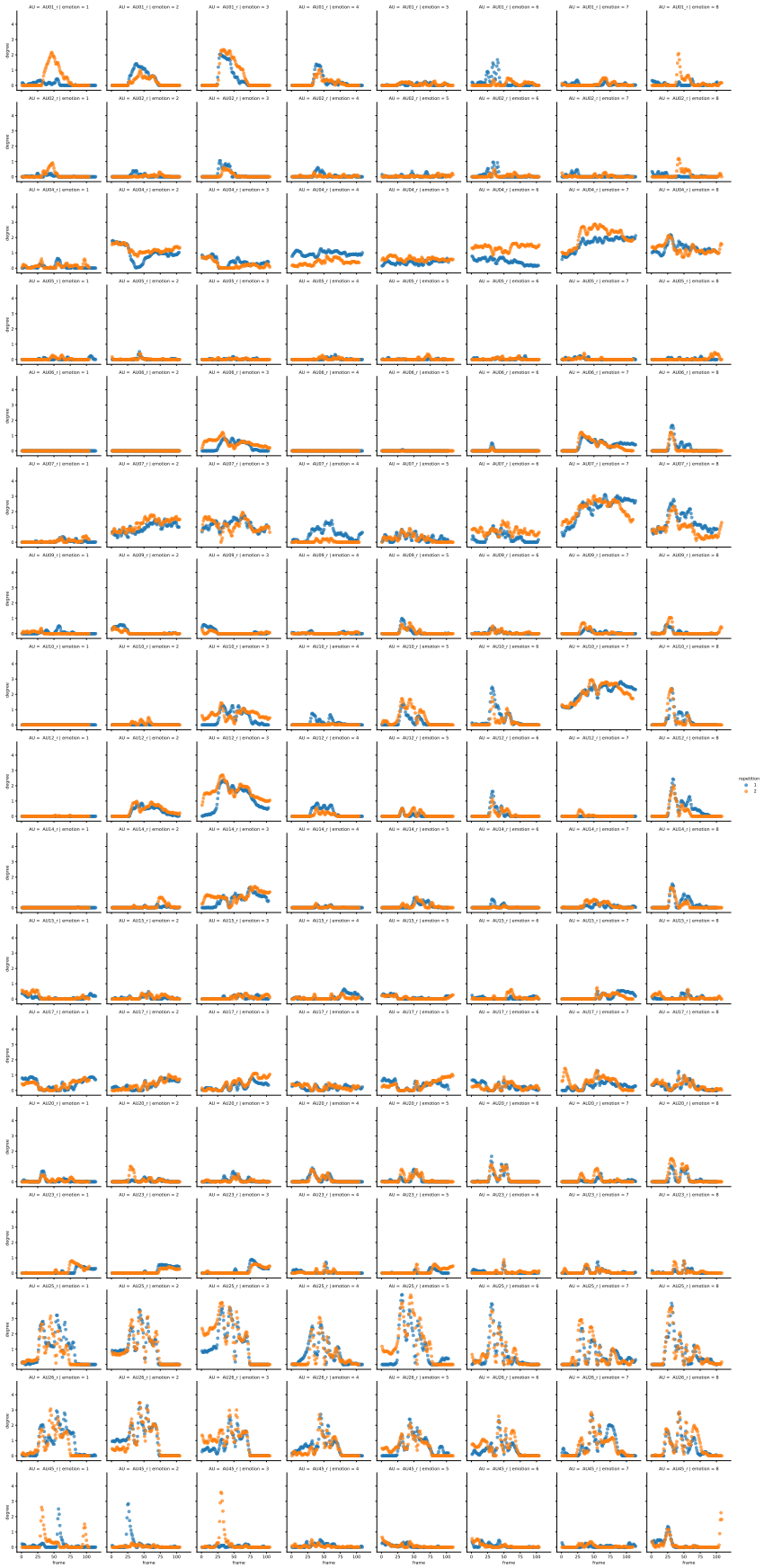


Figure A.31: The evolution curves of AUs detected from the videos of actor17 under different emotions. This actor is considered a good male actor.

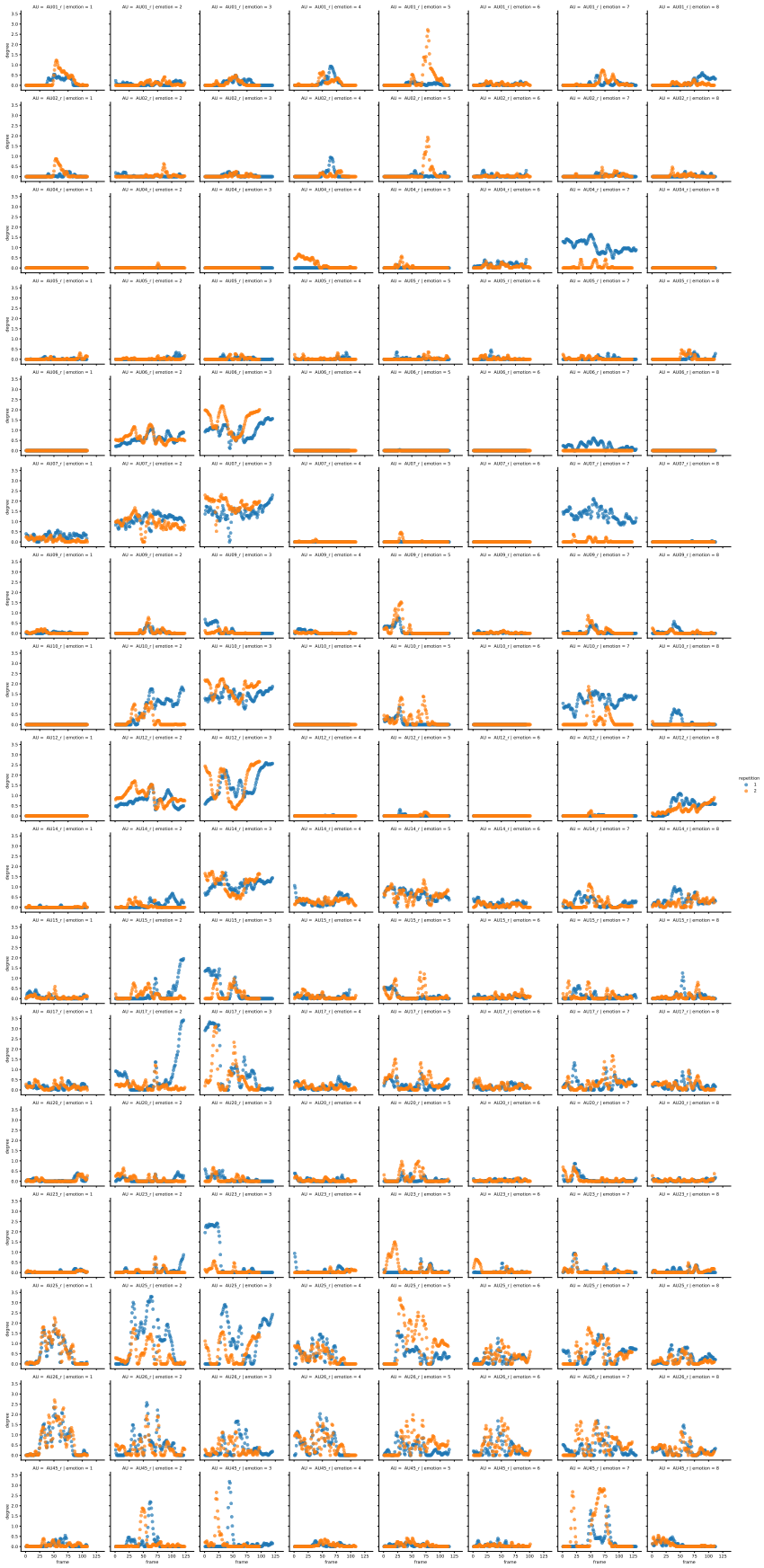


Figure A.32: The evolution curves of AUs detected from the videos of actor5 under different emotions. This actor is considered a bad male actor.

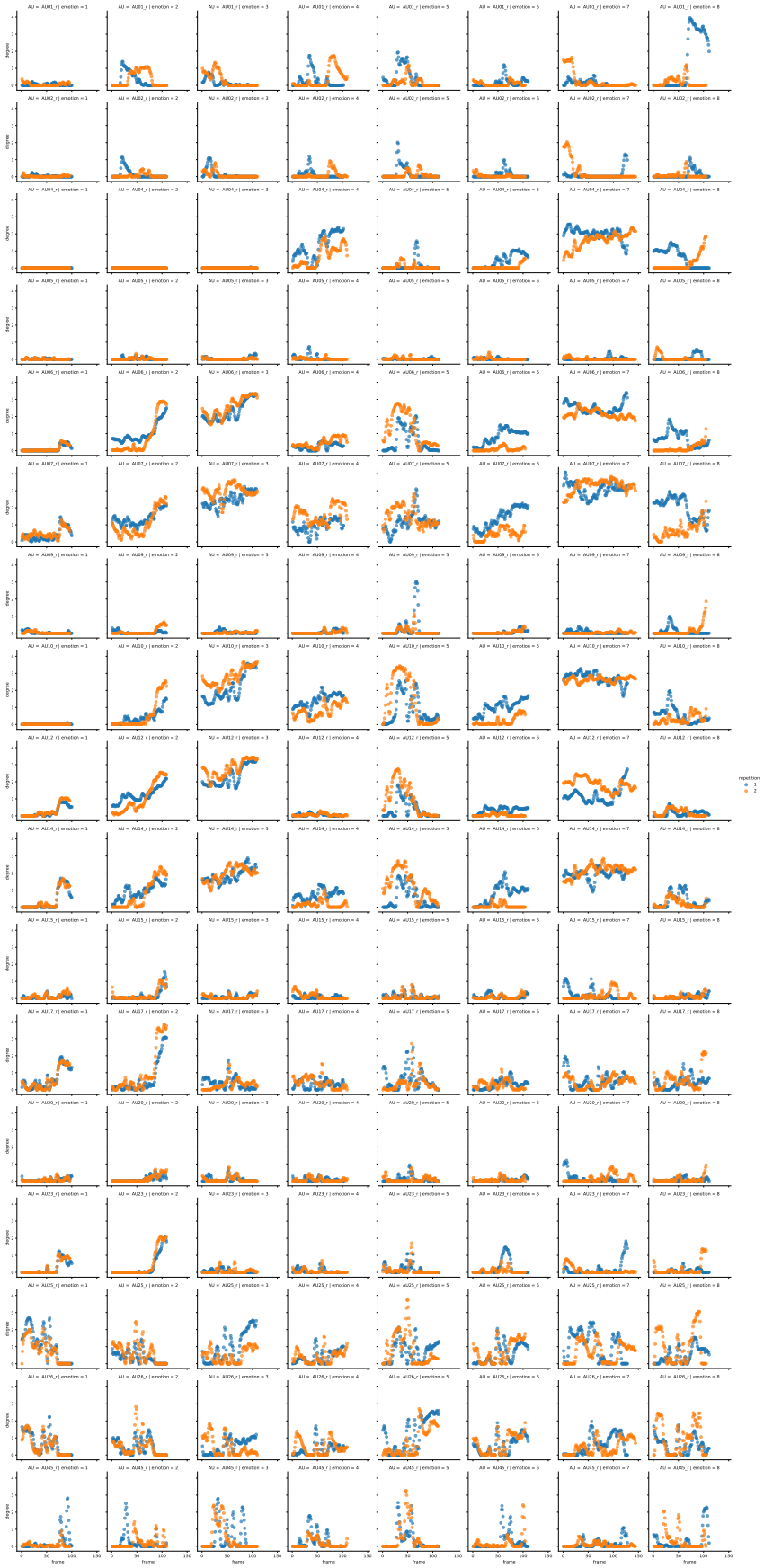


Figure A.33: The evolution curves of AUs detected from the videos of actor19 under different emotions. This actor is considered a bad male actor.

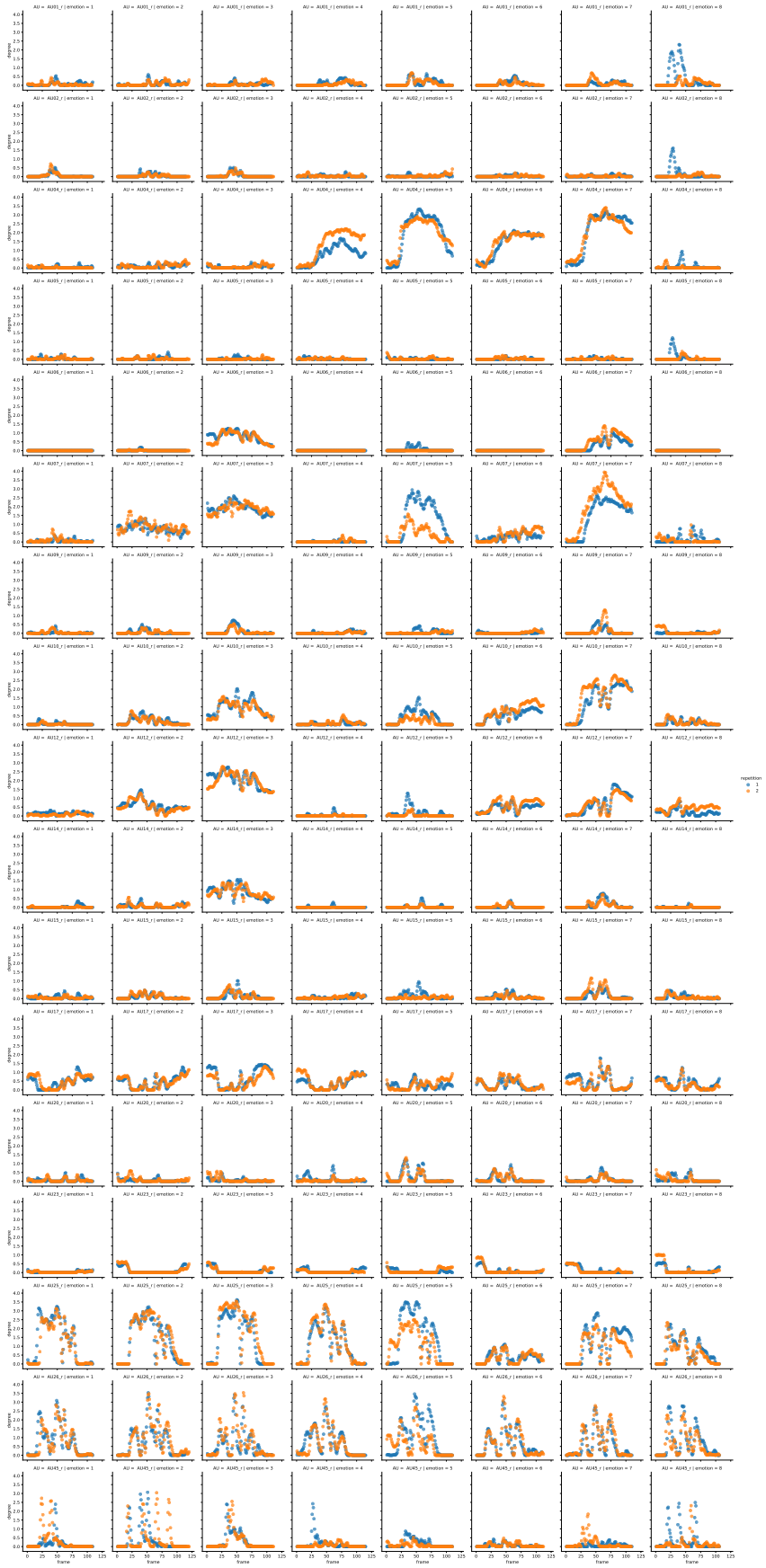


Figure A.34: The evolution curves of AUs detected from the videos of actor2 under different emotions. This actor is considered a good female actor.

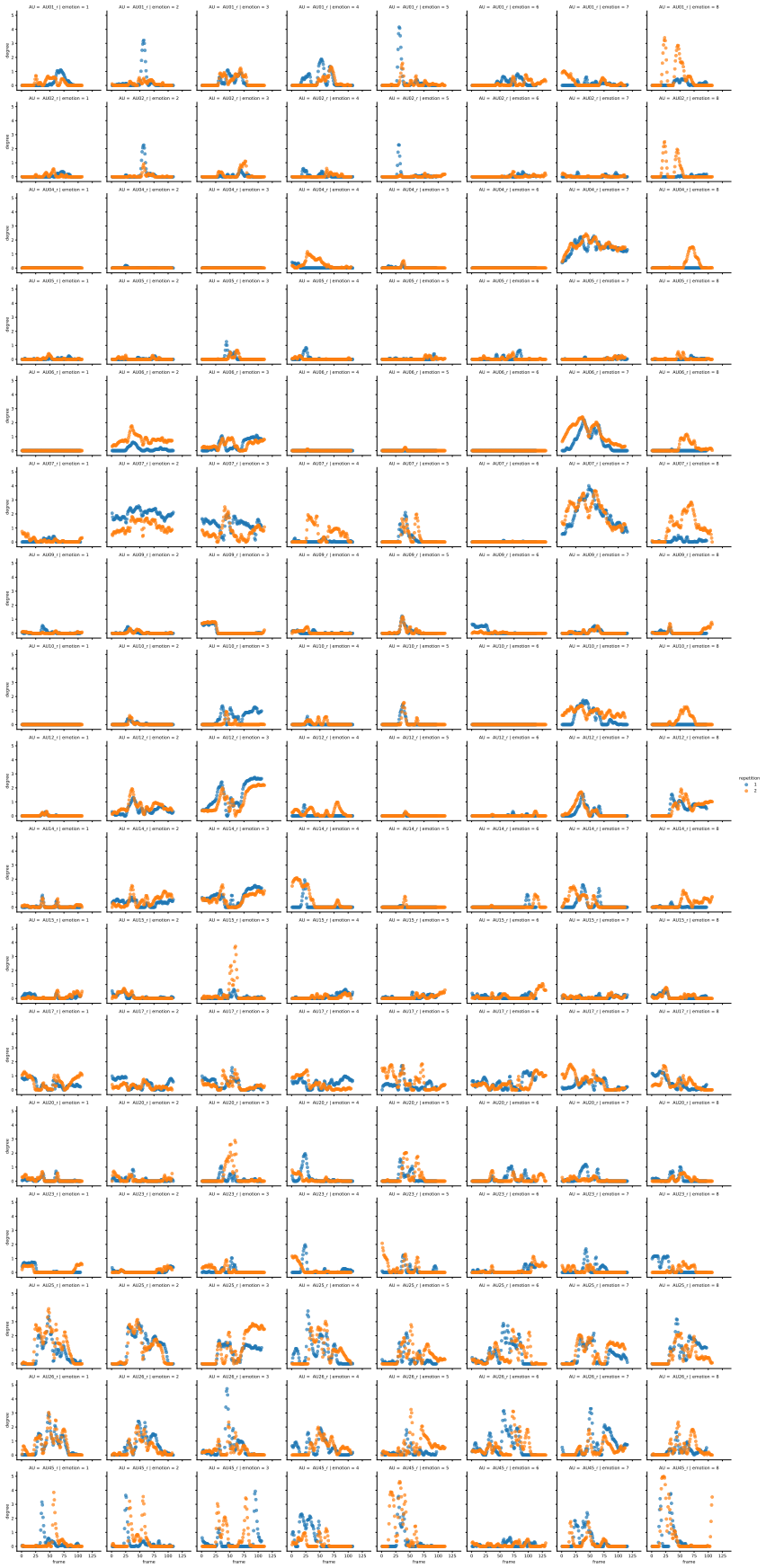


Figure A.35: The evolution curves of AUs detected from the videos of actor20 under different emotions. This actor is considered a good female actor.

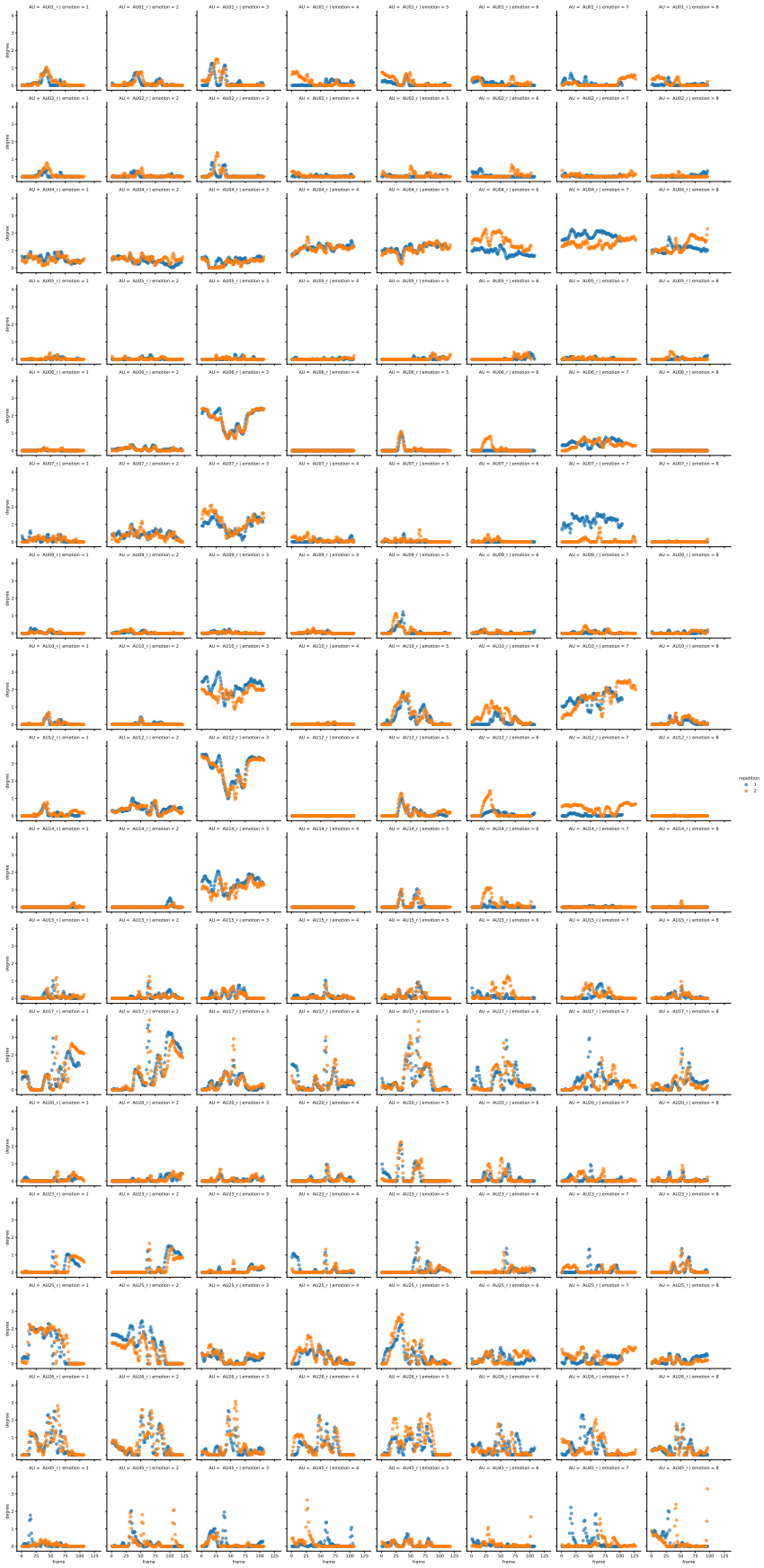


Figure A.36: The evolution curves of AUs detected from the videos of actor6 under different emotions. This actor is considered a bad female actor.

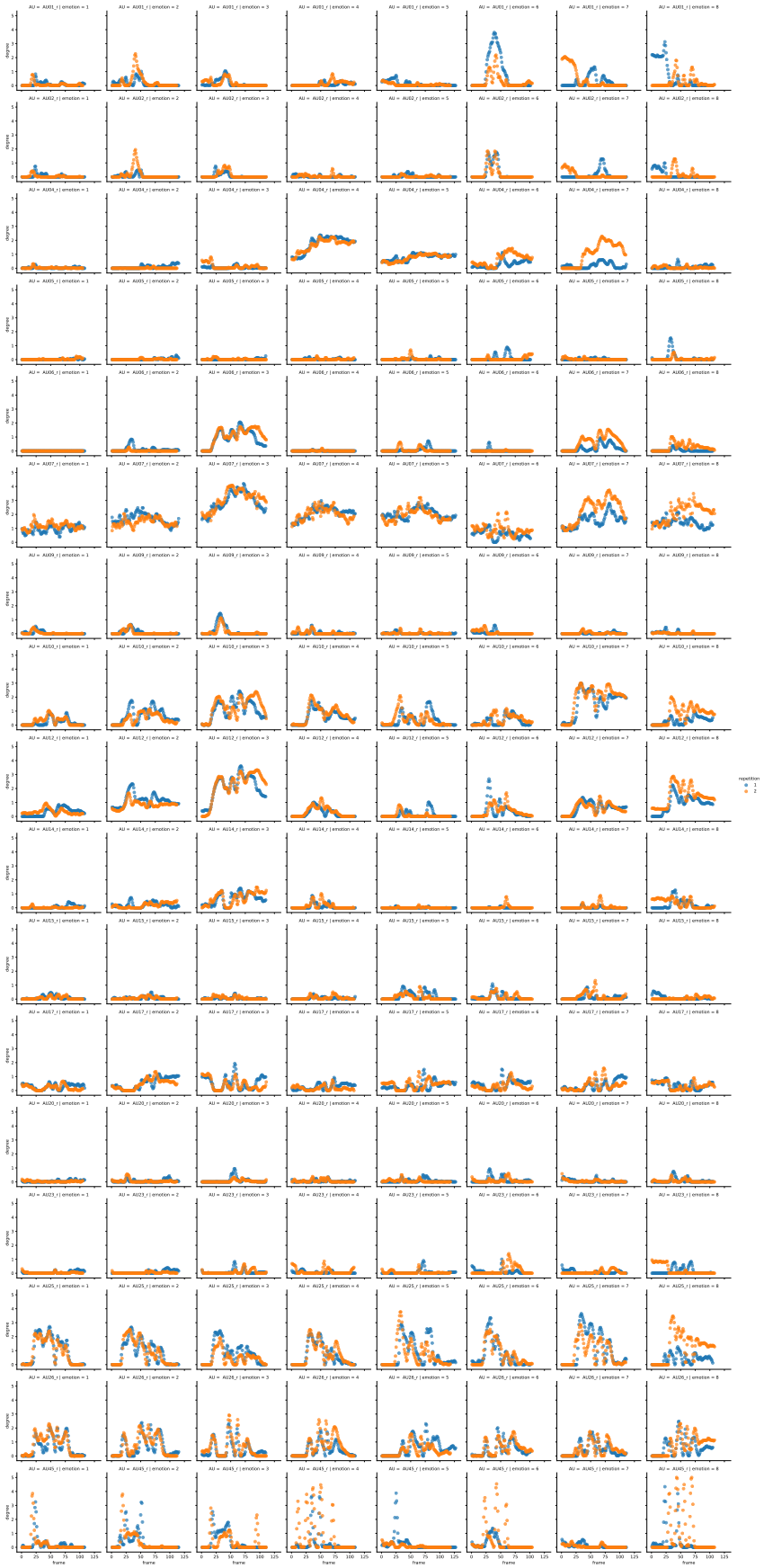


Figure A.37: The evolution curves of AUs detected from the videos of actor12 under different emotions. This actor is considered a bad female actor.