Imperial College London
Department of Computing

Machine Learning for Automatic Analysis of Affective Behaviour

Michael (Mihalis) A. Nicolaou

September 2014

Submitted in part fulfilment of the requirements for the degree of PhD in Computing and the Diploma of Imperial College London. This thesis is entirely my own work, and, except where otherwise indicated, describes my own research.

Abstract

The automated analysis of affect has been gaining rapidly increasing attention by researchers over the past two decades, as it constitutes a fundamental step towards achieving next-generation computing technologies and integrating them into everyday life (e.g. via affect-aware, user-adaptive interfaces, medical imaging, health assessment, ambient intelligence etc.). The work presented in this thesis focuses on several fundamental problems manifesting in the course towards the achievement of reliable, accurate and robust affect sensing systems. In more detail, the motivation behind this work lies in recent developments in the field, namely (i) the creation of large, audiovisual databases for affect analysis in the so-called "Big-Data" era, along with (ii) the need to deploy systems under demanding, real-world conditions. These developments led to the requirement for the analysis of emotion expressions *continuously* in time, instead of merely processing static images, thus unveiling the wide range of temporal dynamics related to human behaviour to researchers. The latter entails another deviation from the traditional line of research in the field: instead of focusing on predicting *posed*, discrete basic emotions (happiness, surprise etc.), it became necessary to focus on *spontaneous*, naturalistic expressions captured under settings more proximal to real-world conditions, utilising more expressive emotion descriptions than a set of discrete labels. To this end, the main motivation of this thesis is to deal with challenges arising from the adoption of continuous dimensional emotion descriptions under naturalistic scenarios, considered to capture a much wider spectrum of expressive variability than basic emotions, and most importantly model emotional states which are commonly expressed by humans in their everyday life. In the first part of this thesis, we attempt to demystify the quite unexplored problem of predicting continuous emotional dimensions. This work is amongst the first to explore the problem of predicting emotion dimensions via multimodal fusion, utilising facial expressions, auditory cues and shoulder gestures. A major contribution of the work presented in this thesis lies in proposing the utilisation of various relationships exhibited by emotion dimensions in order to improve the prediction accuracy of machine learning methods - an idea which has been taken on by other researchers in the field since. In order to experimentally evaluate this, we extend methods such as the Long Short-Term Memory Neural Networks (LSTM), the Relevance Vector Machine (RVM) and Canonical Correlation Analysis (CCA) in order to exploit output relationships in learning. As it is shown, this increases the accuracy of machine learning models applied to this task.

The annotation of continuous dimensional emotions is a tedious task, highly prone to the influence of various types of noise. Performed real-time by several annotators (usually experts), the annotation process can be heavily biased by factors such as subjective interpretations of the emotional states observed, the inherent ambiguity of labels related to human behaviour, the varying reaction lags exhibited by each annotator as well as other factors such as input device noise and annotation errors. In effect, the annotations manifest a strong spatio-temporal annotator-specific bias. Failing to properly deal with annotation

bias and noise leads to an inaccurate ground truth, and therefore to ill-generalisable machine learning models. This deems the proper fusion of multiple annotations, and the inference of a clean, corrected version of the "ground truth" as one of the most significant challenges in the area. A highly important contribution of this thesis lies in the introduction of Dynamic Probabilistic Canonical Correlation Analysis (DPCCA), a method aimed at fusing noisy continuous annotations. By adopting a private-shared space model, we isolate the individual characteristics that are annotator-specific and not shared, while most importantly we model the common, underlying annotation which is shared by annotators (i.e., the derived ground truth). By further learning temporal dynamics and incorporating a time-warping process, we are able to derive a clean version of the ground truth given multiple annotations, eliminating temporal discrepancies and other nuisances.

The integration of the temporal alignment process within the proposed private-shared space model deems DPCCA suitable for the problem of temporally aligning human behaviour; that is, given temporally unsynchronised sequences (e.g., videos of two persons smiling), the goal is to generate the temporally *synchronised* sequences (e.g., the smile apex should co-occur in the videos). Temporal alignment is an important problem for many applications where multiple datasets need to be aligned in time. Furthermore, it is particularly suitable for the analysis of facial expressions, where the activation of facial muscles (Action Units) typically follows a set of predefined temporal phases. A highly challenging scenario is when the observations are perturbed by gross, non-Gaussian noise (e.g., occlusions), as is often the case when analysing data acquired under real-world conditions. To account for non-Gaussian noise, a *robust* variant of Canonical Correlation Analysis (RCCA) for robust fusion and temporal alignment is proposed. The model captures the shared, low-rank subspace of the observations, isolating the gross noise in a sparse noise term. RCCA is amongst the first robust variants of CCA proposed in literature, and as we show in related experiments outperforms other, state-of-the-art methods for related tasks such as the fusion of multiple modalities under gross noise.

Beyond private-shared space models, Component Analysis (CA) is an integral component of most computer vision systems, particularly in terms of reducing the usually high-dimensional input spaces in a meaningful manner pertaining to the task-at-hand (e.g., prediction, clustering). A final, significant contribution of this thesis lies in proposing the *first* unifying framework for *probabilistic* component analysis. The proposed framework covers most well-known CA methods, such as Principal Component Analysis (PCA), Linear Discriminant Analysis (LDA), Locality Preserving Projections (LPP) and Slow Feature Analysis (SFA), providing further theoretical insights into the workings of CA. Moreover, the proposed framework is highly flexible, enabling novel CA methods to be generated by simply manipulating the connectivity of latent variables (i.e. the latent neighbourhood). As shown experimentally, methods derived via the proposed framework outperform other equivalents in several problems related to affect sensing and facial expression analysis, while providing advantages such as reduced complexity and explicit variance modelling.

Acknowledgements

I would like to thank both my supervisors, Prof. Maja Pantic and Dr. Stefanos Zafeiriou for their invaluable help and guidance over these years. I would also like to thank Prof. Vladimir Pavlovic for our fruitful collaboration and all the help that he provided me with. I thank Dr. Hatice Gunes for her help and guidance during the first part of my PhD, as well as all my colleagues for any dialectic encounters, interactions and collaborations. I would like to express my gratitude to my close family and close friends, not just for putting up with me, but also for supporting me. I would like to especially thank my parents, not only for providing me with my unique sequence of deoxyribonucleic acid, but mostly for constantly being an omnipotent core of love and support. This thesis is dedicated to Antonios, Andriani, and to time.

Contents

Chapter 1

Introduction

Contents

The study and understanding of human affect has been a long standing problem, troubling the human race since its infancy. The earliest testimonies on the philosophical enquiries towards the understanding of emotions can be attributed to the Stoics (3rd century BC) [88], who claimed that human affect can be separated into coarse categories such as pleasure, appetite and fear. The Chinese encyclopaedia *Li Chi* (1st century BC), attempts a more detailed discrimination into emotion classes, while also proposing a theory that has dominated the modern psychology of emotions centuries later: that some emotions are *biologically hard-wired* to humans, rather than being acquired through social interactions and learning [217]. Philosophic inquisitions on the understanding of emotions continued throughout the centuries, with pioneering works by Descartes [61] and Spinoza [238], with what Descartes called *passions* being synonymous to the modern definition of emotions. More directly related to emotions is the seminal work of Charles Darwin, who extensively studied expressions of the face and gestures of the body in mammals [55], thus setting the foundations of the study of affect in psychology as well as greatly influencing what we now call *affect sensing*.

A remarkable milestone in the study of affect in psychology, is the work of Paul Ekman and his colleagues, who put forward the claim that there exists a set of six basic emotions (anger, fear, disgust, happiness, sadness and surprise) which are biologically hard-wired to

humans and are common across different cultures, thus rendering them global in terms of both understanding and expressing them. Ekman and his colleagues empirically studied this phenomenon in various works [69, 70, 74], providing the ground for what later evolved as the basis of affective computing. In particular, starting from the mid 1990s, researchers in diverse fields such as computer science, psychology and the cognitive sciences started to take interest in the *analysis* of human affect, be it recognising, interpreting or simulating emotions [198]. This trend has risen out of necessity, since tools generated from the computational analysis of affect can be considered as a requirement for the further evolution of modern scientific fields, such as human-computer interaction, robotics, ambient computing and medicine. The study of affective computing and human behaviour, as it has been defined in the mid 1990s and evolved throughout the past-decades, essentially defines the main topic of this thesis; we propose and develop various techniques, based on machine learning, computer vision and pattern recognition, which particularly fit specific idiosyncratic characteristics of problems commonly dealt with when processing human affect and behaviour, without loss of application generality. In particular, this thesis follows several recent shifts in the field of affective computing [97, 95]: moving away from data acquired in particularly constrained laboratory settings, with actors or other subjects *posing* the emotion expressions (i.e. being told to replicate what they believe to be a specific expression such as anger) to more real-world settings, where the conditions are not so constraint and the emotion expressions by the subjects are naturalistic, usually elicited by conversation or interaction with other subjects. As we thoroughly discuss in what follows, this particular direction entails other radical changes in the problem settings, such as the processing much larger amounts of data (often in the form of videos instead of static images) as well as the adoption of different descriptions of emotions, moving away from the rigid, basic emotion theory initially employed in the field.

The remainder of the introductory chapter is organised as follows. Firstly, in Section 1.1 we refer in more detail to the problem space on which the thesis builds on. Specifically, in Section 1.1.1 we detail the typical structure of affect sensing systems, and subsequently, in Section 1.1.2 we discuss the field shift towards learning continuous dimensional emotion descriptions. Subsequently, in Section 1.2 we analyse a set of significant challenges which have risen in the field, and are specifically tackled in this thesis. Finally, in Section 1.3 we provide a detailed listing of the thesis contributions, providing a summary of methodologies along with the specific application contributions.

1.1 Problem Space

1.1.1 Affect Sensing Systems

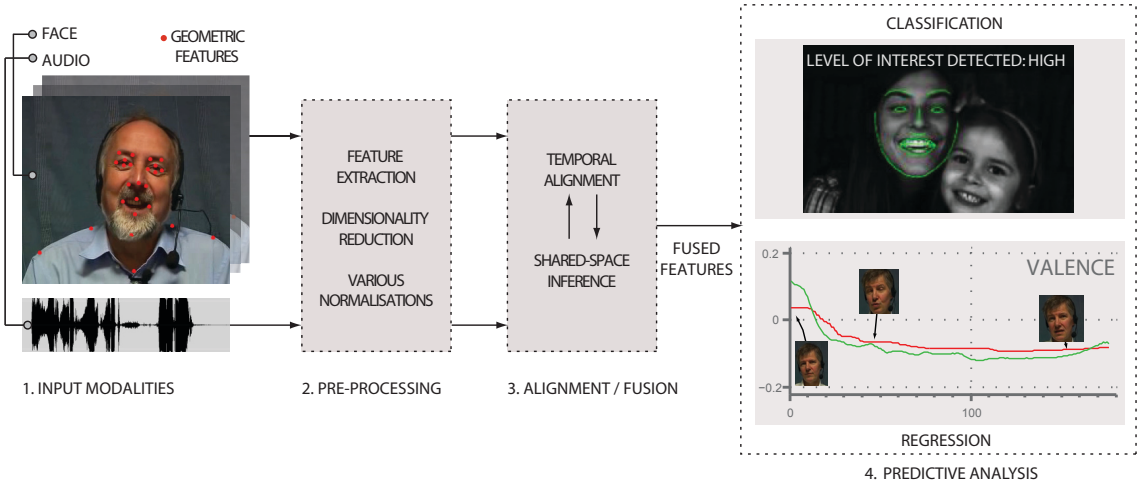


Figure 1.1: Illustration of the commonly utilised pipeline in automatic behaviour analysis and affect sensing. (1) Given a set of observations (features) possibly from multiple modalities, step (2) refers to pre-processing the features to facilitate the task at hand. Furthermore, in step (3), if the observation sets are temporally ordered and not synchronised in time, a temporal alignment process follows, along with the fusion of the features into one set containing all the necessary information pertaining to the task at hand. In the final step (4), predictive analysis takes place, most commonly classification (into discrete classes) or regression (into continuous values).

A typical system aimed towards affect sensing usually follows the pipeline depicted in Fig. 1.1. Firstly, a set of features are obtained depending on the modality utilised (e.g., visual, auditory). In case of e.g., facial images/videos, this can be a collection of coordinates encapsulating the location of various interest points, such as the corners of the eyes, the lips and the eyes. Features derived from such a collection of points are called *geometric features*, while features based on the image pixels are defined as *appearance-based features*. In case of *audio*, this can be prosody features such as pitch or energy, as well as other spectrum-based representations. Secondly, the pre-processing step follows, where the features obtained from each modality are extracted, usually by applying some sort of dimensionality reduction technique, with the goal being to remove the uninteresting components of the input features, such as signals appearing due to noise and corruption, and enhance some characteristics of the signal which can be deemed beneficial for later use (such as e.g., preserving locality and variance). The third stage consists of the actual fusion of the modalities, where the useful information from all utilised cues is inferred - a common way of doing this being by maximally correlating

the modalities. In case there are temporal discrepancies in the data (i.e. the data are not temporally aligned), this stage may also include an alignment step (for an example of aligning human behaviour, please see Fig. 1.5. The final step is typically some form of predictive analysis, be it classifying into discrete labels (e.g., angry, bored), or regressing, i.e. learning continuous values function mappings. As we will see in what follows, regression is usually employed for dealing with the problem of learning continuous emotions.

1.1.2 Continuous and Dimensional Emotion Description

Most of the work in this thesis is driven by the recent trend in affective computing, that is the adoption of a set of latent dimensions which describe the affective state of an individual. Previously, the research community was mostly focusing on the recognition of six discrete basic emotional states [69], happiness, anger, sadness, surprise, fear and disgust. Nevertheless, the deployment of emotion recognition systems under real-world scenarios indicated that a more expressive vocabulary for emotions is required. In fact, research in psychology [129, 138] has hinted that the six basic emotional states correspond only to a small subset of the emotions humans express during their everyday life (see also Fig. 1.2). This lead to the adoption of a different representation for affective states, based on *continuous and dimensional* emotion descriptions. Traced back to the seminal work of Russell in 1980 [216], the most commonly used latent dimensions are Valence and Arousal, with Valence indicating how positive (e.g., happiness, optimism) or negative (e.g., unhappy, depressed) the emotional state is, and Arousal describing how active or passive the emotional state is. This essentially transformed the problem from a classification task to learning *real-valued* functions, i.e. performing regression.

During this paradigm shift in the area of affective computing, another, greater change was taking place in the entire field of data sciences, including machine learning and computer vision. The so-called "Big Data" era led to the gathering of vast amounts of data. In turn, this led researchers to adopt continuous annotations over *time*. That is, instead of annotating static images in terms of discrete emotions, one would annotate audio-visual sequences continuously over the entire duration of the clip in terms of latent dimensions. This led to the creation of databases such as the Sensitive Artificial Listener (SAL) [64] and SEMAINE [157], which were annotated continuously both over time and space. An example of such annotations is shown in Fig. 1.3.

From the machine learning perspective, the presence of multiple continuous emotion dimensions as outputs leads to a regression problem with multiple-outputs. As we will discuss in what follows, this poses a both a set of opportunities for adapting models to the task-at-hand,

(a)

(b)

Figure 1.2: (a) Posed, discrete emotional states (left to right: disgust, happiness, sadness, anger, fear, surprise). Image adapted from [185]. (b) Spontaneous (induced) emotional states. Stills from the SEMAINE database [157].

as well as a set of further challenges to overcome.

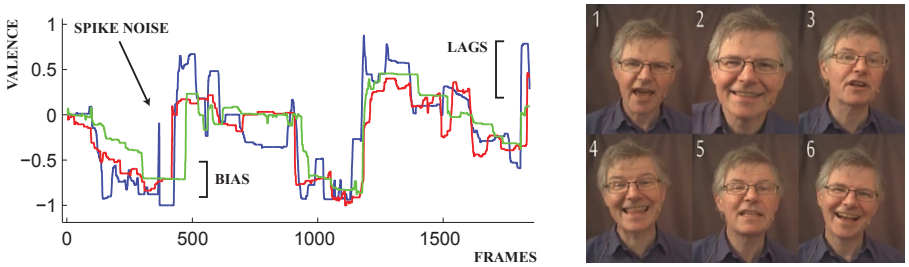


Figure 1.3: Example of multiple valence annotations in the range of [-1,1], with -1 being most negative emotional state and 1 most positive, along with a set of stills from the SEMAINE database. We illustrate a set of challenges arising when dealing with multiple, noisy annotations, as detailed in the text.

1.2 Challenges

In this section, we introduce a set of rising challenges in the field, in order to facilitate later discussions on methodological and application-oriented contributions of our work.

Empirical Analysis of Continuous Emotion Dimensions. The appraisal of emotions utilising latent emotion dimensions is only a recent development in the field of affective computing, and many aspects of the problem can be considered as open problems [95]. Since

Figure 1.4: Temporal phases of Action Unit (AU) activation. From left to right: neutral, onset, apex, offset, neutral. Video from the UvE-Nemo Smile Database.

adopting emotion dimensions such as valence and arousal leads to a vastly different problem setting than the traditional approach of adopting discrete emotion classes, many research questions arise. These questions are of high significance for demystifying several aspects of the problem which in many cases appear to be subjective and ambiguous. A straightforward question can be the correlation of input modalities (e.g., audio or visual cues) to emotion dimensions. This information is essential in order to determine which set of features may be utilised depending on the task at-hand. E.g., as we verify in this work, arousal seems more correlated with audio cues rather than facial expressions and therefore acoustic features can be more suitable for arousal detection. This is actually due to the fact that the frequency and pitch of the voice change accordingly when a person experiences high arousal (e.g., anger, laughter etc.). Secondly, another question which is of interest is the relationship of emotion dimensions to basic emotions. In theory, the values of the latent dimensions which correspond to basic emotions are rather abstract, e.g., happiness corresponds to high valence and high arousal, but no specific value range is defined. Therefore, it is of interest to study how latent dimensions correlate with basic emotions (in effect, the intensity of the presence of these emotions) in order to resolve such ambiguities and provide a better understanding of the problem itself.

Modelling Temporal Dependencies. A recurring challenge in time-series analysis in general, and specifically in behaviour analysis and affect sensing, is the requirement for modelling temporal dynamics. In some settings, such as when analysing the activation of Facial Action Units (AUs)[1], there is a strict sequence of phases which occur in a specific order: neutral, onset, apex, offset and then back to neutral. This is also illustrated in Fig. 1.4, where we visualise the temporal phases of a posed smile activation. Of course, this order of states applies strictly to *posed* expressions, and the situation changes when dealing with *spontaneous* or *elicited* expressions, where the subjects behaviour is more unpredictable, and e.g., an expression

[1] Action Units (AUs) refer to the contraction or relaxation of one or more facial muscles, according to the Facial Action Coding System (FACS) [73]. We discuss AUs more in Chapter 2.

might be interrupted by, a re-activation or the onset of a different expression. Furthermore, in case we are predicting emotions over time, the outputs also exhibit some form of temporal smoothness which needs to be modelled. E.g. in a high valence episode (e.g., corresponding to laughter), temporal phases analogous to the onset and offset of AUs will manifest, the former when the valence value is increased and the latter when it is decreased. In general, taking dynamics into account is crucial for the interpretation of complex, human behaviour as e.g., in many cases the behaviour can be highly ambiguous. An example is a nervous laughter episode during an anger outburst; only via temporal modelling a system can avoid detecting the laughter episode as an example of joy.

Exploiting Emotion Dimension Correlations for Learning. In many learning problems, the setting consists of multidimensional labels (or targets) to be learnt. The problem of dimensional emotion recognition inherently belongs in this class; the latent dimensions which describe the affective state of an individual are multiple, and evidence from psychology hints that the emotion dimensions can be highly correlated [129, 181, 5, 138]. In effect, that means that there is a covariance structure in the multiple dimensions. Since many researchers have adopted the continuous and dimensional emotion descriptions in learning, a research question that naturally arises is whether one can evaluate the correlations which actually arise within emotion dimensions, and actually exploit them for learning. This translates to developing methods which can (i) learn e.g., commonly occurring patterns (over time) between dimensions such as valence and arousal, and (ii) remove the redundancy which is exhibited in the output dimensions in order to construct more parsimonious models. This direction was virtually unexplored in the field of affective computing before the work we present in this thesis.

Fusion of Multiple Continuous Annotations. The fusion of multiple, continuous annotations is arguably the most significant problem which arises when utilising continuous dimensional annotations. While most supervised learning tasks assume the existence of reliable and objective labels, this is very often not the case, especially when dealing with problems related to human behaviour and affect. In particular, the annotation process in such settings can be highly error prone, leading to inaccurate, ambiguous and subjective labels, which in turn are utilised to train ill-generalisable models. Such issues arise both (i) due to the nature of the problem, where many affect-related labels are defined rather ambiguously thus leading to the adoption of personal interpretations, and (ii) due to the impact of human factors, such as the varying perception of emotions and the personal characteristics and experiences of the annotator. The issue becomes even more prominent when the task is temporal, as (i) it renders

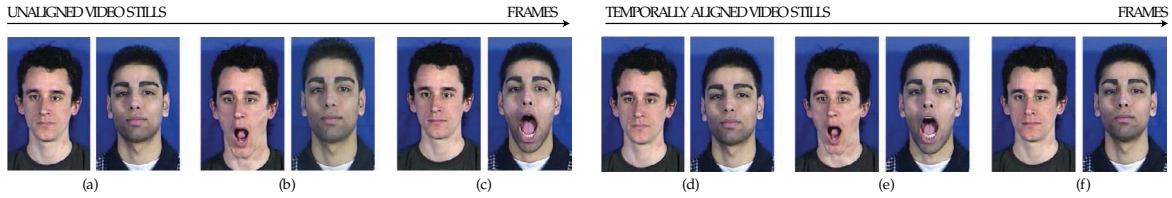


Figure 1.5: The problem of aligning human behaviour from videos. In unaligned videos (a,b,c), the temporal phases of AU 26 (mouth opening) are not synchronised accross the subjects. In the aligned videos (d,e,f) the temporal phases in both subjects are aligned in time, i.e. with (d) being neutral, (e) being apex and (f) back to neutral.

the labelling procedure vulnerable to *varying* temporal lags caused by the varying response times of annotators (depending on factors such as fatigue and stress), while (ii) a delay in most annotators is expected to appear due to the real-time nature of the annotation acting together with the temporal delay exhibited by the annotator when perceiving an emotion and acting towards labelling it. In effect, the annotation signals carry a strong spatio-temporal, annotator specific bias, while also being exposed to other issues such as e.g., noise generated via the input devices used for annotating. These difficulties give rise to various issues in the annotations, such as scale-ambiguities, temporal lags, spike noise and others (see Fig. 1.3, where a set of example annotations from the SEMAINE database are illustrated). In such scenarios, the only information which can be exploited in order to derive a clean, correct version of the ground truth is the existence of multiple annotations. In fact, in such difficult scenarios, multiple experts (usually trained in psychology) are employed as annotators, with the idea being that somehow the "average" annotation will provide the most reliable labels, which will later be used for training machine learning methods for predictive analysis. In fact, the typically employed approach in the field is the most naive, that is, simple averaging. Nevertheless, simple averaging can be deemed suboptimal for such problems for many reasons, such as the lack of a mechanism to rank the annotators and weight the annotations, in effect assessing the confidence level attributed to each annotator. This is a reasonable task, as we expect that some annotators will be more competent than others. Furthermore, simple averaging inherently lacks the ability to compensate for temporal discrepancies amongst the annotators, leading to the manifestation of false peaks in the resulting signal. As can be easily understood, the absence of well defined labels, free from noise and annotator bias deems the learning problem even more difficult and even, in some cases, ill-defined.

Temporal Alignment of Human Behaviour. The manifestation of similar behaviour in multiple sequences is the usual manner in which data are gathered and used in order to

train machine learning models. For example, a set of videos capturing the activation of the same Action Units (AUs) from multiple subjects can be used as observations, leading to the training of a machine learning model which detects the occurrence of the particular set of AUs. The same applies when e.g., training models to detect the occurrence of smiles or the temporal phases of the activation. Nevertheless, a common problem in these scenarios is that although the manifested behaviour is similar *over time* (e.g., both subjects are smiling), this behaviour is not temporally aligned. For example, the peak of the smile in subject one happens at frame t_1, while the peak of the smile in subject two happens at frame t_2, with $t_1 \neq t_2$ (see also, Fig. 1.5). The problem is deemed very challenging due to numerous reasons, such as possible large temporal discrepancies, inter/intra subject variability, as well as the presence of various forms of noise. The most basic of algorithms for solving the temporal alignment problem, Dynamic Time Warping (DTW), is optimal for aligning one-dimensional, clean, temporal signals. Nevertheless, the most common case when dealing with real data lies in the availability of multi-dimensional signals, possibly of different dimensionality. It is natural that some form of dimensionality reduction is utilised to accommodate time-warping in a more robust (to outliers, occlusions and noise) scenario. We discuss more details regarding Time Warping and related work in Chapter 3.

Fusion of Multiple Modalities. It is very common for problems in learning and vision for observations extracted from multiple modalities to be available. An open question is how to optimally fuse the modalities at hand in order to maintain only what can be considered as useful information for a specific task. This can be performed both in an unsupervised manner, when the fused observations are extracted without considering labels but subject to some constraints, or in a supervised manner where the optimisation function includes some kind of penalty when labels are not predicted correctly. A fusion model should be able to isolate corruptions in the data, commonly arising in realistic scenarios, even when the corruptions are not spread evenly across modalities, e.g. in some cases the visual signal might be very noisy due to illuminations or occlusions, while the audio signal may be noise free. A limit case of the problem is when one of the modalities is entirely missing in the test data queries; the model should be able to extrapolate given the training on both modalities and be able to perform inference to determine e.g., the correct label for the test query.

Dimensionality Reduction and Feature Extraction. A typical pre-processing step in most learning and vision applications refers to dimensionality reduction, usually performed via component analysis methods such as Principal Component Analysis (PCA). (See Fig. 1.1). Typically unsupervised, dimensionality reduction methods aim to reduce the number of ran-

dom variables in the given data by projecting them on a latent space which satisfies a set of constraints depending on the problem. For example, PCA transforms the given data into a reduced dimensionality space which preserves most of the data variance, thus minimising the reconstruction error. Linear Discriminant Analysis (LDA) optimally reduces the dimensionality by also considering class labels, while Locality Preserving Projections (LPP) do the same while preserving a notion of locality, usually encoded via a graph. Dimensionality reduction techniques should be flexible enough to accommodate varying problems, while complexity is another factor that should be taken into account, since as a typical pre-processing step the observations will consist of both high-dimensional as well as a large number of samples. While dimensionality reduction via component analysis has been well studied over the past decades, the focus of the research is mostly relating to *deterministic* component analysis. The formulation of novel, probabilistic component analysis models can be very beneficial to many fields, due to advantages such as uncertainty estimation as well as reduced complexity in most cases.

1.3 Contributions

In this section, we list the contributions of our thesis both technically, as well as with respect to the aforementioned problems. The first part of this thesis deals mostly with the empirical analysis of the relatively unexplored problem of continuous dimensional emotion recognition, focusing mostly on learning to predict emotion dimensions via exploiting the correlations exhibited by the emotion dimensions. The contributions arising from this part are mostly driven from the affective computing viewpoint, focusing on the specific application and psychological theory, and ultimately deriving appropriate models to tackle the problem of learning emotion dimensions. The second part of the thesis is more technically oriented, focusing on proposing novel component analysis methods, which are again fitted to specific very crucial problems, such as the fusion of multiple continuous annotations as well as dimensionality reduction. It is important to note that, while at most times the proposed techniques have been developed with a specific application in mind, they remain generally applicable to any problem with similar settings, with possible applications including medical imaging, health assessment, recommender systems, affect-aware and adaptive user interfaces and robotics.

The rest of the section is organised as follows. Firstly, in Section 1.3.1, we introduce the various methodologies presented in this thesis while discussing particular technical novelties. Secondly, in Section 1.3.2, we discuss particular contributions of the aforementioned methodologies with respect to the challenges and problems discussed in Section 1.2.

1.3.1 Proposed Methodologies

In what follows, we summarise the methodologies introduced in this thesis in order to facilitate the following discussion on solving particular challenges and problems via these methods.

Part I: Learning Emotion Dimensions

The first part of this thesis deals particularly with learning emotion dimensions and various challenges met when dealing with predictive analysis in terms of such emotion descriptions. Three methodologies are proposed, which are based on neural networks, the Relevance Vector Machine (RVM) and Canonical Correlation Analysis (CCA), all aiming at exploiting spatio-temporal correlations that manifest in the outputs of a given problem (in this case, in emotion dimensions). The application of these models is without loss of any generality, since they cover a very wide problem class, and are suitable for application within any similar scenario, i.e. where the targets (or outputs) consist of multi-dimensional vectors which are likely to be correlated in time and space. The proposed methods are summarised in what follows.

- **Chapter 5. BLSTM-NN Output-Associative Fusion.** A precursor of a current trend in machine learning, the so-called "deep-learning" methods, the Bidirectional Long Short-Term Memory Neural Networks (BLSTM-NN) are one of the most recent variations of traditional recurrent neural networks. BLSTM-NNs, introduced in [107], are able to model long-term temporal dependencies in observations by modifying the structure of each node in a typical neural network in order to resolve the *vanishing gradient problem*, which led to various issues due to the gradient either vanishing or growing exponentially during learning. We discuss more regarding LSTM in Chapter 3, while in Chapter 5, we utilise BLSTM-NN for (i) fusion of multiple modalities, and (ii) output-associative fusion, that is, learning temporal patterns arising in outputs, not only in inputs.

- **Chapter 6. Output-Associative Relevance Vector Machine (OA-RVM).** The Relevance Vector Machine (RVM) is a formulation of sparse probabilistic regression, introduced by Tipping in [246] and later extended in [249]. Closely related to Gaussian Processes (GP), the RVM provides a fast and sparse alternative to traditional GP learning, at the cost of some unintuitive properties regarding the estimation of uncertainty. Nevertheless, RVMs constitute one of the fastest and sparse Bayesian regression techniques in machine learning, providing both accuracy and robustness to noise. In Chapter 6, we extend RVM by augmenting the design matrix in order to learn correlations which

manifest in multi-dimensional outputs over time. In more detail, we model the correlation of output dimensions over time, and incorporate such representative basis in the model, in order to utilise temporal output patterns in learning. We coin this model the Output-Associative Relevance Vector Machine (OA-RVM).

- **Chapter 7. Correlated-Spaces Regression.** Canonical Correlation Analysis (CCA) is a fundamental component analysis technique which, given two sets of observations discovers a set of loading matrices which project the observation sets onto a latent space where these sets are maximally correlated. Correlated-Spaces Regression (CSR) is a technique we propose in Chapter 7, which is based on CCA. The main idea in CSR is that instead of correlating input observation sets, we correlate inputs and outputs. This simple idea entails several advantages, particularly to the problem of learning continuous dimensional emotion descriptions as we will discuss in this section since in effect, CSR allows us to simultaneously correlate inputs with outputs and reduce output redundancies.

Part 2: Component Analysis for Affective Behaviour

The second part of the thesis deals with the design of novel Component Analysis (CA) methods[2]. In particular, we firstly develop methods belonging to the general category of Shared-Space component analysis. These methods aim to discover a "shared-space" amongst multiple sets of observations, while satisfying particular constraints. Such models are particularly suited for the problems of fusion and temporal alignment, as both problems can benefit from the discovery of a common space of all observations under some constraints (e.g., the derived shared space can act as the fused features). In what follows, we summarise the two novel shared-space CA methodologies introduced in this thesis: Dynamic Probabilistic Canonical Correlation Analysis (DPCCA) and Robust Canonical Correlation Analysis (RCCA).

- **Chapter 9. Dynamic Probabilistic Canonical Correlation Analysis (DPCCA).** In Chapter 9, we propose a novel, dynamic probabilistic model based on a private-shared space formulation. The private-shared space formulation entails that given a set of multiple observations, DPCCA recovers both the common characteristics of the sequence at hand (in the shared space), while isolating portions of the signal which are specific to each sequence (in the private space). Furthermore, by imposing Markovian dependencies on the latent variables, DPCCA is able to model the temporal characteristics of

[2] A detailed introduction to CA is presented in Chapter 3.

the observations. This is, to the best of our knowledge, the first private-shared space technique in the field which models temporal dependencies. Furthermore, DPCCA is augmented with a time-warping process, leading to the DPCCA with Time Warping model (DPCTW) model. Essentially, this is performed by attaching a time-warping processes on the "clean" spaces of each observation sequence, i.e. by removing private, non-shared characteristics and noise, thus enabling the alignment of noisy sequences which contain a commonality set. In effect, this provides an elegant method for the temporal alignment of multiple sequences in a clean, shared space. Summarising, given a set of multiple observations, DPCCA is able to (i) isolate private characteristics belonging to each set observations, (ii) learn the commonality which underlies *all* observations, (iii) model temporal dynamics via Markovian dependencies, and (iv) align the observations in time via a time-warping process.

- **Chapter 10. Robust Canonical Correlation Analysis (RCCA).** Canonical Correlation Analysis (CCA) is a traditional method, commonly utilised in multiple diverse scientific fields. Nevertheless, the Gaussian noise assumption accompanying the original formulation of CCA limits the use under real-world scenarios where gross noise and corruptions are observed. In Chapter 10, we propose the Robust Canonical Correlation Analysis (RCCA) which can better deal with the problem of learning from high-dimensional, grossly corrupted data. This is accomplished by robustly estimating a low-rank latent subspace even in the presence of gross noise, by decomposing the observation sets into a low-rank component and a sparse noise component. Similarly to DPCCA, a time warping process can be integrated into RCCA, in order to align the corrupted sequences in the derived error-free low-rank subspace. Summarising, RCCA (i) jointly decontaminates observation sets which have been perturbed by sparse, gross noise, (ii) models the clear, noise-free shared space of the observations, and (iii) allows for the temporal alignment of high-dimensional, grossly corrupted input sequences, all while providing a framework which can be used for e.g., robust fusion and robust multi-modal fusion.

The final method presented in this thesis deals with probabilistic feature extraction via component analysis. This is in fact a more technical work, which nevertheless entails a set of significant advantages when utilised in various applications. In what follows, we summarise the *first* unifying probabilistic component analysis framework in literature thus far.

- **Chapter 11. A Unified Framework for Probabilistic Component Analysis.** Feature extraction and dimensionality reduction is a crucial pre-processing step in the vast majority of machine learning, application-oriented systems, with the most typically used method being Principal Component Analysis (PCA). Although deterministic CA has been well studied in literature so far and several unification frameworks have been introduced [58, 123, 52, 241], to this date no probabilistic unification framework has been proposed for component analysis. This is of crucial importance, not only because unified frameworks offer various insights into the workings of the methods at hand, but also since probabilistic formulations (e.g., based on Expectation Maximisation (EM)) usually pose several advantages, such as lower per iteration complexity as well as probabilistic inference facilitating explicit noise and uncertainty estimation. In Chapter 11, we propose a unified framework which covers all component analysis methods whose corresponding deterministic formulation can be posed as a trace optimisation problem without domain constraints for the parameters. Besides PCA, we formulate other, commonly used methods including Linear Discriminant Analysis (LDA) and Locality Preserving Projections (LPP), while other more recent approaches are also incorporated, such as Slow Feature Analysis (SFA). The contributions derived from this framework, beyond the theoretical insights on component analysis, are as follows: (i) probabilistic models for certain CA methods are proposed for the first time, such as probabilistic LPP, (ii) explicit per dimension variance modelling, (iii) reduced per-iteration complexity in comparison to deterministic equivalents, as well as (iv) a flexible framework upon which novel component analysis methods can be straightforwardly generated.

1.3.2 Application-oriented Contributions

We have so far described both the challenges that this thesis deals with (Section 1.2), as well as the methodologies introduced in this thesis in order to tackle such challenges (Section 1.3.1). In what follows, we proceed to discuss several application-oriented contributions, inspired mostly by the aforementioned challenges, typically encountered when deploying affect analysis systems under real-world conditions.

Learning Continuous Emotion Dimensions via Exploiting Output Correlations (Chapters 5, 6, 7)

The work presented in this thesis shows the *first* empirical results on large audio-visual corpora which experimentally verify that emotion dimensions manifest various correlations which prove beneficial for the task of learning continuous and dimensional emotional states. Furthermore,

novel models fitted to the task are presented. In more detail, such output correlations are exploited in our work by (i) learning output-dependencies during training in order to improve accuracy on testing data, and (ii) removing redundancy in the output dimensions. In more detail, regarding (i), in Chapter 5 we approach the problem in a straightforward manner; we show how by appropriately utilising the so-called Bi-directional Long Short-Term Memory Neural Networks (BLSTM-NN), we are able to exploit both temporal information as well as temporal correlations. This is, to the best of our knowledge, the first work[3] which explicitly proposes and models the relationship of output dimensions for the task of learning continuous emotion dimensions, and has influenced several works by other researchers, such as [13]. This concept is further studied in Chapter 6, where we propose an extension of the Relevance Vector Machine (RVM) which is able to learn output-dependencies over time. In more detail, we propose a novel Output-Associative Relevance Vector Machine (OA-RVM) regression framework that augments the traditional RVM regression by being able to learn *non-linear input and output dependencies*. Instead of depending solely on the input patterns, OA-RVM models output covariances within a predefined temporal window, thus capturing past, current and future context. As a result, output patterns manifested in the training data are captured within a formal probabilistic framework, and subsequently used during inference. We successful apply our model to the problem of dimensional continuous prediction of emotions, and evaluate the proposed framework by focusing on the case of multiple nonverbal cues, namely facial expressions, shoulder movements and audio cues. We demonstrate the advantages of the proposed OA-RVM regression by performing subject-independent evaluation using the SAL database that constitutes of naturalistic conversational interactions. The experimental results show that OA-RVM regression outperforms the traditional RVM and SVM regression approaches in terms of accuracy of the prediction (evaluated using the Root Mean Squared Error) and structure of the prediction (evaluated using the correlation coefficient), generating more accurate and robust prediction models. Regarding (ii), in Chapter 7, we present a simple methodology based on the least-squares formulation of Canonical Correlation Analysis (CCA), which is able to project features extracted from multiple modalities as well as output emotion dimensions onto a common space, where their inter-correlation is maximised[4]. In effect, this entails that (a) the observations become correlated to the output-dimensions, significantly reducing their dimensionality, while (b) removing the output redundancy by projecting the emotion dimensions on a diagonal covariance latent space. As we show in Chapter 7, this is highly beneficial

[3] At time of publication [174].

[4] Note that since CCA is a static method, the output modelling in CSR is spatial and not temporal (in contrast to OA-RVM, where spatio-temporal modelling is achieved via a temporal window and BLSTM-NN where previous inputs and outputs are recurrently fed into the model). CSR can be easily extended to accommodate for temporal relationships by utilising temporal windows, as in case of OA-RVM.

in terms of accuracy. Finally, it is important to state that in Chapter 7, we also contribute in terms of empirically analysing emotion dimensions, answering questions such as (a) which modality best correlates with particular emotion dimensions in comparison to other emotion dimensions, and (b) analysing the correlation of emotion dimensions to (the intensity of) basic emotions. As mentioned in Section 1.2, (a) is of high importance when designing systems aiming for particular emotion dimensions (such as e.g., arousal), as one can utilise the cues which best correlate with the target dimension (E.g., audio cues for arousal). Furthermore, as we show in our work, it turns out that emotion dimensions are better correlated to other emotion dimensions than to feature sets (E.g., valence is better correlated to arousal, power, intensity and expectation rather to facial expressions or audio cues). In turn, these findings further motivate our work on learning output-associations amongst emotion dimensions.

Level of Interest as a Continuous Emotion Dimension (Chapters 7, 10 and 11)

The modelling of the level of interest constitutes a problem with very large applicability. The demand for the detection of interest under real-world conditions (e.g., in museums) has led to great attention from researchers in affective computing and machine learning [195, 227, 228]. In most related work, interest is not considered as an emotion dimension, but is usually studied similarly to basic emotions, i.e. as a discrete label. In this thesis, we attempt to treat Interest as an emotion dimension. That is, firstly in Chapter 7, we define Interest as an emotion dimension. In a normalised range $[-1, 1]$, we define the dimension of interest as ranging from disinterested (-1) to interested (1) while we gather the relevant annotations of the Level of Interest by eight annotators. Subsequently, we study the correlation of interest to emotion dimensions. In conclusion, we find that the continuous annotations of interest are well correlated with emotion dimension annotations, despite the disjoint set of annotators used in different sets of annotations. In agreement to findings in psychology, we find that the level of interest best correlates with arousal and secondly with valence. Furthermore, in Chapter 10, we provide a novel robust feature fusion technique, Robust Canonical Correlation Analysis (RCCA), which we apply to the problem of audio-visual interest prediction. As we show in relevant results, RCCA is able to outperform other feature techniques for this task. Furthermore, by utilising other emotion dimensions in the comparison, we find that although there is an overlap, the Interest measurements contain unique information with respect to other emotion dimensions. Finally, in Chapter 11, we utilised a set of quantised interest annotations and evaluate the Probabilistic Linear Discriminant Analysis (EM-LDA) we propose in the same chapter on the problem of feature extraction for the detection of interest.

Spatio-temporal Fusion of Multiple Annotations (Chapter 9)

It is a common scenario for applications which pertain to subjective labels to attain multiple annotators in an attempt to reduce the subjectivity, and person-specific bias of the annotations. As discussed in Section 1.2, this is a matter of crucial importance since the "ground truth" derived from the multiple annotations is subsequently used in order to train machine learning models to predict continuous emotions. Therefore, if the ground truth is not obtained correctly, it is unavoidable that the relevant learning techniques employed will be unable to model the true latent functions which map e.g., facial expressions to continuous emotions, but rather will be negatively influenced by both annotator bias and fallacies of human judgement, noise and other temporal discrepancies. Understandably, these issues establish the problem of fusing multiple continuous annotations as perhaps the most significant challenge of adopting continuous emotion annotations. Furthermore, it is important to note that most researchers simply average the annotations in order to obtain what will be considered as the ground truth, a quite suboptimal approach to the problem as it renders the annotations susceptible to various types of noise.

An attempt to solve this problem is presented in Chapter 9 where a novel probabilistic method is presented for inferring the ground truth based on a set of imperfect annotations. In more detail, we present a novel dynamic private-shared space probabilistic model based on Canonical Correlation Analysis (CCA), which we dub Dynamic Probabilistic Canonical Correlation Analysis (DPCCA). This approach offers a complete solution to the problem of fusing multiple annotations, as it is suitable for tackling the most significant of problems which commonly arise in such settings. Firstly, the private-shared space formulation entails a significant advantage fitting to the inherent nature of this problem: the shared space represents the underlying annotation which is common to all annotators, while the private space is able to isolate the portions of the signal which are uninteresting and specific only to one annotator. Furthermore, the dynamic nature of the model enables smoothing over noise of various nature apparent in such annotations (e.g., false positives, errors originating from the imperfect handling of input devices etc.). Note that a significant issue in fusing continuous *in-time* annotations is the various temporal discrepancies that are exhibited by annotators, a consequence of varying human response times, the level of concentration of the annotator and so on. To accommodate for this issue, we extend DPCCA by incorporating a *time-warping* process in the model, which corrects the temporal misalignments manifested in the annotations. Moreover, as we show in Chapter 9, the specific formulation adopted allows for automatic ranking of annotations, including automatically discarding malicious annotations

(or spam). Finally, DPCCA is extended in order to include features and other observations in the derivation of the ground truth. This is crucial in many problems where the actual observations can be the only true reference to the actual annotated sequence, especially when the annotations are very noisy. In effect, DPCCA tackles all problems that arise when fusing multiple continuous annotations. Firstly, DPCCA exploits both (i) the existence of *multiple* annotations as carriers of portions of the true annotations, and (ii) the availability of any features which can act as objective references to the sequence at hand, such as audio and visual features. These are in fact the only information at hand which can disambiguate the existing annotations. Summarising, DPCCA (i) isolates annotator-specific spatial bias, (ii) nullifies temporal discrepancies of annotators, (iii) exploits any available features, (iv) models dynamics (v) ranks the efficacy of annotators and finally, (vi) provides a probabilistic estimation of the "ground truth" as a representation of a clean, shared space underlying all annotations.

Temporal Alignment of Human Behaviour (Chapters 9 and 10)

As mentioned in Section 1.2, the problem of temporal alignment of human behaviour, and sequences is general, carries particular significance and is often encountered within the realms of computer vision. In this thesis we approach this problem with two different models, which both share the same principles of design (see also, Section 1.3.1). The first method, DPCCA, is a probabilistic approach which, as discussed above, is particularly fitted for the problem of fusing annotations, providing the modelling of latent dynamics as well inferring a probabilistic measure of uncertainty. The second method, which we coin Robust Canonical Correlation Analysis (RCCA), is particularly suited to high-dimensional data which are corrupted by non-Gaussian noise, as are e.g., occlusions and other forms of gross noise.

- **Probabilistic Temporal Alignment via DPCCA (DPCTW).** In Chapter 9, we derive a probabilistic, private/shared space model which can be used in order to temporally align sequences. Unlike previous works targeting temporal alignment, this method can handle an arbitrary number of sequences, model temporal dynamics, as well as infer the shared space of all sequences in a probabilistic manner. The advantage of this formulation is that information which is private to a specific sequence is isolated in the private space and does not influence the shared space. With the private space modelling noise and bias, the shared space captures the commonality of the observations, which is subsequently temporally aligned. This effectively allows for temporally aligning the shared characteristics of temporal sequences, even though each may carry some unique

information. Finally, DPCCA allows for a ranking of the observation sequences in terms of the contribution to the "shared" information conveyed by the entirety of observation sequences.

- **Robust Temporal Alignment via RCCA (RCTW).** Most of the CCA variants in literature are based on a the Gaussian-noise assumption. Nevertheless, in many real-world applications, the presence of gross types of noise is observed (E.g., gross errors due to incorrect localization and tracking, presence of partial occlusion, as well as outliers). These types of errors rarely follow a Gaussian distribution, which as aforementioned is the de-facto assumption in most methods. To this end, in Chapter 10 we propose a robust variant of Canonical Correlation Analysis, utilising low-rank approximation and sparse errors. Given a set of high-dimensional observations corrupted by gross noise and by incorporating a temporal alignment step, the method is able to temporally align the observation sequences in a clean (from gross errors) latent space. As we show in both real and synthetic experiments, this method is able to outperform other static variants of CCA which appear unable to cope with gross noise.

Modelling Temporal Dynamics (Chapters 5, 6, 9, 11)

As discussed, the concept of modelling temporal dynamics and in effect temporal dependencies is extremely crucial in terms of analysing human behaviour, especially when the observations consists of video or audio sequences and not of static images. The modelling of dynamics is a common feature which is exhibited by the models proposed in this thesis. In Chapter 5, the employed BLSTM-NNs are inherently able to model such dependencies, and in-effect the proposed output-associative fusion, consisting of BLSTM-NNs, is able to model both short and long term temporal dependencies. The probabilistic OA-RVM presented in Chapter 6 is able to do so by applying a temporal window to the output features, thus utilising neighbouring temporal information. The DPCCA (Chapter 9) model is equipped with latent spaces which model temporal dependencies, in effect by modelling the temporal evolution of the signal in latent states following a first-order, directed Markov chain. In fact, in Chapter 11, where we propose a unified framework for probabilistic component analysis, imposing temporal dependencies in probabilistic component analysis models becomes straightforward, by simply incorporating a Markov Random Field (MRF) with temporal connectivity in the latent space. Note that other models which do not model dynamics are primarily tested on static data and are based on deterministic CCA, which is inherently a static model. Nevertheless, temporal dependencies can be modelled by employing sliding temporal windows or other feature transformations, such as temporal kernels. Finally, we summarise by highlighting that the results

presented in this work crystallize the significance of modelling dynamics in the problem of continuous emotion dimensions and human behaviour in general, as in all cases the inclusion of temporal capabilities increases the obtained accuracy.

The Fusion of Multiple Modalities (Chapters 5, 7, 9, 10)

As aforementioned, many open problems in affective computing and human behaviour analysis revolves around fusion, where typically information from multiple modalities (such as audio and video) convey complementary information. In the first part of thesis, and namely in Chapters 5, 6 and 7, we perform an experimental evaluation of various fusion techniques while providing answers to questions such as *which modality is better for predicting which emotion dimension*, and *which fusion method is more suitable for predicting specific emotion dimensions*. We compare against several widely used fusion techniques such as model-level fusion, where fusion is performed as an added layer on already trained models and feature-level fusion, where features from multiple cues and modalities are simply concatenated. We propose fusing multiple modalities via model level BLSTM-NN fusion 5, while we also show how one can fuse multiple observation sets by utilising a block matrix formulation in CCA (Chapter 7. Subsequently, driven by a fundamental idea proposed in this thesis, namely that emotion dimensions exhibit spatial and temporal correlations which can be utilised to improve the accuracy of predictive analysis, in the related chapters we propose fusion techniques which essentially incorporate output information and dependencies during learning from observations. This line of work has been described in the subsection *Learning Continuous Emotion Dimension via Exploiting Output Correlations* previously in the current section.

Subsequently, in the second part of the thesis, we propose novel fusion techniques, mostly founding on the shared-space principle. Firstly, in Chapter 9, we introduce a novel dynamic model which aims to isolate private information specific for each observation sequence and learn any arising commonality amongst the observation sets. Although the method can be generally utilised for fusion, we apply the method in Chapter 9 to the fusion of annotations with observations in order to assist the inference of the clean "ground truth" signal from multiple noisy observations. Subsequently in Chapter 10, we introduce RCCA, a robust variant of CCA, which aims to isolate gross errors and learn a clean, common subspace for the observation sets. RCCA is utilised for robust fusion of multiple modalities. In Chapter 10, we evaluate RCCA on many fusion related problems, such as the audio-visual fusion for the detection of interest along with heterogeneous face fusion/recognition, where images of subjects attained via different sensors (e.g., visual images, 3D maps (depth information), infrared images as well

as hand-drawn sketches) are fused, and the extracted features are used for classifying. Note that in this scenario, we also examine the very challenging scenario when one of the modalities is missing during testing, as well as the scenario where the set of testing classes and training classes is disjoint, and thus during testing the goal is to match the multiple modalities amongst themselves to obtain a classification. Note that the methods defined within out unifying framework for probabilistic component analysis (Chapter 11) are not currently developed for fusion, but can be easily extended to do so.

Dimensionality Reduction via Probabilistic Component Analysis: Face Analysis and Visualisation (Chapter 11)

In Chapter 11 we present a Unified framework for Probabilistic Component Analysis, suitable for dimensionality reduction and feature extraction. The proposed framework has a great theoretical novelty, as for the first time, a probabilistic framework which unifies most well-known Component Analysis (CA) techniques is presented. In more detail, we present novel probabilistic models for applying Principal Component Analysis (PCA), Linear Discriminant Analysis (LDA), Locality Preserving Projections (LPP) and Slow Feature Analysis (SFA)). The models derived via our framework bear several advantages over equivalent methods, such as reduced complexity in comparison to deterministic equivalents, explicit noise modelling, estimating per-dimension variance (thus being able to rank the derived latent space in contrast to other probabilistic component analysis techniques), as well as allowing for more robust inference due to the probabilistic nature of the model. We evaluate the Expectation Maximisation (EM) based models derived from our framework on various problems. Firstly, we apply our EM Linear Discriminant Analysis (EM-LDA) on the problem of automatically detecting the level of interest of a subject on naturalistic, spontaneous data acquired in the Lisbon Zoo by a robot acting as a virtual guide. Furthermore, we evaluate the methods on the problem of Feature Extraction for Face Recognition on various popular databases such as PIE, YALE and AR under noisy settings. Finally, we evaluate the derived methods on the problem of *high-dimensional data visualisation* on the Frey Faces data. Via our experiments, we show that the theoretical advantages posed by our frameworks greatly reflect on the obtained results, as models derived via our framework outperform other, compared methods.

1.4 Publications

The work presented in this thesis has resulted in the following list of publications.

- **International Conferences**

[1] **M. A. Nicolaou**, S. Zafeiriou, and M. Pantic. A Unified Framework for Probabilistic Component Analysis. In *European Conf. Machine Learning and Principles and Practice of Knowledge Discovery in Databases (ECML/PKDD'14), Nancy, France*, 2014.

[2] **M. A. Nicolaou**, Y. Panagakis, S. Zafeiriou, and M. Pantic. Robust Canonical Correlation Analysis: Audio-visual Fusion for Learning Continuous Interest. In *Proceedings of IEEE Int'l Conf. Acoustics, Speech and Signal Processing (ICASSP)*, 2014.

[3] **M. A. Nicolaou**, Y. Panagakis, S. Zafeiriou, and M. Pantic. Correlated-Spaces Regression for Learning Continuous Emotion Dimensions. In *The 21st ACM International Conference on Multimedia (ACM MM)*, 2013.

[4] Y. Panagakis, **M. A. Nicolaou**, S. Zafeiriou, and M. Pantic. Robust Canonical Time Warping for the Alignment of Grossly Corrupted Sequences. In *Proc 26th IEEE Int'l Conference on Computer Vision and Pattern Recognition (CVPR). Oregon, USA*, 2013.

[5] **M. A. Nicolaou**, V. Pavlovic, and M. Pantic. Dynamic Probabilistic CCA for Analysis of Affective Behaviour. In *Proceedings of the 12th European Conference on Computer Vision (ECCV). Florence, Italy*, 2012.

[6] **M. A. Nicolaou**, H. Gunes, and M. Pantic. Designing Frameworks for Automatic Affect Prediction and Classification in Dimensional Space. In *Proceedings of IEEE Int. Conf. Computer Vision and Pattern Recognition (CVPR-W), Workshop on Gesture Recognition*, Colorado Springs, USA, 2011.

[7] **M. A. Nicolaou**, H. Gunes, and M. Pantic. Output-Associative RVM Regression for Dimensional and Continuous Emotion Prediction. In *Proceedings of IEEE International Conference on Automatic Face and Gesture Recognition (FG)*, Santa Barbara, CA, USA, March 2011.

[8] **M. A. Nicolaou**, H. Gunes, and M. Pantic. Audio-visual Classification and Fusion of Spontaneous Affect Data in Likelihood Space. In *Proceedings of Int'l Conf. Pattern Recognition (ICPR)*, Istanbul, Turkey, 2010.

[9] **M. A. Nicolaou**, H. Gunes, and M. Pantic. Automatic Segmentation of Spontaneous Data using Dimensional Labels from Multiple Coders. In *Proceedings of LREC Int'l Workshop on Multimodal Corpora: Advances in Capturing, Coding and Analyzing Multimodality*, Valletta, Malta, 2010.

- **Journal Articles**

[1] **M. A. Nicolaou**, S. Zafeiriou, and M. Pantic. A Unified Framework for Probabilistic Component Analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI). Submitted - under revision*.

[2] **M. A. Nicolaou**, V. Pavlovic, and M. Pantic. Dynamic Probabilistic CCA for Analysis of Affective Behaviour and Fusion of Continuous Annotations. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 2014.

[3] **M. A. Nicolaou**, H. Gunes, and M. Pantic. Output-associative RVM regression for dimensional and continuous emotion prediction. *Image and Vision Computing (IMAVIS), Special Issue on The Best of Automatic Face and Gesture Recognition 2011*, 2012.

[4] **M. A. Nicolaou**, H. Gunes, and M. Pantic. Continuous Prediction of Spontaneous Affect from Multiple Cues and Modalities in Valence-Arousal Space. *IEEE Transactions on Affective Computing (TAC)*, 2011.

- **Book Chapters**

[1] O. Rudovic, **M. A. Nicolaou**, and V. Pavlovic. *Machine Learning Methods for Social Signal Processing*. Cambridge University Press, 2014. *To appear*.

[2] H. Gunes, **M. A. Nicolaou**, and M. Pantic. *Continuous Analysis of Affect from Voice and Face*. Springer-Verlag, 2011.

1.5 Thesis Outline

The rest of the thesis is structured as follows. In Chapter 2, we present an introduction to the problem of affective analysis, covering mostly continuous dimensional emotion descriptions, the perception of affect from multiple modalities, as well as discuss common feature sets employed in the field. Subsequently, in Chapter 3 we introduce a set of machine learning

techniques which are relevant to this thesis, such as regression, component analysis and time warping for temporal alignment. The main body of this thesis is separated into two main parts. While both parts revolve around both affective computing, machine learning and computer vision, the first is primarily attentive to affective computing while the second is more machine learning oriented.

In more detail, the first part, consisting of Chapters 5, 6 and 7, is more focused on exploring the problems arising from adopting continuous and dimensional emotion descriptions. We perform an empirical analysis of the problem and identify the idiosyncrasies which, when taken into account during the design of a system will prove beneficial. In more detail, we focus on prediction by exploiting output-correlations (i.e., correlations amongst emotion dimensions), an idea which is proposed and implemented in this work for the first time in literature (specifically, in [174]). In more detail, in Chapter 5 we perform an initial approach to the problem by utilising various regression techniques such as the Bidirectional Long-Short Term Memory Neural Networks (BLSTM-NN), while we examine the efficacy of utilising several modalities and cues (visual consisting of facial expressions and shoulder movements, as well as audio cues) in terms of predicting continuous emotions. We propose the utilisation of output-correlations in a form of fusion (which we dub output-associative fusion), aiming to learn output patterns commonly occurring in our data and in effect obtain better models for predictive analysis. In Chapter 6, we formalise the concept of learning output correlations in emotion dimensions further. We introduce a novel, probabilistic framework based on the Relevance Vector Machine (RVM) which can learn spatio-temporal output dependencies while adopting sparse probabilistic learning. Finally, Chapter 7 presents a simple idea on using Canonical Correlation Analysis (CCA), a component analysis method aimed at analysing multiple observation sets, in order to correlate observations to emotion dimensions, while removing any redundancy arising in emotion dimensions. This is achieved by diagonalising the output covariance matrix, and in effect facilitating the utilisation of single-output models without loss of information.

The second part of the thesis is more closely attached to component analysis. In more detail, in Chapters 9, 10 and 11, we propose a set of novel probabilistic and deterministic component analysis techniques and frameworks. While, as in the previous part, we are in many cases driven by an attempt to solve a particular application related problem, this does not constrain in any way the generality of application, as the solutions we provide are fitting to many other fields and domains with similar settings, as we discuss. Firstly, in Chapters 9 and 10, we propose two different component analysis methods which have a common principle: the discovery of a "shared-space", an underlying commonality in all observation sets.

Dynamic Probabilistic Canonical Correlation Analysis (DPCCA), proposed in Chapter 9, is a general method for probabilistically inferring the shared and private information conveyed by observation sequences. Via this method, we tackle the problem of fusing a set of multiple annotations in a formant and elegant framework by preserving the common information underlying all annotations. DPCCA also features temporal warping, ranking of annotations as well as temporal modelling. In Chapter 10, we present the Robust Canonical Correlation Analysis (RCCA), a robust variant of CCA which is able to isolate non-Gaussian noise due to gross errors in the observation sets. RCCA is also extended with time warping, where the temporal alignment takes place in the discovered error-free latent subspace. The applicability of RCCA is evaluated on problems such as the temporal alignment of human behaviour, the multi-modal fusion from multiple sensors (such as e.g., facial images obtained via 3D and infrared sensors) as well as other related problems. Finally, our efforts in Chapter 11 are initially driven by a theoretical problem; the formulation of a unifying, probabilistic framework which unifies all component analysis techniques, which can be formulated as trace optimisation problems with no domain constraints for the parameters. By utilising Markov Random Fields (MRFs), we specify a probabilistic model which can be solved via Expectation Maximisation, and by manipulating the MRF latent prior one can achieve equivalent solutions to CA methods such as Principal Component Analysis (PCA), Linear Discriminant Analysis (LDA), Locality Preserving Projections (LPP) and Slow Feature Analysis (SFA). As discussed and shown via various experiments such as level of interest classification, face recognition and visualisation, the models derived via our framework offer various advantages with comparison to other equivalent (in terms of projection) methods.

Chapter 2

Affect Sensing: Background & the State-of-the-art

Contents

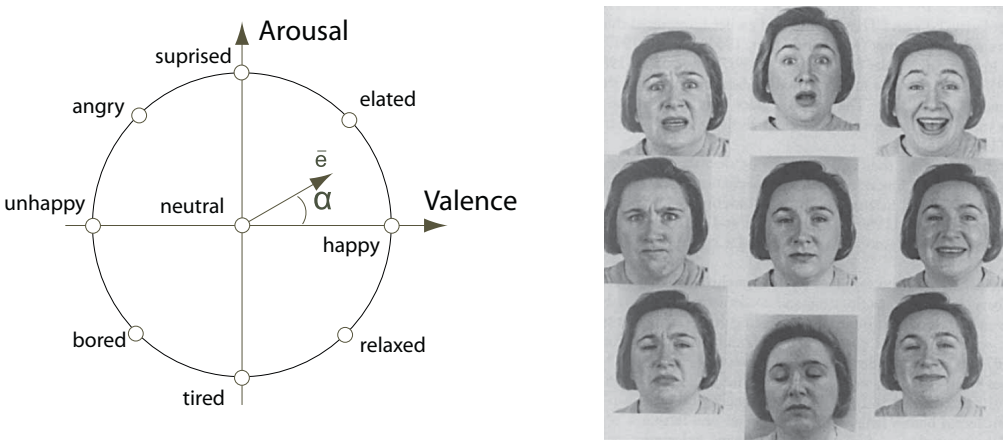
This chapter revolves around affect sensing and analysis. In particular, herein we provide the reader with the necessary background in affect sensing, focusing mostly in terms of continuous emotion dimensions. Furthermore, we provide a review of related literature which is most relevant to the general research directions followed by this thesis[1]. In more detail, in Section 2.1, we firstly discuss the adoption of a dimensional emotion descriptions, along with related work mostly in terms of predictive analysis on continuous emotion dimensions. In Section 2.2, we discuss the transition from posed to spontaneous expressions, while in Section 2.3, we review the basic concepts which relate to the perception of emotions from modalities (such as the visual and audio), while also discussing their fusion. In Section 2.4, we further discuss the significance of modelling temporal dynamics as far as the automatic sensing of

[1] We note that related work which is particular to specific chapters is further analysed in the relevant chapter introduction.

human behaviour is concerned. In Section 2.5, we detail the process of feature extraction from relevant modalities, with particular emphasis placed on facial expressions from the visual modality. Finally, we discuss issues related to the data, such as describing various commonly employed databases (Section 2.6), while also referring to the problem of obtaining reliable annotations (Section 2.7). Finally, we conclude the chapter in Section 2.8. For more details, the reader may refer to [285], [98].

2.1 Continuous and Dimensional Emotion Descriptions

As discussed in Chapter 1, the adoption of continuous, dimensional emotion descriptions arises as a direct consequence of several recent trends emerging in affect sensing, such as the need to accommodate a wider variability of emotion descriptions along with capturing emotions most often encountered in everyday life. The description of affect via the utilisation of latent dimensions dates back to the work of Russell [216], with similar approaches taken on in many works in psychology, such as [132, 244] and [242] (c.f., Figures 2.1 and 2.2).



(a) (b)

Figure 2.1: (a) Russell's valence-arousal space. The angle is represented by α while the vector $\bar{e}$ represents the emotion (point) as a parameter of valence and arousal . (b) Nine facial expressions arranged in the ordering of (a). Image adapted from [200].

While the main basic dimensions of emotion, valence and arousal, are deemed to capture most affective variability encountered in interactive scenarios, other dimensions are often defined in psychology literature, such as *potency* or *power*, referring to the degree of control that the individual feels with respect to the emotional state [56, 163, 183]. In fact, several works consider the dimensions of power and expectation to be significant carriers of affective information [79]. While the first databases adopting elicited spontaneous behaviours along with continuous

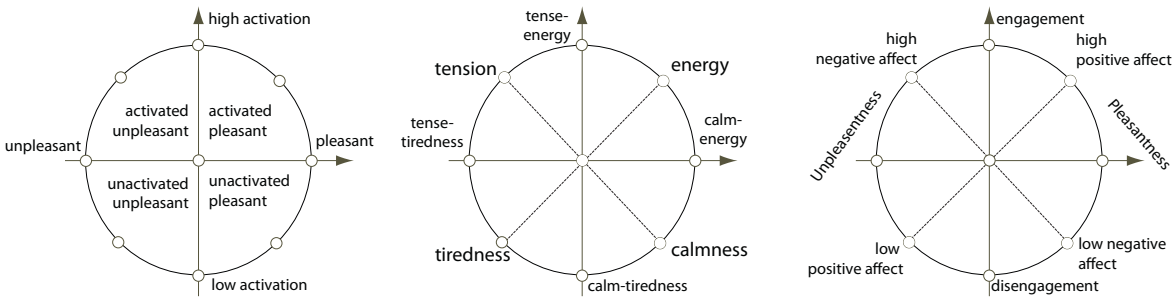


Figure 2.2: Other 2D emotion categorisation approaches: (a) Approach of Larsen and Diener [132] (b) Thayer [244], (c) Watson and Tellegen [242].

annotations where only annotated in terms of valence and arousal (e.g., the Sensitive Artificial Listener, [64]), the introduction of more recent databases such as SEMAINE [157] adopted affective annotations in terms of 5 total emotion dimensions, summarised in what follows.

- **Valence** refers to the positive or negative feeling of the subject's emotional state.
- **Arousal/Activation** points to the subject's feeling of dynamism or lethargy, i.e. how passive or active the emotion state of the subject is.
- **Power** dimension consists of both power and control over the emotion, with more emphasis placed on the power which the emotion holds over the subject.
- **Anticipation/Expectation** relates to control in terms of the domain of information, i.e. expecting an event or dialogue or not.
- **Intensity**, closely interweaved with arousal, points to how far the emotional state of the subject diverts from a rational, cool state.

We clarify that each emotion dimension is usually normalised between [-1, 1], ranging from e.g., negative to positive for Valence, passive to active for Arousal and so on. In cases where an emotion dimension can be either present or absent, (and not ranging the spectrum between two polar opposites), the normalisation is usually done between [0,1], actually representing the intensity of the presence of the emotion (i.e. with 0 corresponding to not present and 1 present). For further details on different approaches to modelling human emotions and their relative advantages and disadvantages, the reader is referred to [220, 87].

There has been a significant increase of work in modelling continuous and dimensional emotions during the past years. Since annotations are provided in a continuous space, many systems that target automatic dimensional affect recognition, generally tend to quantize the continuous range into certain levels. A commonly employed strategy is to map the problem of classifying the six basic emotions to a three-class valence-related classification problem: positive, neutral, and negative emotion classification (e.g., [283]). A similar simplification is to reduce the dimensional emotion classification problem to a two-class problem (positive vs. negative or active vs. passive classification) or a four-class problem (classification into the quadrants of 2D V-A space; e.g., [39], [80, 85, 110, 275]). For instance, [267] analyses four emotions, each belonging to one quadrant of the V-A emotion space: high arousal positive valence (joy), high arousal negative valence (anger), low arousal positive valence (relief), and low arousal negative valence (sadness). Furthermore, [122] discriminates between high-low, high-neutral and low-neutral affective dimensions, while [151] uses the SAL database and quantizes the V-A into 4 or 7 levels and uses Conditional Random Fields (CRFs) to predict the quantized labels.

Methods for discriminating between more coarse categories, such as low, medium and high [126], excited-negative, excited-positive and calm neutral [41], positive vs. negative [172], and active vs. passive [39] have also been proposed. Of these [39] uses the SAL (Sensitive Artificial Listener) database and combines information from audio (acoustic cues) and visual (Facial Animation Parameters used in animating MPEG-4 models) modalities. The authors of [172] focus on audiovisual classification of spontaneous affect into negative or positive emotion categories, and utilize 2- and 3-chain coupled Hidden Markov Models and likelihood space classification to fuse multiple cues and modalities. Kanluan et al. [116] combine facial expression and audio cues exploiting SVM for regression (SVR) and late fusion, using weighted linear combinations and discretized annotations (on a 5-point scale, for each dimension).

The works which model dimensional emotion descriptions in a continuum are even more recent. Many of these works deal exclusively with speech (i.e., [275], [151], [91]). The work presented in [275] utilizes a hierarchical dynamic Bayesian network combined with BLSTM-NN performing regression and quantizing the results into four quadrants (after training). The work by Wöllmer et al. uses Long Short-Term Memory neural networks and Support Vector Machines for Regression (SVR) [151]. Grimm and Kroschel use SVRs and compare their performance to that of the distance-based fuzzy k-Nearest Neighbour and rule-based fuzzy-logic estimators [91]. The work of [96] focuses on dimensional prediction of emotions from spontaneous conversational head gestures by mapping the amount and direction of head

motion, and occurrences of head nods and shakes into arousal, expectation, intensity, power and valence level of the observed subject using SVRs. Several recent works focus on multiple modalities (e.g., combining visual and auditory cues). For instance, Eyben et al. [75] propose a string-based approach for fusing the behavioural events from visual and auditive modalities (i.e., facial action units, head nods and shakes, and verbal and nonverbal audio cues) to predict human affect in a continuous dimensional space (in terms of arousal, expectation, intensity, power and valence dimensions). Metallinou et al. in [164] focus on analysing the vocal and body language behaviour (via MoCap features) of pairs of actors improvising dyadic interactions. For each actor's recording, they computed the Spearman correlation coefficient between the mean annotation and the MLE curve. Activation and dominance were predicted from visual and audiovisual cues reasonably well. Another representative approach is that of Gilroy et al. [84] where a dimensional multimodal fusion scheme is proposed, in order to support detection and integration of spontaneous affective behaviour of users (in terms of audio, video and attention events) experiencing arts and entertainment. At this point, we note that as discussed in Chapter 1 an important contribution of this thesis is the idea of utilising relationships exhibited in emotion dimensions for learning. This has led to the adoption of this idea by other researchers in the field, including recent works such as [206] and [13], which utilise Conditional Random Fields (CRFs) to this end. Finally, it is interesting to note the recent establishment of various workshops dedicated to the topic in related conferences, such as the *Emotion Synthesis, Representation, and Analysis in Continuous space* workshop, dealing particularly with the topic of continuous dimensional emotion descriptions, as well as the the Audio/Visual Emotion Challenge and Workshop (AVEC) [257] which includes evaluation in the continuous domains of valence and arousal.

2.1.1 Modelling the Level of Interest

Although the level of interest is not traditionally considered as part of the latent dimensions which describe the affective state, the automatic detection of interest in audiovisual sequences has been gaining rising attention amongst researchers, in both the fields of affective computing and pattern recognition and machine learning [195, 227, 228]. From a psychology perspective, interest has been extensively studied since 1910 [7], and has since then been considered as an *emotion* by various experts [250, 236], while several works have stated that interest is correlated to emotion dimensions, mostly with arousal and secondly with valence [130]. Interest is commonly defined as *an emotion that causes the subject to focus his or hers attention to the event taking place* [236], and in conclusion, one can consider the magnitude of interest as a continuous dimension. As can be understood, the detection of interest (and, similarly, *engage-*

ment) is crucial for a vast number of applications, ranging from virtual guides to interactive learning systems as well as enhancing the experience of human-computer interaction. Most of related work on the automatic detection of interest [228, 226, 227] treats interest as a discrete emotion, focusing on classification in terms of discrimination between interest/non-interest, as well as discriminating amongst classes e.g., disinterest, indifference and interest. This is in line with traditional research in affective computing and emotion theory, which focuses only on a set of discrete emotions, such as anger and joy, but lacks the expressive variability of a dimensional approach. We note that, as discussed in the introduction (Chapter 1), in this thesis we attempt to treat interest as a continuous emotion dimension (c.f., Chapter 7, Chapter 10).

2.2 Posed vs. Spontaneous Emotional States

Affect recognition systems are often criticised in terms of the difficulties which arise when deploying them under real-world conditions. This arises not only due to the constraints and assumptions that are usually undertaken when training such systems (e.g., constraint, laboratory environment etc.), but also to the *type* of behaviour which is utilised for training. In more detail, affect sensing traditionally focused on *posed* emotion expressions, i.e. where actors or subjects where asked to exhibit an expression. As discussed in Chapter 1, this leads to behaviours which are quite unlike their spontaneous equivalents, since *spontaneous* emotion expressions are more complex and do not follow a set of strict temporal phases, e.g., beginning and finishing in *neutral*, with all facial muscles relaxed. As a result, many researchers shifted their attention to modelling spontaneous human affect, in order to accommodate the increasing demand for robust affect sensing under real-world conditions. The practical implications of this shift lead to a multitude of challenges, arising mostly from recordings taking place in much more unconstrained scenarios, where there is less control over lighting conditions, the movement of subjects is much less constrained while various occlusions may manifest in the recordings, e,g., by body parts, other persons or even foreign objects (such as microphones or headsets).

Many recent studies focus on the analysis of spontaneously manifesting affective states, by utilising both facial expressions [18, 44, 256, 8] as well as acoustic features [20, 134]. Interesting findings relate to the differences between spontaneous and posed expressions. There has been a lot of work in detecting differences between spontaneous and posed behaviour by the Affect Analysis Group[2], while the temporal characteristics of phases as described in Section 2.4 have been found important in detecting spontaneous vs. posed smiles [44, 258]. It is also

[2] http://www.pitt.edu/ emotion/publications.html

significant to note the importance of modality fusion in discriminating between posed and spontaneous emotions. It is typical for spontaneous body expressions to be manifested along with an agreeing facial expression. There are different views on whether body motions or facial expressions are most expressive of the spontaneity of the emotional expression. The two main factors that contribute to this is the difficulty of control and the conscious censoring that humans can impose. Darwin's views support the facial expressions, since as he claimed, the body expressions are more easy to control. Looking at this problem from a different angle, Ekman [71] supports that humans usually try to censor their face (since as Ekman supports, humans are more concious of their facial expression) thus the body expressions would be more prone to expressing uncensored information. There has been work that also suggests that truthful and deceptive behaviour differs on the number of head movements [34, 33], or the lack of accompanying gestures [60].

Some examples of systems discriminating spontaneous from posed behaviour include [258], which discriminates spontaneous from posed smiles by utilising geometric features and multi-modal fusion using head movement, facial expressions and shoulder gestures. Based on the data, the temporal facial states are detected, along with the activated AUs, while GentleBoost and Support Vector Machines (SVM) are used for the classification. Experimentation also occurs with modifying the abstraction level of fusion (early, mid-level and late), while the authors conclude that from the specific results, the head pose seems to be the most important modality. Another example is that of Littlewort et al. [145], discriminating between real vs. fake pain. The system utilises Action Units (AUs) to encode facial expressions, using 20 AU classifiers with input data images of posed and spontaneous facial expressions. The authors presented better accuracy compared to human FACS experts (72% to 52%), while they argue that such a method could be also used for other spontaneous expressions. It is important to note that in general, research on spontaneous vs. posed expressions, whether it is from psychology or in developing affect recognition systems agrees that the temporal dynamics appear to be highly significant in terms of determining one from another [285].

2.3 Modalities and Emotion Perception

In this section, we refer to the various modalities typically employed for affect sensing. We firstly focus on the *visual* modality, where we discuss facial expressions and body gestures. Subsequently, we discuss the perception of emotion from audio cues, as well as provide a brief reference to measuring emotions from physiological parameters. Finally, we discuss the issue of multi-modal fusion.

2.3.1 Visual Modality

Facial Expressions

In order to model the multiple, complex human facial expressions, Ekman and Friesen developed the Facial Action Coding System (FACS) [72] in 1978. This model provided a taxonomy of facial expressions, and is widely accepted as a de-facto standard utilised in order to categorise the facial expressions of emotions. Based on Carl-Herman Hjortsjö's book on the anatomy of facial features [104], the FACS model consists of 32 atomic facial muscle actions, (Action Units, AUs), which in turn represent the contraction or relaxation of one of the facial muscles (Fig. 2.4). An important advantage of the FACCS model is that the annotation of facial expressions is moved away from a subjective, personal interpretation of the annotator to an objective representation of human expressions, which is observer-independent - although usually an expert is required to correctly identify the activated facial muscles and thus, the activated AUs. A list of facial AUs can be found in Fig. 2.3.

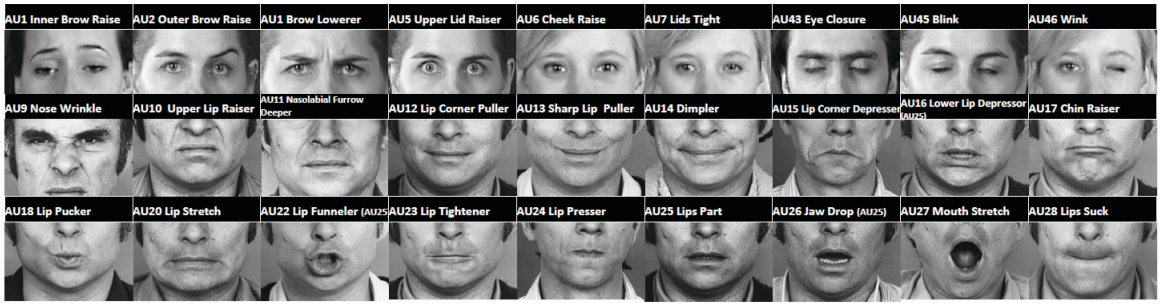


Figure 2.3: Facial Action Units (AUs), with 9 AUs for the upper face and 18 for the lower, containing images from [72] and [189]. Figure adapted from [214].

Body & Gestures

Researchers have long attributed the expression of emotional states through body movement and bodily gestures (e.g. [100, 3, 167]), originating from the work of Darwin on the description of animal and human emotion expression. Various research has also supported that emotional states can be disambiguated via analysing body expressions [259], while also indicating that a better appreciation of emotional states can be achieved by analysing the entire body. In some limited cases, studies have shown that body gestures can be as significant as voice and facial expression modality [47]. There has been research in combining posture and body information

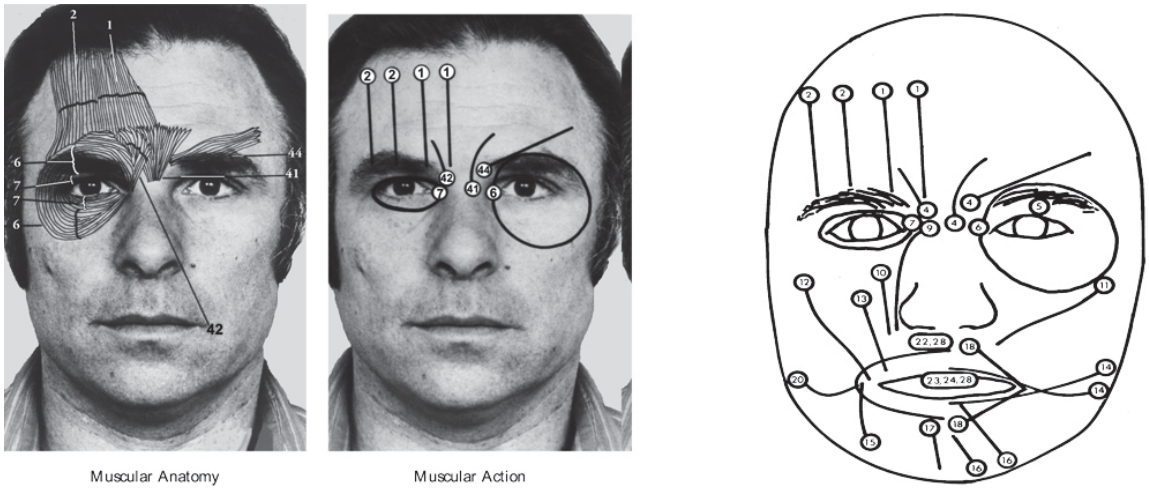


Figure 2.4: Left: Relation between muscular anatomy and muscular actions (Action Units). Right: The AUs of FACS. Circle represents fixed point towards which skin is pulled along the line during activation while number represents the AU. Both images adapted from [72].

with kinematics [67, 92, 93, 94], while there were also attempts to relate emotions to kinematic[3] data (e.g., joint angle data for head tilt, rotation, neck flexion, shoulder abduction, elbow flexion and knee flexion) and gait parameters[4] (velocity, cadence or steps per minute). Results for such attempts varied and demonstrated a difficulty in recognising emotions such as anger, while attaining best performance in recognising sadness. The most characteristic parameters expressing emotion were related to limb motion and general posture. It is important to note that, in contrast to facial expressions, there is no standardised method in interpreting human postures and gestures (like FACS) and no equivalent to AUs, although there have been efforts in that direction (e.g. [127]).

2.3.2 Audio

Audio and speech are essential carriers of human affect. The acoustic behaviour of humans is separated into the transfer of linguistic, paralinguistic and extralinguistic information, although only linguistic and paralinguistic are communicative [133]. The *linguistic* part is refers to language itself, being precisely the explicit verbal part of the communication. The *paralinguistic* element refers to the non-verbal part of the communication, which is used as to modify the verbal meaning, or convey emotion (e.g. falsetto in mocking), whether it is expressed unconsciously or consciously. Features such as volume, pitch and intonation are related to

[3] Kinematics is a branch of classical mechanics which relates to the description of motion

[4] Gait analysis is related to the quantization of parameters in order to help athletes improve their performance or identify posture related problems

paralinguistics. The *extralinguistic* element refers to informative but not communicative information which might e.g. identify the speaker from overall pitch and loudness of speech. The extralinguistic part refers to information which has no conventional meaning, but is unintentional, for example pitch differentiation based on age and sex [51]. Usually in emotion [285, 97] and speech recognition [223], the discrimination is between verbal (linguistic) and non-verbal (paralinguistic, extralinguistic) elements of speech. Important information with respect to the expression of emotions is deemed to be conveyed in the paralinguistic part, while it has been reported that spoken messages are not reliable in expressing affective behaviour [169], as e.g. a different selection of words is used by different persons in order to express the same affective state, while other difficulties can be for instance, in cases where human speakers refer to emotional states which are irrelevant to their current emotional state. Despite the difficulties, there have been attempts to generate dictionaries of words and affective states, e.g. Whissell's dictionary of affect in language [270], which is essentially a list of 4000 words, with a 2D rating in the activation/evaluation space.

On the other hand, implicit paralinguistic messages are deemed to provide significant contribution towards emotion recognition, while parameters which have been identified as strong indicators of emotions are continuous acoustic measures, especially those who relate to the pitch (fundamental frequency) such as frequency range, the mean, median and variability values [97]. Further detailed surveys in this area include [210] and [137], while a survey of acoustic features is presented in [49]. It is important to note that while the identification of the optimal feature set is yet an open problem, human listeners are accurate in detecting basic emotions from prosody features (rhythm, stress, intonation) [210] and some non-basic affective states from non linguistic vocalisations like laugh, cries, sighs and yawns [203]. A recent, systematic survey on *computational paralinguistics* including tools and techniques can be found in [225].

2.3.3 Physiological Parameters & Heat

There have been other methods of attaining results and measurements of human affective states, to which we will refer briefly in this section. Firstly, we refer to measuring physiological parameters or bio-potential signals. The range of parameters ranges from measuring brain signals by functional Near Infrared Spectroscopy (fNIS), scalp signals by electroencephalogram (EEG), peripheral signals such as cardiovascular activity, electrodermal activity, Galvanic Skin Response (GSR) and the electromyograph (EMG). It is believed that these measurements can be translated to the valence-arousal emotion space. Furthermore, research results suggest a correlation between emotional states and core body temperatures of mammals, e.g. the change

in the facial temperature of monkeys when they are under stressful situations, or the body temperature of rats under similar fearful situations. It is also notable that a correlation has been found between measurements in blood flow and changes of affective states [252, 192], due to thermo-muscular activity. Thus, by obtaining objective measurements of the skin temperature change, there is a possibility of obtaining information for affect states of subjects. Again, a generic framework for these measurements is yet to be defined. For more details, the reader can refer to [97].

2.3.4 Fusing Modalities

A significant issue relating to affect sensing and automatic behaviour analysis lies in the appropriate fusion of multiple modalities. Clearly, in human-to-human communication the combination of information conveyed from speech, gestures and facial expressions is essential in order to disambiguate the actual conveyed emotion [160, 159]. In human communication, the modality information is fused either consciously or subconsciously. McNeil emphasises what he calls the *conceptual expression* of gestures in combination with language, as he claims that the speaker is thinking in images and in words, expressing words by language and images by gestures. It is suggested facial expressions and vocal characteristics (tone of voice, prosody) strongly influence each other ([155, 57, 239]). It has also been reported that body expressions disambiguate the classification of facial expressions, as well as influence vocal features such as tone [259]. Summarising, these findings point to the significance of properly fusing modalities when analysing human behaviour. This includes balancing the contribution of modalities (i.e., properly weighting cues which are better for analysing particular behaviours). Also, in many cases the modality information can be incongruent (i.e., disagreeing information). Meeren et al. [162] investigate the agreement and conflict of facial and body modalities, by presenting images of faces on body's to participants, with agreeing (e.g. happy face on happy persons body) or conflicting information (sad face on an exited persons body). The human participants opted towards the trusting the body expression where the information was conflicting, leading to an indication of the importance of bodily expression in the presence of ambiguous facial expressions. The most common employed fusion techniques are feature-level fusion (where the features are simply concatenated and normalised) and decision-level fusion, where a predictive analysis algorithm is trained on single modalities, and the results are subsequently fused [285].

2.4 The Significance of Temporal Features

As has been hinted in previous sections, the modelling of temporal dynamics is of crucial importance to affect recognition, since such information provides further indications regarding the affective state of the subject, which may not be conveyed if one observes each temporal instance in isolation. For example, in [44, 258], it is shown that the timing of smiles can demonstrate whether a subect is *posing* a smile or not. The significance of modelling such temporal characteristics has been shown in many studies, such as [6] where the importance of time slices against stills in personality judging is denoted and in [221], where discussion involves temporal features of "social" expressions such as smiles and other expression components such as yawning and eyebrow flashing. The sequence of temporal phases of facial expressions (based on Ekman's work [68]) can be described as follows.

- **Neutral.** The neutral phase is when there are no manifestations of muscle activation and the face is considered to be relaxed.
- **Onset.** Onset phase occurs at the beginning of an action, where the activity in the facial muscles begins, and gradually increases in intensity.
- **Apex.** Apex is the plateau when the intensity of the motion stabilises.
- **Offset.** The last phase is the offset phase, where the muscular action begins to relax.

Typically, human facial expressions follow the above pattern, especially when the expressions are posed. In cases where the emotion can be spontaneous, it is likely that the sequence will not follow the precise steps defined above (e.g., two consequent smiles, with the second onset initiating during the first offset, i.e. offset $\rightarrow$ onset $\rightarrow$ apex). In Fig. 2.5, we show an example of such a case by illustrating a plot annotated with the intensity changes as well as the temporal phases. Furthermore, in Figure. 2.6 we show an example of a spontaneous smile, where as can be seen, an offset face is preceded by another onset and apex, instead of resting to the neutral position as usually happens in posed data.

Regarding the temporal structure of body gestures, there have been similar studies although much less explored. In general, a gesture can assume up to five temporal phases [97, 161]. These are defined as (i) the preparation phase, where the body parts move to the posture where the gesture stroke will commence from, (ii) the pre-stroke hold state, which occurs when the body parts hold in position, (iii) the where the peak of intensity is acquired in the stroke phase, (iv) the post-stroke hold, where the final gesture position is reached, and (iv) the

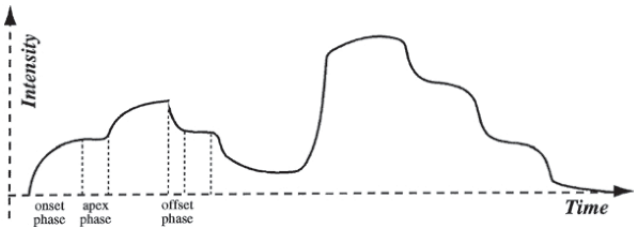


Figure 2.5: A hypothetical example from [74], where temporal facial phases are portrayed as functions of intensity. The neutral state is assumed to occur when intensity is around zero.

Figure 2.6: An example of a spotaneous smile from the UvE Nemo database. Note that the sequence of temporal phases during activation is not so strict in spontaneous behaviour. As can be seen, the expression changes from apex to apex in the second and third frames without firstly going through the neutral state.

retraction phase, where the body parts returns to the previous state. As argued in [161, 272] the only required part in this transitional process is the stroke, while all other phases are optional.

2.5 Feature Extraction and Pre-processing

In this section, we will refer to the typically employed methodologies used to extract features from various recordings. We mostly focus on facial expressions, which are utilised in the vast majority of work presented in this thesis. Two separate steps are usually employed when utilising facial expressions: the detection of the face, and subsequently the extraction of features. For completion, we also briefly refer to feature extraction from body movements and the audio modality.

2.5.1 Facial Expressions

Face Detection

In order to extract features from facial images, the first step consists of two parts: (i) determining whether a face exists or not in a given image, and (ii) determining the actual location of the face. This process is typically called *Face Detection*. While seemingly a relatively easy prob-

lem, especially when compared to highly complex modern computer vision problems, it can become highly challenging when studied under real-world conditions, with the manifestation of occlusions, uncontrolled variations in the head pose and varying illumination conditions. Typically, the problem is simplified by various assumptions, e.g., that there is only one face in the image [97] or by limitations in the posture of the person (front or profile view). Face detection is usually based on classifiers trained with positive and negative examples of faces, while modern methods for face detection are typically based on the Viola-Jones algorithm [263], which has been extended and improved in [142, 77]. Detailed surveys regarding the advances in face detection can be found in [294, 286].

Tracking

Having detected a face, a set of points must be localised on the face, e.g., via tracking. This process, often called facial point detection, may be be omitted in case only the texture of the face is required for the task-at-hand, but it is most often required since in most applications faces need to be spatially aligned and registered. Such methods can be based on texture or both texture and shape, whereas techniques based on shape also propagate information which essentially constraints the solution space, by disregarding e.g., anatomically impossible results. An example of a facial point detector based on local Gabor wavelets is presented in [266], later improved in [255, 154] by introducing graph-based constraints in order to validate the face shape. The main disadvantage of this family of detectors lies in being unable to cope with non near-frontal images. More recent methodologies which can deal with large pose variations as well as deal better with occlusions and varying illuminations have been proposed in [301, 278]. A very well known method based on both texture and shape refers to Active Appearance Models (AAM) [156]. In particular, AAMs define facial shapes via a 2D triangular mesh, while Principal Component Analysis (PCA)[5] is applied within each triangle in order to model the variation within. The reconstruction error is optimised in order to recover the optimal parameters via iterative gradient-descent. It is important to note that other commonly employed tracking methods such as Eigentracking [27], Lucas-Kanade [148] and Constraint Local Models (CLMs) [9] belong in the same family as AAMs, namely the Parametrised Appearance Models (PAMs) [171].

Feature Extraction

Having detected a face, the next step consists of extracting the desired features. In what follows, we discuss the most commonly utilised feature sets, commonly categorised into geometric

[5] Component Analysis (CA) is more formally introduced in Chapter 3.

and appearance based features [59, 285].

Geometric Features. As the name suggests, geometric features usually consist of encoded information regarding the location and the shape of the face and facial features (e.g., location of lips, eyes, nose and brows). The simplest feature set consists of 2D or 3D Cartesian coordinates, although more complex representations have been utilised in related work, such as basis fitting (e.g., polynomial and exponential), angle and distance-based representations. Furthermore, other works [117] employ the parametrisation of shape components as a feature set, which in many cases provides several useful invariant properties. There are many examples of works which utilise geometric features in related work. To name a few, in [187, 188] a set of 20 facial points is used in order to describe facial expressions. The derivation of further features such as angles, distances and velocities has been used in works such as [256, 258]. Specifically, in [256], the features are derived from 58 facial points, aiming to to capture the temporal structure of Action Units, while in [258], 12 fiducial points along with head motion are utilised in order to distinguish posed from spontaneous smiles. In [43], a model-based face tracked is utilised in order to track facial features such as eyebrows, eyelids and mouth, along with head motion, for the analysis of basic emotions. While geometric features are deemed unable to capture particular Action Units (AUs) such as AU28 (inward lip sucking), since the change is only visible in terms of texture, geometric features have been successfully utilised in the analysis of facial expressions in many works [149, 186, 188, 256].

Appearance Features. Appearance features are essentially based on representations of texture information. Therefore, in contrast to geometric features, appearance features can capture skin texture changes such as wrinkles and bulges [112, 197]. There is a multitude of appearance feature sets utilised for affect sensing, including feature extraction via Component Analysis methods such as Principal Component Analysis (PCA), Linear Discriminant Analysis (LDA) and Independent Component Analysis [19, 21]. A set of very commonly used descriptors (e.g., in [53, 295]) include the gradient-based descriptors such as Histogram of Gradient Orientations [54] and the Scale Invariant Feature Transform (SIFT) [147]. An alternative feature set based on the description of pixels relative to their neighbours is the Linear Binary Patterns (LBP) method introduced in [180] and applied in many works pertaining to affect sensing, such as [293, 4] as well being utilised in the provided feature set for the Audio/Visual Emotion Challenge 2013 [257]. Finally, we note that other feature sets which have been greatly applied in affect sensing include the Gabor wavelets and Haar Filters [292, 150, 19, 271]. Each particular feature set comes both with a set of advantages and disadvantages, in most cases determin-

ing the tradeoff between accuracy, complexity and robustness to various transformations and noise. For example, Gabor wavelets [135] essentially involve the modulation of a since wave with a Gaussian envelope at multiple spatial scales, orientations and locations. Good results have been obtained by utilising Gabor filters, while theoretically the process is assumed to be similar to the human visual system [135]. While Gabors are deemed robust to misalignments, they are also deemed very computationally expensive due to the presence of the convolution operator. Also, a redundancy of features is also generated, which in turn is remedied by applying dimensionality reduction, usually via component analysis. Haar filters [271] correspond to more coarse features, being more computationally efficient but less accurate in terms of texture details. This is also an issue with the DCT [1], where texture variation from the frequency domain is utilised. Since the high frequency coefficients are usually discarded, as they are considered to be noise, the DCT may lead to a loss of texture details. Gradient-based feature sets such as the Histograms of Oriented Gradients (HOG) [54] as well the Scale Invariant Feature Transform (SIFT) [147] are based on pixel gradients and are deemed especially robust in terms of varying illumination and scale changes. Finally, Local Binary Patterns (LBP), [180] encode a vector of 8 dimensions for each pixel, describing the pixel's intensity with respect to the neighbouring ones. LBPs are deemed computationally efficient and simple and robust to illumination changes due to the relative description of the pixels intensity, while being less robust to image rotations.

It is not clear whether appearance-based or feature-based extraction is best, since there have been surveys suggesting the better performance of either appearance-based [17] or feature based methods [245, 187, 256]. There have been attempts to produce hybrid systems (e.g. [245, 289]), and it has been suggested that methods which combine the two approaches could provide better results [184]. The main advantage of appearence-based feature sets lies in representing subtle texture changes which can not be detected via geometric features, but it is also a question of whether these changes are vital to a particular task. Geometric features on the other hand are more intuitive as the descriptions can be easily grasped by humans, and also facilitate the more direct modelling of dynamics and facial movement, since the features are essentially spatial coordinates. The further interested reader can refer to [184] for more details regarding feature extraction.

2.5.2 Body and Gesture

There have been many attempts in interpreting and capturing human gestures and body posture, combining techniques from fields such as computer vision and image processing,

mostly targeting Human Machine Interaction (HMI) systems. Specific systems that make use of these capabilities are sign language recognition systems, computer control through gestures, alternative computer interfaces and systems which target emotion recognition.

According to [97], methodologies relating to gesture and body recognition can be separated into three categories:

- **model-based**, which depend on the body or body parts by modelling them or recovering 3D configuration from vision processing.
- **appearance-based**, which base the recognition process on 2D information, e.g. by tracing edges which could form body contours..
- **motion-based**, where the main characteristic tracked is related to motion.

In general, gesture recognition is one of the most difficult tasks in computer vision, due to difficulties commonly appearing in related scenarios (illumination, background/foreground separation, edge tracing, background, occlusions). There is also the issue of separating out irrelevant body motions (which may occur during a proper gesture), determining when a gesture begins manifesting and when it terminates, while also another problem is when a gesture overlaps another.

There is quite a variety of techniques used for tracking, as covered in [65], while an example of a system related to tracing specific features can be found in [178], where the system detects shoulder positions by fitting a parabola to detected horizontal lines in the image and then using the weighted Hough Transform to detect the shape. In [258], head motion is detected with a cylindrical head tracker [277], while a 12 point tracker is used to capture facial features. In order to track shoulder motion, a particle filtering technique is employed.

In general, body gesture recognition requires the calculation of different features, such as the measuring the amount of motion compared to outline changes, hand velocity etc. It is again noted that these methods are optimised for very constraint environments and the development of generic body gesture systems is still an open issue. Relevant extensive surveys on these areas include Yilmaz et al. [282] on general object recognition and specifically vision-based human motion analysis, Mitra and Acharya's [166], specific to hand gestures and facial expressions, and Poppe [201], which surveys modern approaches to vision-based human motion while also discussing theoretical issues of human motion in relation to modelling (e.g. kinematic models, silhouettes, contours). There is also a discussion on the issue of *estimation*, i.e. finding the

set of pose parameters to minimise the observation error in relation to the model (or example set or projection function) used to estimate it. The field has advanced rapidly due to the introduction of easily accessible 3D cameras (or scanners) such as the one utilised by the Microsoft Kinect project [291].

2.5.3 Audio

The optimal set of acoustic features always depends on the particular problem at hand, as well as the inherent characteristics of the dataset employed. In general, commonly employed features include the fundamental frequency (or pitch), as well as the signal energy [285]. A summary of acoustic features in relation to emotion expressions is presented in Table 2.1 (adopted from [49]). Regarding spectral features, the Mel-Frequency Cepstrum Coefficients (MFCCs) are deemed one of the most commonly used feature sets. The mel-frequency bands are equaly spaced on the mel scale[6] and thus are considered to better approximate the response of the human auditory system. Other examples of acoustic features include the voice quality [36], as well as the measurement of pauses and silences [62]. Following the shift towards spontaneous emotion detection, several approaches combined acoustic features and spoken words, while others used linguistic features to improve spontaneous emotion recognition. A notable example of a popular acoustic feature extraction toolkit is described in [76], while further details with respect to acoustic features can be found in [285], [225].

Table 2.1: Sound features in relation to emotional states. Table adopted from [49].

	Anger	Happiness	Sadness	Fear	Disgust
Rate	Slightly faster	Faster or slower	Slightly slower	Much faster	Very much faster
Pitch Average	Very much higher	Much higher	Slightly lower	Very much higher	Very much lower
Pitch Range	Much wider	Much wider	Slightly narrower	Much wider	Slightly wider
Intensity	Higher	Higher	Lower	Normal	Lower
Voice Quality	Breathy, chest	Breathy, blaring tone	Resonant	Irregular voicing	Grumble chest tone
Pitch Changes	Abtrupt on stressed	Smooth, upward	Downward in ections	Normal	Wide, downward
Aritculation	Tense	Normal	Slurring	Precise	Normal

2.6 Databases

An important problem that researchers in this field are often confronted with is the proper acquisition and labelling of data. We have already referred to the problem of determining spontaneous vs. posed data (Section 2.2) and in general, the long-term goal of realising systems which perform automatic spontaneous emotion recognition. In fact, strictly speaking

[6] The mel scale is defined as a scale of pitches which are judged to be equi-distant from one another by human listeners.

the available databases can be separated into the following, depending on the setting of the recordings.

- **Posed,** where the participants are requested to produce the affective state on demand, usually in laboratory settings.
- **Induced,** where the experiment takes place in controlled environments which are designed in order to induce the affective states, e.g. by projecting video clips to the participants or capturing human-to-human or human-to-machine interaction [14].
- **Spontaneous,** as in occurring in real-life settings, e.g. in naturalistic human to human communication.

Recording the subjects in such databases requires the use of cameras for facial and body expressions and microphones for recording the audio signals, while often motion capture systems are used to record 3D postures and gestures. Ideally, these sensors should be minimally intrusive to the actual recording process in order to minimise the effect on the subjects behaviour. Issues relate to variant noise levels in the audio signal as well as various occlusions, e.g., of the face by various equipment or body parts.

Most existing affective databases contain *posed* data, where the expressions exhibited by the participants follow the neutral-onset-apex-offset-neutral transition of facial expressions. This is due to the fact that posed data are easier to squire than spontaneous or induced, while there are many difficulties in terms of capturing spontaneous manifestations, as they are more difficult to elicit or capture, they are more influenced by the context and therefore more difficult to analyse and track, are more noisy (e.g., more occlusions by body parts, different angles and distances from the camera etc.) while even the annotation process (labelling) of the data becomes more difficult.

Due to the rising interest in detecting spontaneous emotions there have been attempts to generate databases of spontaneous emotions. While typically, the basic emotions are used for categorisation in posed databases, spontaneous databases often use the may utilise more descriptive approaches, such as dimensional emotions, including dimensions such as valence and arousal. A database which contains both spontaneous and posed data is the MMI Database [189], considered to be one of the most comprehensive set of facial behaviour recordings, providing both images and videos depicting frontal and profile views. It includes more than 1500 samples, while the samples are encoded utilising the FACS system.

In this thesis, we mostly utilise databases annotated in terms of continuous emotion dimensions, namely the Sensitive Artificial Listener (SAL), as well as the Sustained Emotionally Coloured Machine-human Interaction using Nonverbal Expression (SEMAINE) databases. SAL is essentially the first database which adopts a human-to-human interactive scenario with a goal of eliciting spontaneous emotions, while also adopting continuous and dimensional emotion annotations. SAL has been superseded by SEMAINE, which offers a similar scenario while offering various advantages, including the improvement of annotation quality and quantity, better input device quality with a reduction of noise, as well as more subjects and sessions. We discuss more regarding these databases in what follows.

The SAL database

The Sensitive Artificial Listener Database (SAL-DB) [64] contains audio-visual, naturalistic affective conversational data taking place between a participant and an avatar (operated by a human): Poppy (happy), Obadiah (gloomy), Spike (angry) and Prudence (pragmatic). Each avatar is considered to have a different personality: Poppy is happy, Obadiah is gloomy, Spike is angry and Prudence is pragmatic.

Figure 2.7: Stills from the SAL database, where sessions involving the above subjects have been annotated in the valence-arousal space.

The audiovisual sequences have been recorded at a video rate of 25 fps (352 x 288 pixels) and at an audio rate of 16 kHz. The recordings were made in a controlled laboratory setting, using one camera, a uniform background and constant lighting conditions. The SAL data has been annotated by a set of annotators who provided continuous annotations with respect to valence and arousal dimensions using the FeelTrace annotation tool [48]. Feeltrace allows annotators to watch the audiovisual recordings and move their cursor, within the 2-dimensional emotion space (valence and arousal) confined to $[-1, 1]$, to rate their impression about the emotional

state of the subject. Although there are approximately 10 hours of footage available in the SAL database, V-A annotations have only been obtained for two female and two male subjects. This is the portion of data we utilise throughout the experiments on SAL in this thesis (Chapters 5, 6). Example frames from the SAL database are shown in Fig. 2.7.

The SEMAINE database

The SEMAINE (Sustained Emotionally Coloured Machine-human Interaction using Nonverbal Expression) database [157], contains a set of audio-visual recordings focusing on dyadic interaction scenarios, similarly to SAL. The recording scenario is similar to SAL, with the adoption of HD video and a smoother frame rate (50 frames per second) in SEMAINE. The dyadic interaction scenarios consist of a human subject, conversing with an operator, who assumes the role of an avatar. Each operator assumes a specific personality, which is defined by the avatar he undertakes: happy, gloomy, angry or pragmatic. This is in order to elicit spontaneous emotional reactions by the subject that is conversing with the operator. As discussed in Section 2.1, SEMAINE has been annotated in terms of several emotion dimensions, particularly in terms of valence, arousal (activation), power, expectation (anticipation) and intensity. Stills from the SEMAINE database are shown in Fig. 2.8.

2.7 Continuous Annotations: Obtaining the Ground Truth

As discussed in the introduction of this thesis, obtaining annotations continuously in time is a tedious and error-prone task, leading to many open challenges. In this section, we summarise the aforementioned set of issues with respect to the database described above, namely SAL and SEMAINE. The typical annotation tool which has been employed in both SAL and SEMAINE is the Feeltrace tool [48], which allows the affective state of the individual to be evaluated in terms of dimensions such as valence and arousal. In the case of audio-visual recordings, the annotator which is responsible for the annotation observes and listens to the recording. The annotator moves the mouse and in effect the cursor indicating the current annotation. The annotation is usually performed real-time and later normalised from -1 to 1. The agreement of annotators with regards to the mapping of the observed emotional stimulus in to a dimensional space is difficult to achieve. Problems adopting labels related to emotions carry an inherent issue of *label subjectivity*. When measuring quantities such as subject *interest* or emotion dimensions such as *valence*, it is natural for some ambiguity to arise, especially when utilising spontaneous data in naturalistic, interactive scenarios. This is essentially the trade-off between capturing a larger spectrum of expressions, and minimising the space in order to reduce label

ambiguity. Furthermore, the on-line nature of the annotation process renders the resulting annotations vulnerable to various temporal lags which depend on the response time of the annotator. In more detail, the annotator has to first interpret the emotional state observed, subsequently map it to the emotion dimension annotated, and then perform a movement with an input device (here, mouse) in order to reflect his understanding of the emotional state of the subject. As can be understood, this leads to a temporal lag in the annotation with respect to the video itself, which is dependent on many parameters such as the complexity of the emotion being portrayed, a set of annotator specific human factors as well as any extra effort required by the input device. Clearly, the task of obtaining a "gold standard" (i.e., the true

Figure 2.8: Frames grabbed simultaneously from the five video streams offered in SEMAINE. The operator appears on the left, while the user on the right. The image has been adapted from [157].

annotation, given a set of possibly noisy annotations) is it is clear that the task of obtaining a "gold standard" (i.e. the "true" annotation, given a set of possibly noisy annotations), is a quite tedious task, and researchers in the field have not been agnostic regarding this in previous work [164]. In the majority of past research related to affect sensing though, usually a form of averaging is employed for this task, assuming that the true annotations is represented by a simple average of the multiple annotations [274], or utilising weighted averages, e.g., by the correlations of each annotator to the rest ([174], Chapter 5). Nevertheless, majority voting (for discrete labels) or averaging (for continuous in space annotations) makes a set of explicit assumptions, namely that all annotators are *equally* good, and that the majority of the annotators *will* identify the correct label eliminating any ambiguity/subjectivity. Nevertheless, in most in real-world problems these assumptions typically do not hold. As we discussed, even in the case of SEMAINE and SAL, where the annotators are trained experts, they are not infallible when it comes to a *subjective* process which incorporates all the pitfalls discussed above, indicating the existence of a strong spatio-temporal bias. On top of that, in many cases though, annotators can be inexperienced, naive or even uninterested in the annotation task. This phenomenon has been amplified by the recent trend of *crowdsourcing* annotations (via services such as Mechanical Turk), which allows gathering labels from large groups of users, who usually have no formal training in the task-at-hand, shifting the annotation processes from

a small group of experts to a massive but weak-annotator scale. In general, besides experts, we can consider that annotators can be assigned to classes such as *naive* which commonly make mistakes, *adversarial* or *malicious* annotators, that provide erroneous annotations on purpose, or *spammers* that do not even pay attention at the sequence they are annotating. It should be clear that if e.g., the majority of annotators are adversarial then majority voting will always obtain the wrong label. This is also the case if the majority of annotators are *naive*, and on a difficult/subjective data all make the same mistake. This phenomenon led to particular interest manifesting in modelling annotator performance, c.f.[208, 209]. Note that due to the discussed temporal lag exhibited by the annotators, simply averaging the annotations without eliminating temporal discrepancies is very likely to lead to both *phase* and *magnitude* errors (such as false peaks). We clarify here, that temporal lags depending on annotator response times are always "later" in time (i.e. are positive temporal shifts). In effect, this means that if we adopt simple averaging, there will always be a misalignment between the annotation and the sequence-at-hand.

The idea of shifting the annotations in time in order to attain maximal agreement has been touched upon in [173] and later in [152]. Nevertheless, these works refer to a constant time-shift, which assumes that the annotator-lag is constant. This does not appear to be the case, as the annotator-lag depends on time-varying conditions. Note that in Chapter 9, we present a novel probabilistic model aiming to resolve such temporal errors in the annotations.

Finally, we discuss the method proposed in [208] towards the fusion of multiple annotations and labels. In this work, an attempt is made to model the performance of annotators, who assign a possibly noisy label. The latent "true" (binary) annotation is not known, and should be discovered in the estimation process. By assuming independence of all annotators and furthermore, assuming that annotator performance does not intrinsically depend on the annotated sample, each annotator can be characterised by his/her sensitivity and specificity. In this naive Bayes scenario, the annotator scores are essentially used as weights for a weighted majority rule, where if all annotators have the same annotator characteristics it collapses to the majority rule[7]. Note that the more general approach of [208] indicates that in the absence of a gold standard, neither simple nor weighted majority voting is optimal. In fact majority voting can be seen only as a first guess aimed at assigning an uncertain consensus "true" label, which is then further refined using an iterative Expectation Maximisation (EM) process, where both the "true" label and the annotator performance are recursively estimated.

[7]Detailed analysis of majority voting, including its weighted version, can be found in [128, 218].

2.8 Conclusions

In this section, we provided a thorough examination of the background which relates to affect sensing and analysis, covering the general directions employed by this thesis, as well as referring to various related work. In what follows, we briefly summarise the relationship of the aspects covered in this chapter with respect to the work presented in this thesis. In more detail, in the first part of this thesis (Chapters 5, 6 and 7), we focus on presenting a set of methodologies aiming at learning continuous emotion dimensions by further utilising relationships amongst the output dimensions. We also present novel methodologies which utilise the fusion of multiple modalities (including facial expressions, shoulder gestures and audio cues), as well as provide an empirical analysis to the problems which arise from utilising continuous emotion dimensions. In the second part of the thesis, we firstly focus on the problem of fusing multiple continuous annotations (Chapter 9), and propose an approach which aims to deal with the multitude of problems arising in this scenario. In Chapter 10, a robust, multi-modal fusion technique is proposed, which is evaluated in terms of predicting continuous interest. In Chapter 11 we propose a unifying framework for probabilistic component analysis, giving rise to many methods which can be applied for feature extraction in affect sensing. In the same chapter, we apply the proposed EM-LDA to the problem of interest classification.

Chapter 3

Learning Techniques

Contents

3.1 Introduction

In this chapter, we refer to a set of machine learning techniques which are closely to the content of this thesis. Firstly, in Section 3.1.1 we discuss the issue of supervised and unsupervised learning, as well as refer to generative and discriminative models. In Section 3.1.2 we provide a high level introduction to the methods described, including regression and component analysis. In Section 3.2, we discuss methods based on Recurrent Neural Networks (Section 3.2.1) and Bayesian Regression (Section 3.2.2), while we briefly refer to Support Vector Regression (SVR) in Section 3.2.3. Subsequently, in Section 3.3, we shift to component analysis and detail various, commonly employed, component analysis methods. In Section 3.4, we discuss time warping (temporal alignment) and provide a connection to component analysis, while finally, in Section 3.5 we conclude the chapter.

3.1.1 Supervised vs. Unsupervised Learning

As in all learning problems, machine learning problems tangent to automatic behaviour analysis consist of a set of observations (features) and in many cases, a set of labels (annotations).

In case a set of labels is available, the goal is to learn a mapping from the feature space to the labels (e.g., learn a mapping from a face image to the emotion expressed by the image). Otherwise, the goal becomes the extraction of a subspace from the original observation space which preserves particular desired properties of the signal (e.g., given a set of images (observations), remove the features which characterise the identity of the subject and keep only the features which pertain to the expression of the subject).

The above setting also determines, to a large degree, the type of learning method employed for a particular task. In general, a learning problem can be approached either by *supervised* or *unsupervised* learning. In case annotations (labels) are available, one can resort to the so-called *supervised* learning methods. This implies that for a given problem, a set of annotations has been obtained either manually or automatically. In automatic behaviour analysis, the typical case is that the annotations have been manually annotated - a costly task, as we discuss throughout in this thesis. Usually, supervised learning leads to the adoption of *discriminative* learning methods, which model the conditional distribution of the labels given the observations. In many cases, this has been shown to be beneficial in terms of classification accuracy, since this distribution is exactly what is required in order to classify. Nevertheless, this comes at a sacrifice of model flexibility. In case no labels are available, one has to resort to unsupervised learning techniques in order to extract *interesting* information from the observations. Unsupervised learning is highly affiliated with generative models, which focus on modelling joint distributions (instead of conditional as for discriminative). In general though, it should be clarified that discriminative methods can be extended for unsupervised tasks, and generative models can be extended to supervised or semi-supervised scenarios. In what follows, we discuss regression and component analysis, with regression being a predictive analysis method which is inherently supervised and discriminative, and component analysis, where methods are inherently generative. Furthermore, Component Analysis methods can be both supervised and unsupervised depending on the constraints that are imposed, e.g., if the constraints are simply to maximise the variance of the data then no labels are needed; if while doing so we are required to conform to class label constraints, then the method becomes supervised.

3.1.2 Regression and Component Analysis

In this section, we will introduce regression and component analysis more formally. Interestingly enough, an incredible amount of research in machine learning over many decades is based on a seemingly simple linear equation:

$$\boldsymbol{\chi} = \mathbf{w}^T \boldsymbol{\psi}. \tag{3.1}$$

Essentially a deterministic linear mapping, the goal of a problem utilising such a mapping is only defined when both specific interpretations are assigned to each of the random variables $\boldsymbol{\chi}$ and $\boldsymbol{\psi}$, as well as the desired behaviour of $\mathbf{w}$ is specified. In general, in machine learning scenarios we have one or more set of observations, or input data. These essentially represent *any* information we have with regards to the problem, and usually they will also be available during testing, where we have already trained a system and deployed it in the target application domain. This information can be e.g., in the form of features (or observations), which are usually extracted via the procedures detailed in Section 2.5. Furthermore, in predictive analysis scenarios (regression, classification), during training we also have a set of class labels or outputs (annotations), which essentially encapsulate a form of class or target value which may correspond to each training observation sample (in continuous scenarios), or can carry an entire value for an entire segment or sequence. These labels essentially represent the targets of the linear function presented in Eq. 3.1 (in case they are continuous). In other words, we aim to learn a function mapping from the *observations* to the *labels*, or put simply, from the *inputs* to the *outputs*. Once this function is learned, the inputs should enable the accurate prediction of the outputs. Adopting the aforementioned scenario, let us assume we are given observations $\mathbf{x}_i$ and target values for each observation, $\mathbf{y}_i$. In this regression setting, Eq. 3.1 becomes

$$\begin{aligned} \mathbf{y}_i &= \mathbf{w}^T \mathbf{x}_i \\ \text{input: } & \mathbf{y}_i, \mathbf{x}_i. \end{aligned} \tag{3.2}$$

essentially meaning that one wants to obtain the best $\mathbf{w}$ which map the inputs $\mathbf{x}_i$ as close as possible to the given outputs $\mathbf{y}_i$. Having learnt the correct $\mathbf{w}$, is the only requirement for predicting $\mathbf{y}^*$ given a test datum, $\mathbf{x}_i^*$. Interestingly enough, most of the state-of-the-art predictive analysis techniques employed in modern research and industry are based on optimising this simple functional, ranging from simple linear regression to the Relevance Vector Machine (RVM) [246] as well as the Support Vector Machine [66].

Regression and classification though are not the only techniques which are based on learning a simple mapping. Component Analysis (CA), a significant branch of statistics and machine learning, consists of a set of techniques which aim at factorising a given signal in a manner which facilitates an employed task, e.g. clustering or even regression and classification. The differentiating factor in *unsupervised* CA, is that essentially there are no target labels to learn a mapping to[1]. CA techniques essentially infer a latent, unobserved space which satisfies a particular set of desired properties, with the most common being Principal Component Analysis

[1]While label information may be available and component analysis techniques can be easily adapted in

(PCA), which essentially recovers a parsimonious explanation of the observations, providing a latent space which maintains the variability of the input features while decorrelating them and usually reducing their dimensionality. As described, in CA we usually have a set of observations ($\mathbf{x}_i$), and aim to discover the latent space which satisfies the desired properties (in the form of constraints), $\mathbf{y}_i$ with $\mathbf{w}$ again being the loadings. Eq. 3.1 can now be expressed as

$$\begin{aligned} \mathbf{y}_i &= \mathbf{w}^T \mathbf{x}_i \\ \text{input: } &\mathbf{x}_i. \end{aligned} \tag{3.3}$$

where the actual loadings are commonly found by formulating a trace-optimisation or least-squares problem under particular constraints. Examples of such formulations will be discussed in more depth in Section 3.3.

3.1.3 Non-linear Mappings

While, as aforementioned, the linear mapping is a basic functional commonly employed, in many scenarios features need to be mapped in higher dimensional spaces: this is because the data in their current form are simply not linearly separable. Typically, this is performed by utilising the kernel trick in regression scenarios, that is, by estimating an implicit feature space without actually estimating the coordinates of the data in the feature space, but rather simply computing the inter products between the images of all pairs of data. This makes the projection in many cases feasible, and also computationally efficient. E.g., in the RVM which employs the regression linear mapping (Eq. 3.2), we simply need to replace $\mathbf{x}_i$ for $\phi(\mathbf{x}_i)$, leading to the mapping

$$\mathbf{y}_i = \mathbf{w}^T \phi(\mathbf{x}_i) \tag{3.4}$$

where $\phi(\mathbf{x}_i) = [K(\mathbf{x}_i, \mathbf{x}_1), K(\mathbf{x}_i, \mathbf{x}_2), \ldots, K(\mathbf{x}_i, \mathbf{x}_N)]$, with N being the number of samples and K a non-linear function (kernel), such as e.g., the Radial Basis Function (RBF), defined as

$$K(\mathbf{x}_i, \mathbf{x}_j) = \exp\left\{-\frac{||\mathbf{x}_i - \mathbf{x}_j||}{l}\right\}. \tag{3.5}$$

with l being the length scale. This process is similar in SVM and other non-linear regression techniques and many kernel types may be utilised [262]. Summarising, this allows us to simply pre-compute the kernels between all pairs of data and still apply a linear method.

many cases to account for labels (i.e. supervised component analysis), the methods do not map the features *to* the labels but rather utilise the labels for optimally recovering the projections.

3.1.4 Handling Noise

Another typical desired characteristic in machine learning is resilience to noise. This is a common situation encountered in almost all learning scenarios, since data is nearly never perfect. The typically assumed form of noise is Gaussian. In probabilistic scenarios, as e.g., in RVM or Probabilistic PCA, this can be easily added to the model functional, by incorporating a noise term ϵ. E.g., in RVM, the regression functional defined in Eq. 3.2 is extended as

$$\begin{aligned} \mathbf{y}_i &= \mathbf{w}^T \mathbf{x}_i + \epsilon_i \\ \epsilon &\sim \mathcal{N}(0, \sigma^2), \end{aligned} \tag{3.6}$$

where ϵ represents the noise as independent samples from a zero-mean Gaussian noise process with variance σ^2. Interestingly enough, if one takes the Maximum Likelihood (ML) solution of Eq. 3.6, one easily finds that the actual term being minimised (up to a constant) is the least-squares penalty,

$$\sum_{i=1}^{N} (\mathbf{y}_i - \mathbf{w}^T \mathbf{x}_i). \tag{3.7}$$

In effect, this shows that least squares estimates are actually equivalent to producing the maximum likelihood solution of Eq. 3.6, where the parameters and variables are linearly related up to Gaussian noise.

3.2 Related Regression Techniques

In this section, we describe in more detail a set of related methods which are utilised in the thesis. Firstly, in Section 3.2.1 we describe Recurrent Neural Networks (RNNs), and in more detail one of the most recent reincarnations, Long Short-Term Memory Neural Networks (LSTM-NNs). LSTMs are actually the first RNN variant being able to model long range temporal dependencies, a crucial aspect in terms of analysing the inherently temporal characteristics of human behaviour (as discussed in Chapter 2). Nevertheless, neural networks have been heavily criticised in the past decades, mostly due to (i) the lack of efficient training algorithms and (ii) the inherent lack of model interpretability; a mapping was learnt, but this provided no information regarding the relative importance of the data as well as no information regarding uncertainty of predictions. The first issue, that of efficient training, was recently resolved via the introduction of Hinton's contrastive divergence method, significantly speeding up the learning procedure, as well as the increase of computational capabilities of modern computers and the utilisation of GPUs in training. Nevertheless, the second issue

of interpretability remains, while many of the algorithms employed are based on empirical evaluations and are not theoretically justified in terms of e.g., convergence or even the actual approximation targets. In effect, this has led many researchers to characterise deep neural networks as simply powerful empirical feature-extractors (often utilised as a "black-box"), providing just a single step in the design of a large complex system. This leads to the second model we discuss. The Relevance Vector Machine (RVM), detailed in Section 3.2.2 is a very popular probabilistic regression technique (or more accurately, a Sparse Bayesian Regression technique), which infers probabilistic distributions of datums utilising Bayesian Regression in a fast and robust manner. RVM utilises only the set of data which are highly relevant to the output datums, while providing parsimonious explanations of the data at hand.

3.2.1 Recurrent to Long Short-Term Memory Neural Networks

Recurrent Neural Networks

Recurrent Neural Networks (RNN) are significant tools in the analysis of time series. While traditional feedforward neural networks are allowed to only have forward connections (i.e. from the input to the output), recurrent neural networks also employ feedback connections, thus permitting the formation of cycles and loops. This adjustment facilitates the adaptation of RNNs to past inputs, therefore incorporating temporal dependencies in the learning procedure, thus enabling the analysis of temporally enriched sequences.

In mode detail, assuming that we have a regular feedforward network, given an input x at time t, the network learns the following mapping:

$$y(t) = \mathcal{F}(x(t)) \tag{3.8}$$

That is, the network, which has an internal configuration consisting of weights on connections between neurons along with the family of activation functions used, will map the input $x(t)$ at any time t to the output $y(t)$. It is important to stress that the output depends only on the current configuration and input. On the other hand, a recurrent network can operate on an internal state space, which ideally contains all relevant information from the past behaviour of the system. This extends the network capabilities by allowing it to capture temporal information and exploiting them during learning. Thus, the recurrent network's output at time t, $y(t)$ would be a function of the current state of the network $s(t)$, which in turn depends on the previous state $s(t-1)$ and the current input $x(t)$:

$$y(t) = \mathcal{F}'(s(t)) \tag{3.9}$$

$$s(t) = \mathcal{G}'(s(t-1), x(t)) \tag{3.10}$$

It is interesting to observe that this Markovian-like dependencies expressed in the above equations are very similar to typical Linear Dynamical Systems, only we essentially have a neural network function instead of a simple linear mapping. To contrast the computational power of recurrent neural networks in comparison to regular feedforward networks, it is enough to say the following: while a feedforward network, given enough hidden nodes can approximate any spatially finite function, recurrent neural networks (again assuming any number of hidden nodes) can represent any Turing Machine [102], while if real weights are used, the network can function as a super-Turing Machine [234], notions which are much more powerful than approximating finite functions.

In this section, we will refer to a neural network with one hidden layer, the input layer and the output layer. For referring to a node in the hidden or output layer, the subscripts h and o will be used respectively. We consider the input to have a size of n, while we consider m nodes in the hidden layer and m nodes in the output layer. The activation of a neuron belonging to the hidden layer of such a feedforward network will have an activation value $y_h(t)$:

$$y_h(t) = \sigma(net_h(t)) \tag{3.11}$$

$$net_h(t) = \sum_{i}^{n} x_i(t) w_{hi} + \beta_h \tag{3.12}$$

That is, the output is the *net* input to the neuron applied to the activation function σ (typically a non-linear such as the logistic function). The *net* input to the hidden node is the sum of the weights coming to node h from each input i (the input vector $\mathbf{x}$ has a size of n), while β is the bias of node h.

Assume a simple recurrent network, where besides the feedforward connections, the nodes of the hidden layers have one step delay feedback connections, that is the previous activation of the nodes in that layer is taken into account. Since there are more connections, a new set of weights v_{ij} is required. Again looking at the activation of a node in the hidden layer, $y_h(t)$, Equation 3.11 remains the same. What changes is the $net_h(t)$:

$$net_h(t) = \sum_{i}^{n} x_i(t) w_{hi} + \sum_{j}^{m} y_j(t-1) v_{hj} + \beta_h \tag{3.13}$$

where m is the number of nodes which have the feedback connection to node h and $y_j(t-1)$ is the previous activation of each of them. In the example presented in the section, we stated

that feedback loops occur only in the hidden layer, so the equations for the output nodes of the network are the same as the feedforward networks:

$$y_o(t) = \sigma(net_o(t)) \tag{3.14}$$

$$net_o = \sum_{j}^{m} y_j(t) u_{oj} + \beta_o \tag{3.15}$$

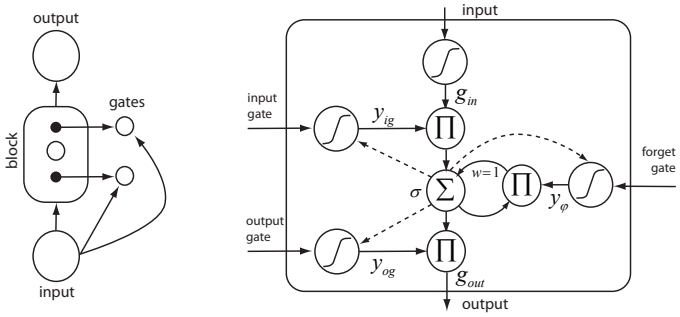
Where u_{oj} are weights from the hidden nodes j to the output node o and again, σ is the activation function and β_o the bias of the output node.

There is a vast amount of literature concerning recurrent neural networks, e.g., state-space models where the previous activation of the hidden layer is considered part of the next input, input-output recurrent models where the actual output of the network is being fed back to the input, recurrent multilayer perceptrons where each computation layer has a feedback, and second order networks where the previous outputs are actually multiplied. For more details regarding recurrent neural networks along with optimisation details (utilising Back-Propagation Through Time) we refer the reader to [125] and [113].

Long Short-Term Memory Neural Networks

One of the most significant issues when utilising RNNs was the apparent inability to model temporal dependencies longer than a few time steps away due to to the so-called vanishing-gradient problem [106]. Essentially, the problem refers to the inability of conventional training algorithms for RNNs to keep the error signals which are flowing backwards in time from either vanishing exponentially or increasing exponentially, leading to an inherent inability to model long range dependencies. This has been shown extensively in Hochreiter's analysis [105] while also discussed in [23, 106]. To this end, the LSTM Neural Networks (LSTM-NNs) were introduced by Graves and Schmidhuber [89] to overcome this issue. Essentially, the LSTMs are the most recent incarnation of RNNs before the rise to prominence of the recent "Deep Learning" trend.

LSTMs introduce recurrently connected memory blocks instead of traditional neural network nodes, which contain memory cells and a set of multiplicative gates. The gates essentially allow the network to learn when to maintain, replace or reset the state of each cell. As a result, the network can learn when to store or relate to context information over long periods of time, while the application of non-linear functions (similar to transfer functions in traditional NN) enables learning *non-linear dependencies*



(a) (b)

Figure 3.1: Illustration of (a) the simplest LSTM network, with a single input, a single output, and a single memory block in place of the hidden unit, and (b) a typical implementation of an LSTM block, with multiplication units (Π), an addition unit (Σ) maintaining the cell state and typically non-linear squashing function units.

In more detail, in Fig. 3.1b, the three types of gates are visualised: the input, output and forget gates. As aforementioned, they can be thought of as providing write, read and reset access to what is called a cell state (σ), which represents temporal network information. This can be seen from examining the state updates at time t:

$$\sigma(t) = y_{\phi}(t)\sigma(t-1) + y_{ig}(t)g_{in}(t)$$

The next state $\sigma(t)$ is defined as the sum of the forget gate at time t ($y_{\phi}(t)$) multiplied by the previous state, $\sigma(t-1)$ and the squashed input to the cell $g_{in}(t)$ multiplied by the input gate $y_{ig}(t)$. Thus, the forget gate can reset the state of the network, i.e. when $y_{\phi} \approx 0$ then the next state does not depend on the previous one:

$$\sigma(t) \approx y_{ig}(t)g_{in}(t)$$

This is similar when the input gate is near zero. Then, the next state depends only on the previous state and the forget gate value. The output of the cell is is the cell state, as regulated by the value of the output gate (Fig. 3.1b). This configuration enforces constant error flow and overcomes the vanishing gradient problem.

Bidirectional LSTMs

In addition, traditional RNNs process input in a temporal order, thus learning input patterns by relating only to past context. Bidirectional RNNs (BRNNs) [230, 12] instead modify the learning procedure to overcome the latter issue of the past and future context: they present each of the training sequences in a forward and a backward order (to two different recurrent

networks, respectively, which are connected to a common output layer). In this way, the BRNN is aware of both future and past events in relation to the current timestep. The concept is directly expanded for LSTMs, referred to as Bidirectional Long Short-Term Memory neural networks (BLSTM-NN). BLSTM-NN have been shown to outperform unidirectional LSTM-NN for speech processing (e.g., [89]) and have been used for many learning tasks. They have been successfully applied to continuous emotion prediction from speech (e.g., [151], [275]) proving that modelling the sequential inputs and long range temporal dependencies appear to be beneficial for the task of automatic emotion prediction.

3.2.2 Relevance Vector Machine

The Relevance Vector Machine (RVM), introduced by Tipping in [246] is a Bayesian regression technique, aimed at providing parsimonious, probabilistic solutions for regression and classification. In more detail, we assume a regression problem with N training examples, $(\mathbf{x_i}, t_i)$. As briefly mentioned in Section 3.1.2, within the Bayesian framework applied in RVM, our goal is to learn the functional

$$t_i = \mathbf{w}^T \phi(\mathbf{x_i}) + \epsilon_i \tag{3.16}$$

where the ϵ_i are assumed to be independent Gaussian samples with zero mean and σ^2 variance, $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$, and ϕ is a typically non-linear projection of the input features, $\mathbf{x_i}$. The method infers the set of weights $\mathbf{w}$ along with the noise estimation, given the training data. In general, in most regression techniques one wishes to penalise the growth of the weights $\mathbf{w}$ in order to constrain the complexity of the inferred function and thus obtain more parsimonious solutions. In deterministic scenarios, i.e. in SVM, this can be employed by e.g. $l2$-regularisation, by directly penalising the norm of the weights, i.e. $||\mathbf{w}||_2$. In a probabilistic scenario as in RVM, this is performed by utilising *prior* probability distributions on $\mathbf{w}$, thus expressing our preference for smoother and less complex functions. Specifically in RVM, the weight prior is defined by utilising a zero-mean Gaussian distribution

$$P(\mathbf{w}|\alpha_i) = \mathcal{N}(0, \alpha_i^{-1}), \tag{3.17}$$

where α_i describes the precision (i.e. inverse variance) of each weight, w_i. In effect, this controls the strength of the prior individually for each weight, since essentially the prior is data dependent. An important property of the RVM is that the α_i hyperpriors are *hierarchical*,

i.e. a set of scale parameters in the form of Gamma distributions are employed:

$$P(\alpha) = \prod_{i=1}^{N} \text{Gamma}(\alpha_i|\gamma_{\hat{\theta}}) \tag{3.18}$$

$$P(\sigma^{-2}) = \text{Gamma}(\sigma^{-2}|\gamma_\theta) \tag{3.19}$$

where the γ_θ stand for the parameters of the gamma distribution. By setting these parameters to small values, uniform hyperpriors are obtained. An advantage of adopting these "improper priors" lies in the provided scale-invariance, since all scales are equally likely. Furthermore, these priors are essentially a form of *automatic relevance determination* priors. Put simply, these broad priors over the hyperparameters allow for the posterior mass to concentrate at very large values of α_i and thus sending the weight posterior to zero. This essentially defines the sparse properties of the model: the weights for specific data which is deemed unnecessary will be sent to zero, thus ignoring the sample and being able to learn simpler, less complex models. It is interesting to study the sparseness of RVM a bit further. If we consider the distribution of w when marginalising out the hyperparameters, i.e. the α, we have

$$P(w_i) = \int P(w_i|\alpha_i)p(\alpha_i)d\alpha_i. \tag{3.20}$$

This in fact results in a *Student-t* distribution, thus justifying the sparseness properties. This is due to compounded a Gaussian distribution with an unknown variance following an inverse gamma distribution, which has been subsequently marginalised out. This is illustrated in Fig. 3.2. Finally, we note that the sparse property of RVM, along with the existence of fast,

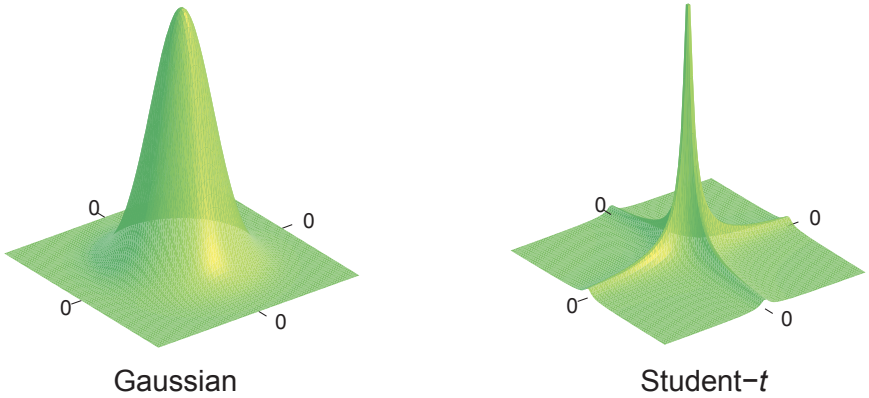


Figure 3.2: Comparing a two dimensional Gaussian prior with a two dimensional Student-t prior. The probability mass is concentrated at the origin and along the spines, where one of the weights is zero. Image adapted from [246].

computationally efficient and incremental methods for learning [249], deem RVM a suitable

model for processing large amounts of data under realistic conditions, where a large amount of this data may be corrupted by noise. We will discuss more on extending RVM in Chapter 6, while we refer the interested reader to [246, 249], Chapter 7 of [26] and Chapter 13 of [168] for more details.

3.2.3 Support Vector Machines for Regression

In this section, we briefly summarise a commonly employed technique, Support Vector Machines for Regression [66] (SVR). In SVR, a non-linear function (conceptually similar to RVM) is optimised by the model, in a mapped feature space, induced by the kernel used (as discussed in Section 3.1.2). An important advantage of SVMs is the convex optimization function employed which guarantees that the optimal solution is found. The goal is to optimize the generalization bounds for regression by a loss function which is used to weight the actual error of the point with respect to the distance from the correct prediction. To this aim, various loss functions maybe employed (e.g., quadratic loss function, Laplacian loss function, and ϵ-insensitive loss function). The ϵ-insensitive loss function, introduced by Vapnik, is an approximation of the Huber loss function and enables a more reliable generalization bound [50]. This is due to the fact that unlike the Huber and quadratic loss functions (where all the data may be support vectors), utilising an ϵ-insensitive loss function leads to a sparse selection of support vectors. Sparse data representations have been shown to reduce the generalization error [264] (see Chapter 3.3 of [222] for details). Finally, SVM is commonly used in related work on predicting continuous affect (e.g., [151, 91, 116]).

3.3 Component Analysis

A major part of this thesis is based on Component Analysis (CA), a set of statistical techniques aimed at factorising observations into components, based on certain constraints which capture desirable properties of the resulting spaces. As mentioned earlier, CA constitutes an important step in systems tangent to computer vision and machine learning. The roots of CA can be traced back to 1901, with the introduction of Principal Component Analysis (PCA) by Karl Pearson [193]. PCA was later developed independently in 1933 [108] by Hotelling, three years before Hotelling introduced Canonical Correlation Analysis (CCA) [109]. While the main goal of PCA is to identify the principal directions of maximal variance of a set of observations[2] CCA generalises this to two observation sets, by finding the projection directions

[2]The actual definition of PCA as initially posed by Pearson [193] was defined as the linear projection minimising the average projection cost, which is defined as the mean squared distance between points and

under which the sets are maximally correlated. CCA has essentially risen from the need to study multiple observation sets, and led to other significant work in the following decades, such as Tucker's Inter-Battery Factor Analysis (IBFA) [253]. In what follows, we summarise two basic component analysis techniques, PCA and CCA, in order to facilitate discussions in later chapters. Throughout this description we consider, without any loss of generality, a zero mean set of F-dimensional observations of length T, $\{\mathbf{x}_1, \ldots, \mathbf{x}_T\}$, represented by a matrix $\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_T]$. All CA methods discover an N-dimensional latent space $\mathbf{Y} = [\mathbf{y}_1, \ldots, \mathbf{y}_T]$ which preserves certain properties of $\mathbf{X}$.

3.3.1 Principal Component Analysis (PCA)

PCA discovers a lower dimensionality space (the principal subspace), where the variance of the observations is maximised. The deterministic model of PCA finds a set of projection bases $\mathbf{W}$, with the latent space $\mathbf{Y}$ being the projection of the training set $\mathbf{X}$ (i.e., $\mathbf{Y} = \mathbf{W}^T\mathbf{X}$)). Since we aim to maximise the projected variance, the optimization problem can be defined as

$$\mathbf{W}_o = \arg\max_{\mathbf{W}} \operatorname{tr}\left[\mathbf{W}^T\mathbf{S}\mathbf{W}\right], \text{ s.t. } \mathbf{W}^T\mathbf{W} = \mathbf{I} \tag{3.21}$$

where $\mathbf{S} = \frac{1}{T}\sum_{i=1}^{T}\mathbf{x}_i\mathbf{x}_i^T$ is the total scatter matrix and $\mathbf{I}$ the identity matrix. One can alternatively arrive at the same optimisation problem by formulating the analogous problem of minimising the reconstruction error and end up in the same algorithm. The above trace is maximised by setting $\mathbf{W}$ to the N projection basis corresponding to the N eigenvectors of $\mathbf{S}$ corresponding to the largest N eigenvalues.

PCA has also been studied in terms of probabilistic formulations. In more detail, approaches towards Probabilistic PCA (PPCA) were proposed independently in [211] and [248]. In [248] a linear Gaussian generative model was adopted as:

$$\mathbf{x}_i = \mathbf{W}\mathbf{y}_i + \boldsymbol{\epsilon}_i, \ \mathbf{y}_i \sim \mathcal{N}(\mathbf{0}, \mathbf{I}), \ \boldsymbol{\epsilon}_i \sim \mathcal{N}(\mathbf{0}, \sigma^2\mathbf{I}) \tag{3.22}$$

where $\mathbf{W} \in \Re^{F\times N}$ is the matrix that relates the latent variable $\mathbf{y}_i$ with the observed samples $\mathbf{x}_i$ and $\boldsymbol{\epsilon}_i$ is the noise which is assumed to be an isotropic Gaussian model. The motivation is that, when $N < F$, the latent variables will offer a more parsimonious explanation of the dependencies between the observations. The Maximum Likelihood (ML) and Expectation Maximisation (EM) solutions for parameter and moments $\mathbb{E}[\mathbf{y}_i]$ and $\mathbb{E}[\mathbf{y}_i\mathbf{y}_i^T]$ can be found in

their projections. Hotelling defined PCA as the orthogonal projection of the data onto a lower dimensional linear space (the principle suspace) where the variance of the projected data is maximised [108]. Both of these definition lead to the same algorithm.

[26, 248]. Several variations have from the proposed since, e.g. by incorporating sparseness and non-negative constraints [235] or utilising joint generative/regression frameworks (the so-called Supervised Probabilistic Principal Component Analysis (SPPCA) [284]). In SPPCA, a model $\mathbf{x}_i = \mathbf{W}_x\mathbf{y}_i + \boldsymbol{\epsilon}_i^x$ is assumed to generate the data, while a second generative framework models a set of outputs $\mathbf{z}_i$ on the latent variables $\mathbf{y}_i$ as $\mathbf{z}_i = \mathbf{W}_z\mathbf{y}_i + \boldsymbol{\epsilon}_i^y$, $\mathbf{y}_i \sim N(\mathbf{0}, \mathbf{I})$, $\boldsymbol{\epsilon}_i^x \sim N(\mathbf{0}, \sigma^2\mathbf{I})$, $\boldsymbol{\epsilon}_i^y \sim N(\mathbf{0}, \sigma^2\mathbf{I})$. $\mathbf{z}_i$ can represent outputs from a regression task or can stand for continuous class labels.

3.3.2 Canonical Correlation Analysis (CCA)

CCA has risen out of the need to study samples from "multiple batteries"[3]. Since CCA deals with multiple sets of observations, we assume the observation matrices $\mathbf{X}_1$ and $\mathbf{X}_2$. The projected data should be maximally correlated, i.e.

$$\arg \max_{\mathbf{W}_1, \mathbf{W}_2} \frac{\mathbf{W}_1^T \boldsymbol{\Sigma}_{X_1X_2} \mathbf{W}_2}{\sqrt{\mathbf{W}_1^T \boldsymbol{\Sigma}_{X_1X_1} \mathbf{W}_1} \sqrt{\mathbf{W}_2^T \boldsymbol{\Sigma}_{X_2X_2} \mathbf{W}_2}}. \tag{3.23}$$

where $\boldsymbol{\Sigma}_{XY}$ corresponding to the empirical covariance matrix on sample matrices $\mathbf{X}$ and $\mathbf{Y}$, i.e. $\boldsymbol{\Sigma}_{XY} = cov(X, Y)$. Due to scale invariance of the correlation with respect to the loadings, the problem can be posed as

$$\max_{\mathbf{W}_1, \mathbf{W}_2} \ \mathbf{W}_1^T \mathbf{X}_1 \mathbf{X}_2^T \mathbf{W}_2 \tag{3.24}$$

$$s.t.\ \mathbf{W}_1^T \mathbf{X}_1 \mathbf{X}_1^T \mathbf{W}_1 = \mathbf{I}, \ \ \mathbf{W}_2^T \mathbf{X}_2 \mathbf{X}_2^T \mathbf{W}_2 = \mathbf{I}. \tag{3.25}$$

where the solution is found by solving the generalised eigenvalue problem

$$\mathbf{X}_1 \mathbf{X}_2^T (\mathbf{X}_2 \mathbf{X}_2^T)^{-1} \mathbf{X}_2 \mathbf{X}_1^T \mathbf{w}_1 = \lambda \mathbf{X}_1 \mathbf{X}_1^T \mathbf{w}_1 \tag{3.26}$$

and using the top eigenvectors for the loadings (where λ is the eigenvalue corresponding to the eigenvector $\mathbf{w}_1$). Most related to our work is the least-squares formulation of this problem [58, 240], where the solution of CCA can by found by solving

$$\arg \min_{\mathbf{W}_1, \mathbf{W}_2} ||\mathbf{W}_1^T \mathbf{X}_1 - \mathbf{W}_2^T \mathbf{X}_2||_F^2$$

$$s.t.\ \mathbf{W}_1^T \mathbf{X}_1 \mathbf{X}_1^T \mathbf{W}_1 = \mathbf{I}, \ \ \mathbf{W}_2^T \mathbf{X}_2 \mathbf{X}_2^T \mathbf{W}_2 = \mathbf{I}. \tag{3.27}$$

[3] Battery (tests) refers to a series of psychological, behaviour or cognitive assessment tests. This term was often used in statistics since data from multiple batteries were essentially the one of the first datasets which consisted of multiple modalities, leading to several significant publications in the field of statistics being published, e.g., in Psychometrika, a psychology oriented journal.

Probabilistic formulations of CCA have been explored in various forms, with the most recent being the work of Bach and Jordan [11], where a latent variable model with a maximum likelihood solution co-directional to deterministic CCA is defined as

$$\begin{aligned} \mathbf{z} &\sim \mathcal{N}(0, \mathbf{I}) \\ \mathbf{x}_1|\mathbf{z} &\sim \mathcal{N}(\mathbf{W}_1\mathbf{z}, \mathbf{\Psi}_1) \\ \mathbf{x}_2|\mathbf{z} &\sim \mathcal{N}(\mathbf{W}_2\mathbf{z}, \mathbf{\Psi}_2). \end{aligned} \tag{3.28}$$

where $\mathbf{\Psi}_i$ stands for the covariance matrix. The interest in this particular formulation lies in the fact that the common space (linked to the random variable $\mathbf{z}$) is explicitly represented, instead of discovered by minimising the sum-of-squares of the projected observation sets, as in the deterministic formulation. This work has later been extended by Klami and Kaski to include private spaces (i.e. modelling information specific to one observation set), thus making the model more similar to Inter Battery Factor Analysis (IBFA) [253], and its probabilistic interpretation by Browne [32]. In more detail, the model is defined as

$$\begin{aligned} \mathbf{z} &\sim \mathcal{N}(0, \mathbf{I}), \\ \mathbf{z_1} &\sim \mathcal{N}(0, \mathbf{I}), \\ \mathbf{z_2} &\sim \mathcal{N}(0, \mathbf{I}) \\ \mathbf{x}_1 &\sim \mathcal{N}(\mathbf{W}_1\mathbf{z} + \mathbf{B}_1\mathbf{z}_1, \mathbf{\Sigma}_1) \\ \mathbf{x}_2 &\sim \mathcal{N}(\mathbf{W}_2\mathbf{z} + \mathbf{B}_2\mathbf{z}_2, \mathbf{\Sigma}_2) \end{aligned} \tag{3.29}$$

with $\mathbf{\Sigma}_i$ representing a diagonal covariance matrix, indicating the independence of the noise component over the features. In this particularly useful formulation, the shared space is modelled in the latent variable $\mathbf{z}$, while the remaining variation is modelled via the latent variables $\mathbf{z}_i$, with both latent spaces being transformed to the observation space via linear mappings, specific to the observation set.

3.4 Time Warping

The problem of temporally aligning multiple signals is commonly encountered in many manifestations, and can be a common problem in cases of analysing signals obtained from multiple modalities (e.g., unsynchronised audio and video). In general, the alignment of temporal sequences is a very challenging problem, where besides computer vision [86, 114, 265, 297, 298, 296], has also been raised in the fields of bioinformatics [144] and speech processing [120, 219].

One can define the temporal alignment problem as finding the temporal coordinate transformation which renders the given sequences to be aligned in time. Traditionally, this problem can be solved via dynamic programming, utilising the so-called *Dynamic Time Warping* (DTW) technique. Given sequences $\mathbf{X}_1 \in \mathbb{R}^{D \times T_1}$ and $\mathbf{X}_2 \in \mathbb{R}^{D \times T_2}$, DTW can be defined as a least-squares problem as follows:

$$\arg\min_{\mathbf{\Delta}_1, \mathbf{\Delta}_2} ||\mathbf{X}_1\mathbf{\Delta}_1 - \mathbf{X}_2\mathbf{\Delta}_2||_F^2 \tag{3.30}$$

whereas now $\mathbf{\Delta}_1 \in \{0,1\}^{T_1 \times T_\Delta}$ and $\mathbf{\Delta}_2 \in \{0,1\}^{T_2 \times T_\Delta}$ are binary selection matrices, with T_Δ the aligned, common length. In this way, the warping matrices $\mathbf{\Delta}$ effectively re-map the samples of each sequence. The $\mathbf{\Delta}$ matrices are essentially a matrix representation of the warping path $\mathbf{p}$, which is a vector of the mapped indices from the original sequence to the resulting time-warped (i.e. $\mathbf{X}_1\mathbf{\Delta}_1^T = \mathbf{X}_1(\mathbf{p}_1)$). Although the number of possible alignments is exponential in T_xT_y, employing dynamic programming can recover the optimal path in $\mathcal{O}(T_xT_y)$ (i.e. polynomial time), with the optimal T_Δ automatically inferred. Furthermore, a set of constraints must be satisfied, namely (i) the boundary conditions: the first index of each p must be 1, and the last should map to the last frame of each sequence (T_1, T_2), (ii) the monotonicity condition: the p vectors must be in increasing order (not *strictly* increasing, since repetitions are allowed), and (iii) the continuity condition: $[p_1^{t+1}, p_2^{t+1}] - [p_1^t, p_2^t] \in \{[0,1],[1,0],[1,1]\}$. Although DTW provides an optimal solution, at least for 1 dimensional time-series, it comes with many disadvantages, such as the inability to process sequences with varying dimensionality (i.e. $D_1 \neq D_2$) as well as being highly susceptible to various forms of noise. As we will discuss in what follows, a solution that makes time warping a much more flexible method, and thus more appropriate for high-dimensional data usually associated with human behaviour analysis comes through incorporating time-warping with component analysis.

3.4.1 Time Warping and Component Analysis

The incorporation of Component Analysis (CA) and Time Warping (TW) is a natural consequence of the need to process and align high-dimensional data in modern scenarios, with more corruptions and noise. By utilising CA one can avoid applying time-warping to portions of the signal which are uninteresting: one can time-warp only the part of the signal which is relevant to the task, for example, one can align only the portion of each signal which is shared amongst all warped sequences while removing noisy components. The incorporation of CA and TW comes naturally: most CA methods assuming a generalised eigenvalue problem are also subject to least-squares formulations [58, 241]. This is very important in terms of incorporating them with time-warping algorithms, since both problems can be naturally combined

into one least squares problem. For example, by taking the DTW (Eq. 3.27) and the CCA formulations (Eq. 3.30), one can arrive at the following problem:

$$\begin{aligned}
&\underset{\mathbf{W}_1,\mathbf{W}_2,\boldsymbol{\Delta}_1,\boldsymbol{\Delta}_2}{\arg\min} \; ||\mathbf{W}_1^T\mathbf{X}_1\boldsymbol{\Delta}_1 - \mathbf{W}_2^T\mathbf{X}_2\boldsymbol{\Delta}_2||_F^2 \\
&s.t. \;\; \mathbf{W}_1^T\mathbf{X}_1\boldsymbol{\Delta}_1\boldsymbol{\Delta}_1^T\mathbf{X}_1^T\mathbf{W}_1 = \mathbf{I} && (3.31)\\
&\mathbf{W}_2^T\mathbf{X}_2\boldsymbol{\Delta}_2\boldsymbol{\Delta}_2^T\mathbf{X}_2^T\mathbf{W}_2 = \mathbf{I} && (3.32)\\
&\mathbf{W}_1^T\mathbf{X}_1\boldsymbol{\Delta}_1\boldsymbol{\Delta}_2^T\mathbf{X}_2^T\mathbf{W}_2 = diag, && (3.33)\\
&\mathbf{X}_1\boldsymbol{\Delta_1}\mathbf{1} = \mathbf{0}, \mathbf{X}_2\boldsymbol{\Delta_2}\mathbf{1} = \mathbf{0}. && (3.34)
\end{aligned}$$

where the added constraints ensure rotation, scaling and translation invariance. This leads to the *Canonical Time Warping* (CTW) model [298], successfully combining multi-series component analysis (CCA) with DTW in a model which allows for aligning signals from multiple modalities and varying dimensionality. At this point, it is worth mentioning that in the related statistical field of *Functional Data Analysis* (FDA) [207, 90], where the observed data are represented as functional data (e.g., utilising basis such as exponential, polynomial etc.), functional PCA has been applied along with time-warping (or registration as it is called in FDA). A related idea of utilising functional basis for time-warping has been introduced in [296], where the generalised time warping methodology introduced has been combined with CCA. Finally, the *dynamic manifold temporal warping* (DMTW) [86] and the *manifold warping* (MW) [265] extend the CTW to handle more complex spatial transformations through manifold learning.

3.5 Conclusions

In this chapter, we discussed a set of related machine learning techniques which are closely related to this thesis, focusing mostly on regression and component analysis. In what follows, we briefly map the techniques discussed in this chapter to the content of this thesis. Regarding the first part of the thesis, BLSTM-NN are utilised in Chapter 5, while the RVM is extended in Chapter 6. Finally, we propose a novel regression framework based on CCA in Chapter 7. As far as the second part of the thesis is concerned, it is entirely devoted to component analysis methods. In particular, Chapter 9 provides a probabilistic, shared-space component analysis method aiming mostly at fusing multiple annotations. Chapter 10 presents a novel, robust variant of CCA, which is able to learn a low-rank subspace while isolating gross errors in a sparse component. Finally, in Chapter 11, we propose a novel, unified framework for probabilistic component analysis, which is able to encapsulate methods such as Principal Component Analysis (PCA), Locality Preserving Projections (LPP), Linear Discriminant Analysis (LDA) and Slow Feature Analysis (SFA).

Part I

Learning Continuous Emotion Dimensions via Exploiting Output Correlations

Chapter 4

Introduction

Risen from the need to analyse emotions occurring spontaneously under real-world conditions, researchers adopted the continuous emotion dimensions in order to facilitate the description of typically encountered emotional states. In this part of the thesis, we explore the newly introduced problem of predicting and analysing human emotional states in terms of emotion dimensions. We are mostly motivated by various research findings in psychology which demonstrate that emotion dimensions exhibit some form of correlation. The idea posed herein, is that be exploiting such correlations and relationships, one can improve the accuracy of the predictive task at hand. The content of this chapter is summarised in what follows.

Chapter 5

In this chapter, we present one of the first studies in related work in terms of learning continuous and dimensional emotions, initially published in [174]. In particular, we present the first approach in literature towards automatic, dimensional and continuous affect predictions in terms of valence and arousal, based on all facial expressions, shoulder gestures and audio cues (at time of publication). Based on Bidirectional Long-Short Term Memory Neural Networks (BLSTM-NN), the presented approach is aimed at both learning long-range temporal dependencies, a crucial requirement for the given problem, as well modelling dependencies in the output dimensions. This work is in fact, to the best of our knowledge, the first work which explicitly aims to improve accuracy by modelling output relationships in emotion dimensions. In mode detail, in Chapter 5, we initially perform a comparison of BLSTM-NN to another, commonly used regression technique in the field, Support Vector Machines (SVMs). Subsequently, we focus on the fusion of multiple modalities, and compare two commonly employed fusion techniques, feature-level and model-level fusion, to the proposed output-associative fusion

based on LSTM-NN. Result-wise, BLSTM-NN and the proposed fusion technique overperform other, compared methods, establishing the significance of properly modelling temporal dependencies in the given problem, as well as exploiting output correlations.

Chapter 6

In Chapter 6 we present one of the first probabilistic methods particularly focused on the predictive analysis of continuous dimensional emotion dimensions. While the work presented in Chapter 5 was based on neural networks (NN), many researchers have criticised the inherent lack of interpretability of trained NN as well as the lack of an estimation of *uncertainty*. In contrast to NN, the Bayesian framework we adopt in this chapter provides an elegant solution to the problem, while estimating a sparse, parsimonious solution. In more detail, in Chapter 6 we present an extension of the Relevance Vector Machine (RVM, c.f., Chapter 3), which we coin Output-Associative Relevance Vector Machine (OA-RVM). By utilising an augmented design matrix with a temporal window, OA-RVM allows for learning temporal output dependencies manifesting in emotion dimensions within a probabilistic robust framework, inheriting the advantages posed by the original RVM framework while remaining in the same computational complexity class. Experiments are performed on all audio, visual and shoulder movement cues, while utilising a small number of data for training. Results show that OA-RVM significantly outperforms other regression techniques such as SVM and RVM.

Chapter 7

Finally, in Chapter 7, we firstly focus on empirically answering several important questions which have remained relatively unexplored in related literature, such as the correlation of each emotion dimension (i) with respect to other emotion dimensions, (ii) to basic emotions (e.g., happiness, anger) as well as (iii) to the level of interest. In more detail, in order to study the level of interest in comparison to other emotion dimensions, we essentially treat interest as a continuous emotion dimension, ranging from *disinterested* to *enthusiastic*. As a measure of comparison, we utilise audiovisual features. Interestingly enough, results show that (i) each emotion dimension is more correlated with other emotion dimensions rather than with face and acoustic features, and similarly (ii) that each basic emotion is more correlated with emotion dimensions than with audio and video features. Regarding interest, we find that interest is most correlated with the emotion dimension of arousal, while secondly with valence.

It is interesting to note here, that since each emotion dimension is *better* correlated to other emotion dimensions rather than to face or audio cues (which are of much higher dimensionality

than annotations), the idea of dimensionality reduction for this problem is further motivated. Furthermore, this empirical study further motivates the idea of exploiting output correlations in this problem. In this light, we present a method based on Canonical Correlation Analysis (CCA) for exploiting output correlations and learning emotion dimensions[1]. This work, which we coin Correlated Spaces Regression (CSR) deviates from the previous methods towards learning emotion dimensions as it is mostly focused on generating the appropriate features for utilising in terms of predictive analysis, therefore acting as a bridge between the more application-oriented, first part of the thesis, to the more technical-based second part, which focuses on component analysis. The basic idea lies in projecting both the features/observations and the outputs onto a latent space on which they are maximally correlated. The implications are two-fold. Firstly, this process maximally correlates the features with the outputs by projecting on a dimensionality reduced latent space, thus providing appropriate features for predictive analysis. Secondly, the output-dimensions are de-correlated via an orthogonal projection, thus enabling the utilisation of single-dimensionality regression to optimally learn the function mapping to the outputs. In essence, this method is highly useful for problems where we have multi-dimensional outputs, since any redundancy in the outputs is removed while the feature space dimensionality is reduced significantly without penalising predictive accuracy. As we show, this type of fusion provides better results than other alternatives employed in related work.

[1] Although Canonical Correlation Analysis is a shared-space component analysis method and in theory this Chapter is also relevant to the second part of this thesis, we describe this method in the first part since (i) the main contribution of this work is in the application domain, specifically to facilitate regression by extracting the appropriate features while capturing the dependencies of emotion dimensions in the form of correlations, and (ii) the method is based on an already existing component analysis technique, namely CCA.

Chapter 5

Continuous Prediction of Spontaneous Affect from Multiple Cues and Modalities in Valence–Arousal Space

Contents

5.1 Introduction

Motivated by evidence in psychology pointing out various correlations and relationships between emotion dimensions such as valence and arousal, in this chapter we introduce a novel, output-associative fusion methodology based on Bidirectional Long-Short Term Memory Neural Networks (BLSTM-NNs), which are able to learn both long and short term temporal dependencies. In this chapter, (i) the first approach towards automatic, dimensional and continuous affect prediction based on facial expression, shoulder gesture, and audio cues is presented[1], and (ii) a framework that integrates temporal correlations between continuous dimensional outputs

[1] At time of publication of [175].

(valence and arousal) to improve regression predictions is proposed. Our motivation for the latter is essentially based on psychological evidence reporting that the valence and arousal dimensions are inter-correlated [181],[5], [129], [138]. Despite this fact, automatic modelling of these correlations has not been attempted before this work.

The chapter is organised as follows. Initially, in Section 5.2 we outline the basic methodology employed. In Section 5.3 we briefly describe the dataset employed along with any pre-processing tasks. Section 5.4 explains audio and visual feature extraction and tracking. Section 5.5 describes the learning techniques and the evaluation measures employed for continuous emotion prediction, and introduces the output-associative fusion framework, while in Section 5.6 we discuss the experimental results. Finally, in Section 5.7 we conclude the chapter.

5.2 Methodology

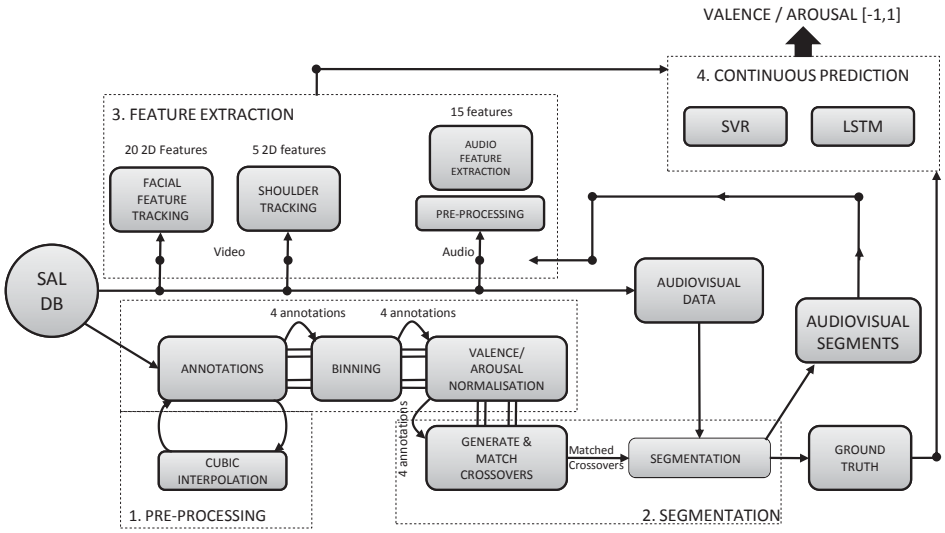


Figure 5.1: Methodology employed: Pre-processing, segmentation, feature extraction and prediction.

The methodology proposed in this chapter consists of pre-processing, segmentation, feature extraction, and prediction components, and is illustrated in Fig. 5.1. The first two stages, that of pre-processing and segmentation, depend mostly on the set of annotations provided with the SAL database (in terms of valence and arousal dimensions). We introduce various procedures to (i) obtain the ground-truth corresponding to each frame by maximizing inter-annotator agreement, and (ii) to determine the audiovisual segments that capture the transition *from* one emotional state *to* another (and back). Essentially, these procedures automatically segment

spontaneous multi-modal data in terms of negative and positive audiovisual segments that contain an offset before and after (i.e., the baseline) the displayed expression (Section 5.3.3).

During the feature extraction stage, the pre-segmented audiovisual segments from the SAL database are used. For the audio modality, the Mel-frequency Cepstrum Coefficients (MFCC) [115], as well as prosody features, such as pitch and energy features are extracted. To capture the facial and shoulder motion displayed during a spontaneous expression we use the Patras - Pantic particle filtering tracking scheme [190] and the standard Auxiliary Particle Filtering (APF) technique [199], respectively.

The final stage, that is based on all the aforementioned steps, consists of affect prediction, multi-cue and multi-modal fusion, and evaluation. SVRs and BLSTM-NNs are used for single-cue affect prediction. Due to the their superior performance, BLSTM-NNs are further used for feature and model-level fusion of multiple cues and modalities. An output-associative fusion framework, that employs a first layer of BLSTM-NNs for predicting V-A values from the original input features, and a second layer of BLSTM-NN using these predictions jointly as intermediate features to learn the V-A inter-dependencies (correlations), is introduced next. Performance comparison shows that the proposed output-associative fusion framework provides a significantly improved prediction accuracy compared to feature- and model-level fusion via BLSTM-NNs.

5.3 Dataset and Pre-processing

5.3.1 Dataset

We use the Sensitive Artificial Listener Database (SAL-DB) [64], which was described earlier in Chapter 2. SAL contains spontaneous data capturing the audiovisual interaction between a human and an operator undertaking the role of an avatar with four personalities: Poppy (happy), Obadiah (gloomy), Spike (angry) and Prudence (pragmatic). We utilise the valence and arousal annotations provided for SAL. Although there are approximately 10 hours of footage available in the SAL database, V-A annotations have only been obtained for two female and two male subjects. We used this portion for our experiments.

5.3.2 Challenges

Using spontaneous and naturalistic data that have been manually annotated along a continuum presents us with a set of challenges which essentially motivate the adopted methodology.

The first issue is known as *reliability of ground truth*. In other words, achieving agreement amongst the annotators (or observers) that provide annotations in a dimensional space is very challenging [95]. In order to make use of the manual annotations for automatic recognition, most researchers take the mean of the observers ratings, or assess the annotations manually. In Section 5.3.3, we describe the process of producing the ground-truth with respect to the annotators' annotations, in order to maximize the inter-annotator agreement.

The second issue is known as *the baseline problem*. This is also known as the concept of having 'a condition to compare against' in order for the automatic recognizer to successfully learn the recognition problem at hand [95]. For instance, in the context of acted (posed) facial expression recognition, the subjects are instructed to express a certain emotional state starting (and ending) with an expressionless face. Thus, posed affect data contain all the temporal transitions (neutral - onset - apex - offset - neutral) that provide a classifier with a sequence that begins and ends with an expressionless display: the baseline. Since such expressionless states are not guaranteed to be present in spontaneous data [95, 136], we use the transition *to* and *from* an emotional state (i.e., the frames where the emotional state changes) as the baseline to compare against.

The third issue refers to *unbalanced data*. In naturalistic settings it is very difficult to elicit balanced amount of data for each emotion dimension. For instance, [38] reported that a bias toward quadrant 1 (positive arousal, positive valence) exists in the SAL database. Other researchers (e.g., [42]) handle the issue of unbalanced classes by imposing equal a priori probability. As classification results strongly depend on the a priori probabilities of class appearance, we attempt to tackle this issue by automatically pre-segmenting the data at hand. More specifically, the segmentation stage consists of producing (approximately equal number of) negative and positive audiovisual segments with a temporal window that contains an offset before and after the displayed expression (i.e., the baseline).

5.3.3 Data Pre-processing and Segmentation

The data pre-processing and segmentation stage consists of (i) producing ground-truth by maximizing inter-annotator agreement, (ii) eliciting frames that capture the transition *to* and *from* an emotional state, and (iii) automatic segmentation of spontaneous audiovisual data. A detailed description of these procedures is presented in [173].

In general, the V-A annotations of each annotator are not in total agreement , mostly due to the variance in human observers' perception and interpretation of emotional expressions.

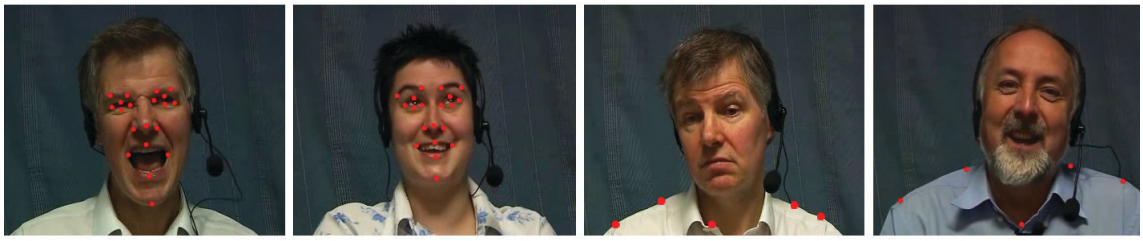


Figure 5.2: Illustration of tracked facial points (T_{f1}–T_{f20}) and shoulder points (T_{s1}–T_{s5}).

Thus, in order to deem the annotations comparable, we normalized the data and provided some compensation for the synchronization issues. We experimented with various normalization techniques and opted for the one that minimized the inter-annotator MSE. To tackle the synchronization issues, we allow the time-shifting of the annotations for each specific segment up to a threshold of 0.5 sec. given that this increases the agreement between annotators.

In summary, achieving agreement from all participating annotators is difficult and not always possible for each extracted segment. Thus, we use the inter-annotator correlation to obtain a measure of how similar one annotator's annotations are to the rest. This is then used as a weight to determine the contribution of each annotator to the ground truth.

More specifically, the averaged correlation cor'_{S,c_j} assigned to annotator c_j is defined as follows:

$$cor'_{S,c_j} = \frac{1}{|S|-1} \sum_{i \in S, c_i \neq c_j} cor(c_i, c_j) \tag{5.1}$$

where S is the relevant session annotated by $|S|$ number of annotators, and each annotator annotating S is defined as $c_i \in S$.

Typically, an automatically produced segment consist of a single interaction of the subject with the avatar (operator), starting with the final seconds of the avatar speaking, continuing with the subject responding (and thus reacting and expressing an emotional state) and concluding where the avatar starts responding. Given that in naturalistic data, emotional expressions are not generally preceded by neutral emotional states [95, 136], we considered this window to provide the best baseline possible. For more details, we refer the reader to [173]. It should be noted that this method is purely based on the annotations, unlike other methods which are based on features, e.g. turn-based segmentation based on voice activity detection [151].

5.4 Feature Extraction

In this section we describe the audio and visual features that have been extracted using the automatically segmented audiovisual SAL data.

5.4.1 Acoustic Features

Extracted acoustic features include Mel-frequency Cepstrum Coefficients (MFCC) [115] and prosody features (the energy of the signal, the Root Mean Squared Energy and the pitch obtained by using a Praat pitch estimator [191]). Mel-frequency Cepstrum (MFC) is a representation of the spectrum of an audio sample which is mapped onto the nonlinear mel-scale of frequency to better approximate the human auditory system's response. The MFC coefficients (MFCC) collectively make up the MFC for the specific audio segment.

We used 6 cepstrum coefficients, thus obtaining 6 MFCC and 6 MFCC-Delta features for each audio frame. We have essentially used the typical set of features used for automatic affect recognition [285], [196]. Along with pitch, energy and RMS energy, we obtained a set of features with dimensionality $d = 15$ (per audio frame). Note that we used a 0.04 second window with a 50% overlap (i.e. first frame 0-0.04, second from 0.02-0.06 and so on) in order to obtain a double frame rate for audio (50 Hz) compared to that of video (25 fps). This is an effective and straightforward way to synchronise the audio and video streams.

5.4.2 Facial Expression Features

To capture the facial motion displayed during a spontaneous expression we track 20 facial feature points (FFP), as illustrated in Fig. 5.2. These points are the corners of the eyebrows (4 points), eyes (8 points), nose (3 points), the mouth (4 points) and the chin (1 point). To track these facial points we used the Patras - Pantic particle filtering tracking scheme [190]. Prior to tracking, initialization of the facial feature points has been done using the method introduced in [266]. For each video segment containing n frames, we obtain a set of n vectors containing 2D coordinates of the 20 points tracked in n frames ($T_f = \{T_{f1} \dots T_{f20}\}$ with dimensions $n * 20 * 2$).

5.4.3 Shoulder Features

The motion of the shoulders is captured by tracking 2 points on each shoulder and one stable point on the torso, usually just below the neck (see Fig. 5.2). The points to be tracked are initialized manually in the first frame. We then use the standard Auxiliary Particle Filtering

(APF) [199] to track the shoulder points. This scheme is less complex and faster compared to the Patras - Pantic particle filtering tracking scheme, it does not require learning the model of prior probabilities of the relative positions of the shoulder points, while resulting in sufficiently high accuracy. The shoulder tracker results in a set of points $T_s = \{T_{s1} \dots T_{s5}\}$ with dimensions of $n * 5 * 2$.

5.5 Dimensional Affect Prediction

5.5.1 Learning Techniques

As aforementioned, in this chapter we utilise BLSTM-NN for regression. BLSTM-NNs are a recent extension of Recurrent Neural Networks (RNNs), that are able to model both long-term and short-term dependencies in observations. Furthermore, for comparison we utilise non-linear Support Vector Regression (SVR), as it is commonly employed in the prediction of continuous affect [151, 91, 116]). For more details with regards to BLSTM-NNs and SVR, please refer to Chapter 3.

5.5.2 Evaluation Metrics

Finding optimal evaluation metrics for dimensional and continuous emotion prediction and recognition remains an open research issue [95]. The mean squared error (MSE) is the most commonly used evaluation measure by related work in the literature (e.g., [151, 91, 116]) while correlation coefficient is also employed by several studies (e.g., [91, 116]).

MSE evaluates the prediction by taking into account the squared error of the prediction from the ground truth. Let $\hat{\theta}$ be the prediction and θ be the ground truth. MSE is then defined as:

$$MSE = \frac{1}{n}\sum_{f=1}^{n}(\hat{\theta}(f) - \theta(f))^2 = \sigma_{\hat{\theta}}^2 + E([\hat{\theta} - \theta])^2 \tag{5.2}$$

As can be seen from Eq. 5.2, MSE is the sum of the variance and the squared bias of the predictor, where E is the expected value operator. Therefore, the MSE provides an evaluation of the predictor based on its variance and bias. This also applies for other MSE-based metrics, such as the root mean squared error (RMSE), defined as RMSE $= \sqrt{\text{MSE}}$. In this work we use the RMSE since it is measured in the same units as our actual data (as opposed to the squared units measuring MSE). We note that MSE-based evaluation has been criticized for heavily weighting outliers [24]. Most importantly, it is unable to provide any structural information regarding how θ and $\hat{\theta}$ change together, i.e. the covariance of these values. The

correlation coefficient (COR), that we employ for evaluating the prediction and ground truth, compensates for the latter, and is defined as follows:

$$COR(\hat{\theta}, \theta) = \frac{COV\{\hat{\theta}, \theta\}}{\sigma_{\hat{\theta}}\sigma_{\theta}} = \frac{E[(\hat{\theta} - \mu_{\hat{\theta}})(\theta - \mu_{\theta})]}{\sigma_{\hat{\theta}}\sigma_{\theta}} \tag{5.3}$$

where σ stands for the standard deviation, COV stands for the covariance while μ symbolises the mean (expected value).

COR provides an evaluation of the *linear* relationship between the prediction and the ground truth, and subsequently, an evaluation of whether the model has managed to capture linear structural patterns inhibited in the data at hand. As for the covariance calculation, since the means are subtracted from the values in question, it is independent of the bias (and differs from the MSE-based evaluation).

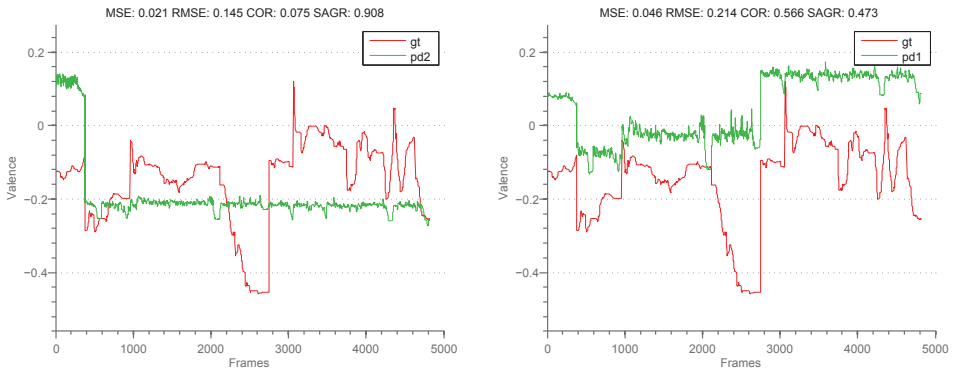
In addition to the two aforementioned metrics, we propose the use of another metric which can be seen as *emotion-prediction-specific*. Our aim is to obtain an agreement level of the prediction with the ground truth by assessing the valence dimension, as being positive (+) or negative (-), and the arousal dimension, as being active (+) or passive (-). Based on this heuristic, we define a sign agreement metric (SAGR) as follows:

$$SAGR = \frac{1}{n}\sum_{f=1}^{n} \delta_{(sign(\hat{\theta}(f)), sign(\theta(f)))} \tag{5.4}$$

where δ is the Kronecker delta function, defined as:

$$\delta_{(a,b)} = \begin{cases} 1, & a = b \\ 0, & a \neq b. \end{cases} \tag{5.5}$$

As a proof of concept, we provide two cases from our experiments that demonstrate how each evaluation metric contributes to the evaluation of the prediction with respect to the ground truth. In Fig. 5.3 we present two sub-optimal predictions from audio cues, for the valence dimension, using two BLSTM-NNs with different topologies. Notice how each metric informs us of a specific aspect of the prediction. The MSE of Fig. 5.3a is smaller than Fig. 5.3b, demonstrating that the first case is numerically closer to the ground truth than the second case. Despite this fact, the first prediction does not seem to follow the ground truth structurally, it rather fluctuates around the mean of the prediction (generating a low bias). This is confirmed by observing COR which is significantly higher for the second prediction case (0.566 vs. 0.075). Finally, SAGR demonstrates that the first prediction case is in high



(a) (b)

Figure 5.3: Illustration of how MSE-based (both MSE and RMSE), COR and SAGR evaluation metrics provide different results for two different predictions on the same sequence (gt: ground truth, pd: prediction).

agreement with the ground truth, in terms of classifying the emotional states as negative or positive. In summary, the MSE and the COR seem to capture the most important structural characteristics of the prediction, while SAGR confirms the previous.

Our empirical evaluations show that there is an inherent trade off involved in the optimization of these metrics, an issue which lies within each employed learning technique. By using all three metrics simultaneously we attain a more detailed and complete evaluation of predictor vs. ground truth, i.e., (i) a variance-and-bias-based evaluation with MSE (how much prediction and ground truth values vary), (ii) a structure-based evaluation with COR (how closely the prediction follows the structure of the ground truth), and (iii) class related evaluation with SAGR (how much prediction and ground truth agree on the positive vs. negative, and active vs. passive aspect of the exhibited expression). Of course, the set of utilised metrics should be compliant to the problem settings at hand.

5.5.3 Single-cue Prediction

The first step in our experiments consists of prediction based on single cues. Let $\mathcal{D} = \{V, A\}$ represent the set of dimensions, $\mathcal{C}$ the set of cues consisting of the facial expressions, shoulder movement and audio cues. Given a set of input features $\mathbf{x_c} = [\mathbf{x_{1_c}}, \ldots, \mathbf{x_{n_c}}]$ where n is the training sequence length and $c \in \mathcal{C}$, we train a machine learning technique f_d, in order to predict the relevant dimension output, $\mathbf{y_d} = [y_1, \ldots, y_n]$, $d \in \mathcal{D}$.

$$f_d : \mathbf{x} \mapsto y_d \tag{5.6}$$

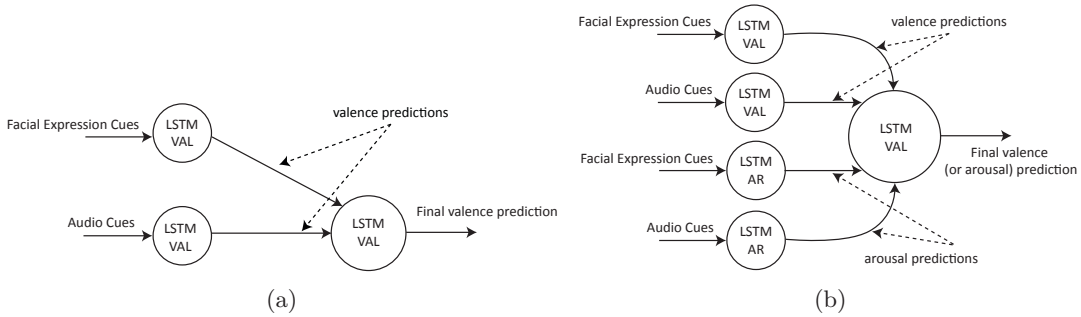


Figure 5.4: Illustration of (a) model-level fusion and (b) output-associative fusion using facial expression and audio cues. Model-level fusion combines valence predictions from facial expression and audio cues by using a third network for the final valence prediction. Output-associative fusion combines both valence and arousal values predicted from facial expression and audio cues, again by using a third network which outputs the final prediction.

This step provides us with a set of predictions for each machine learning technique, and each relevant dimension employed.

5.5.4 Feature-level Fusion

Feature-level fusion is obtained by concatenating all the features from multiple cues into one feature vector which is then fed into a machine learning technique. In our case, the audio stream has a double frame rate with respect to the video stream (50 Hz vs. 25 fps). When fusing audio and visual features (shoulder or facial expression cues) at the feature-level, each video feature vector is repeated twice, and the ground truth for the audio cues is then used for training and evaluation. This practice is in accordance with similar works in the field that focus on human behaviour understanding from audiovisual data (e.g., [196]).

5.5.5 Model-level Fusion

In the decision-level data fusion, the input coming from each modality and cue is modelled independently, and these single-cue and single-modal recognition results are combined in the end. Since humans display multi-cue and multi-modal expressions in a complementary and redundant manner, the assumption of conditional independence between modalities and cues in decision-level fusion can result in loss of information (i.e. mutual correlation between the modalities). Therefore, we opt for model-level fusion of the continuous predictions as this has the potential of capturing correlations and structures embedded in the continuous output of the regressors (from different sets of cues). This is illustrated in Fig. 5.4b.

More specifically, during model-level fusion, a function learns to map predictions to a dimension d from the set of cues as follows:

$$f_{mlf} : f_d(\mathbf{x_1}) \times \cdots \times f_d(\mathbf{x_m}) \mapsto y_d \tag{5.7}$$

where m is the total number of fused cues.

5.5.6 Output-associative Fusion

In the previous sections, we have treated the prediction of valence or arousal as a 1D regression problem. However, as aforementioned, psychological evidence shows that valence and arousal dimensions are correlated [181],[5],[281].

In order to exploit these correlations and patterns, we propose a framework capable of learning the dependencies that exist amongst the predicted dimensional values. We use BLSTM-NN as the basis for this framework as they appear to outperform SVR in the prediction task at hand (see Section 5.6). Given the setting described in Section 5.5.3, this framework learns to map the outputs of the intermediate predictors (each BLSTM-NN as formulated in Eq. 5.6) onto a higher (and final) level of prediction by incorporating cross-dimensional (output) dependencies (see Fig. 5.4a). This method, that we call *output-associative fusion*, can be represented by a function f_{oaf}:

$$f_{oaf} : f_{Ar}(\mathbf{x_1}) \times f_{Val}(\mathbf{x_1}) \cdots \times f_{Ar}(\mathbf{x_m}) \times f_{Val}(\mathbf{x_m}) \mapsto y_d \tag{5.8}$$

where m is again the total number of fused cues. As a result, the final output, taking advantage of the temporal and bidirectional characteristics of the regressors (BLSTM-NNs), depends not only on the entire sequence of input features $\mathbf{x_i}$ but also on the entire sequence of intermediate output predictions $\mathbf{f_d}$ of both dimensions (see Fig. 5.4a).

5.5.7 Experimental Setup

Prior to experimentation, all features used for training the machine learning techniques have been normalized to the range of [-1,1], except for the audio ones, which have been found to perform better with z-normalization (i.e., normalizing to mean=0 and standard deviation=1).

For validation purposes we use a subset of the SAL-DB that consists of 134 audiovisual segments (a total of 30,042 video frames) obtained by the automatic segmentation procedure (see [173]). We employ subject-dependent leave-one-out-validation evaluation as most of the

works in the field report only on subject-dependent dimensional emotion recognition when the number of subjects and data are limited (e.g., [151]).

For automatic dimensional affect prediction we employ two state-of-the-art machine learning techniques: Support Vector Machines for Regression (SVR) and bidirectional Long Short-Term Memory Neural Networks (BLSTM-NN). Experimenting with SVR and BLSTM-NN requires that various parameters within these learning methods are configured and the interaction effect between various parameters is investigated. For SVR we experiment with Radial Basis Function (RBF) kernels ($e^{(-\gamma\|\mathbf{x}-\mathbf{x}'\|^2)}$) as the results outperformed our initial polynomial kernel experiments. To this aim, kernel specific parameters, such as the γ RBF kernel parameter (which determines how closely the distribution of the data is followed) and the polynomial kernel degree as well as generic parameters, including the outlier cost C, the tolerance of termination and the ϵ of the loss function need to be optimized. We perform a grid search (using the training set) and select the best performing set of parameters to be used.

The respective parameter optimization for BLSTM-NNs refers to mainly determining the topology of the network along with the number of epochs, momentum and learning rate. Our networks typically have one hidden layer and a learning rate of 10^{-4}. The momentum is varied in the range of [0.5, 0.9]. All these parameters can be determined by optimizing on the given training set (e.g., by keeping a validation set aside) while avoiding overfitting.

5.6 Experimental Evaluation

5.6.1 Single-cue Prediction

To evaluate the performance of the employed learning techniques for continuous affect prediction, we firstly experiment with single cues. Table 5.1 presents the results of applying BLSTM-NN and SVR (with a radial basis function kernel) for the prediction of valence and arousal dimensions.

We initiate our analysis with the valence dimension. From both BLSTM-NNs and SVR, it is obvious that the visual cues appear more informative than audio cues. Facial expression cues provide the highest correlation with the ground truth (COR=0.71) compared to shoulder cues (COR=0.59) and audio cues (COR=0.44). This fact is also confirmed by the RMSE and SAGR values. Facial expression cues provide the highest SAGR (0.84) indicating that the predictor was accurate in predicting an emotional state as positive or negative for 84% of the frames.

Table 5.1: Single-cue prediction results for valence and arousal dimensions (F: Facial Expressions, S: Shoulder Cues, A: Audio). Evaluation is performed by utilising the Root Mean Squared Error (RMSE), the correlation coefficient (COR) and the sign agreement (SAGR).

		BLSTM-NN			SVR		
		RMSE	**COR**	**SAGR**	**RMSE**	**COR**	**SAGR**
Valence	**F**	**0.17**	**0.712**	**0.841**	0.21	0.551	0.740
	S	**0.21**	**0.592**	**0.781**	0.25	0.389	0.718
	A	**0.22**	**0.444**	**0.648**	0.25	0.146	0.538
Arousal	**F**	**0.25**	**0.493**	**0.681**	0.27	0.418	0.700
	S	0.29	**0.411**	**0.687**	**0.27**	0.388	0.667
	A	**0.24**	**0.586**	**0.764**	0.26	0.419	0.716

Works on automatic affect recognition from audio have reported that arousal can be much better predicted than valence using audio cues [91], [251]. Our results are in agreement with such findings, for prediction of the arousal dimension audio cues appear to be superior to visual cues. More specifically, audio cues (using BLSTM-NNs) provide COR=0.59, RMSE=0.24, and AGR=0.76. The facial expression cues provide the second best results with COR=0.49, while the shoulder cues are deemed less informative for arousal prediction. These findings are also confirmed by the SVR results.

In Table 5.1, we present a comparison of the performance of the employed learning techniques. We clearly observe that BLSTM-NNs outperform SVRs. In particular, COR and SAGR metrics provide better results for BLSTM-NNs (for all cues and all dimensions). The RMSE metric also confirms these findings except for the prediction of arousal from shoulder cues. Overall, we conclude that capturing temporal correlations and *remembering* the temporally distant events (or storing them in memory) is of utmost importance for continuous affect prediction.

5.6.2 Multi-cue and Multi-modal Fusion

The experiments in the previous section have demonstrated that using BLSTM-NNs provide better results (for all cues and all dimensions) than using SVRs. Therefore, BLSTM-NNs are employed for feature-level and model-level fusion, as well as output-associative fusion (described in Section 7.7). Experimental results are presented in Table 5.2, along with the statistical significance test results. We performed statistical significance tests (t-test) using $alpha = 0.05$ (95% confidence interval). We performed t-tests to compare the RMSE results of the proposed output-associative fusion to that of the best of model-level or feature-level fusion result (for each cue combination). Table 5.2 shows the significant results marked with a †. Overall, the output-associative fusion appears to be *significantly* better than the other

Table 5.2: Fusion results for the three methods employed. The best results are obtained by employing output-associative fusion. Significant results are marked with a †. For comparison purposes, the average agreement level between human annotators is also shown. Evaluation is performed by utilising the Root Mean Squared Error (RMSE), the correlation coefficient (COR) and the sign agreement (SAGR).

		output-associative			**model-level**			**feature-level**		
		RMSE	**COR**	**SAGR**	**RMSE**	**COR**	**SAGR**	**RMSE**	**COR**	**SAGR**
Valence	**FS**	0.15	**0.777**	**0.89**	0.16	0.774	**0.890**	0.19	0.676	0.845
	SA	**0.18**	**0.664**	0.825	0.19	0.653	**0.830**	0.21	0.583	0.733
	FA	**0.16**†	**0.760**	**0.892**	0.17	0.748	0.856	0.20	0.604	0.790
	FSA	**0.15**†	**0.796**	**0.907**	0.16	0.782	0.892	0.19	0.681	0.856
	annotators	0.141	0.85	0.86	0.141	0.85	0.86	0.141	0.85	0.86
Arousal	**FS**	**0.24**†	**0.536**	**0.719**	0.25	0.479	0.666	0.27	0.508	0.731
	SA	**0.23**†	**0.602**	**0.763**	0.26	0.567	0.637	0.28	0.461	0.685
	FA	**0.22**†	**0.628**	0.800	0.23	0.605	0.800	0.24	0.589	0.763
	FSA	**0.21**†	**0.642**	**0.766**	0.22	0.639	0.763	0.26	0.500	0.700
	annotators	0.145	0.87	0.84	0.145	0.87	0.84	0.145	0.87	0.84

fusion methods, except for prediction of valence from face-shoulder and shoulder-audio cue combinations.

Looking at Table 5.2, feature-level fusion appears to be the worst performing fusion method for the task and data at hand. Although in theory the cross-cue temporal correlations can be exploited by feature-level fusion, this does not seem to be the case for the problem at hand. This is possibly due to the increased dimensionality of the feature vector along with synchronicity issues between the fused cues.

In general model-level fusion provides better results than feature-level fusion. This can be justified by the fact that the BLSTM-NNs are able to learn temporal dependencies and structural characteristics manifesting in the continuous output of each cue. Model-level fusion appears to be much better for predicting the valence dimension rather than the arousal dimension. This is mainly due to the fact that the single-cue predictors for valence dimension perform better, thus containing more correct temporal dependencies and structural characteristics (while the weaker arousal predictors contain less of these dependencies). Both fusion techniques re-confirm that visual cues are more informative for valence dimension than audio cues. Finally, the fusion of all cues and modalities provides us with the best (most accurate) results.

Regarding the arousal dimension, we observe that the performance gap between model-level and feature-level fusion is smaller compared to that of valence dimension. For instance, for the fusion of face and shoulder cues, the feature-level fusion provided better COR and SAGR

results (but a worse RMSE) than model-level fusion.

Facial expression and audio cues have been the best performing single cues for continuous emotion prediction (see Section 5.6.1). Therefore it is not surprising that fusion of these two cues provides the best feature-level fusion results. For model-level fusion instead, the best results are obtained by combining the predictions from all cues and modalities.

Finally, the proposed output-associative fusion provides the best results, outperforming both feature-level and model-level fusion. Similar to the model-level fusion case, the best results (for both dimensions) are obtained when predictions from all cues and modalities are fused.

We denote that the performance increase of output-associative fusion is higher for the arousal dimension (compared to the valence dimension). This could be justified by the fact that the single-cue predictors for valence perform better than for arousal (Table 5.1) and thus, more correct valence patterns are passed onto the output-associative fusion.

Table 5.2 also shows the average agreement level between human annotators in terms of RMSE, COR and SAGR metrics (calculated for each dimension separately). It is interesting to note that when predicting the valence dimension, the proposed output-associative fusion (i) appears to outperform the average human annotator in terms of SAGR criterion, and (ii) provides prediction results that are relatively close to human annotators (in terms of RMSE and COR).

In Fig.5.5, we illustrate a set of predictions obtained via output-associative fusion. As can be observed from the figure, the prediction results closely follow the structure and the values of the ground truth.

Overall, the temporal dynamics of spontaneous multi-modal behaviour (e.g., when a facial or a bodily expression starts, reaches an apex, and ends) have not received much attention in the affective and behavioural science research fields. More specifically, it is virtually unknown whether and how the temporal dynamics of various communicative cues are inter-related (e.g., whether a smile reaches its apex while the person is shrugging his shoulders). The facial, shoulder and audio cues explored in this chapter possibly have different temporal dynamics. Accordingly, the BLST-NN are able to incorporate and model the temporal dynamics of each modality independently (and appropriately) in the output-associative and model-level fusion schemes. This may be one reason why output-associative and model-level fusion appear to perform better than feature-level fusion.

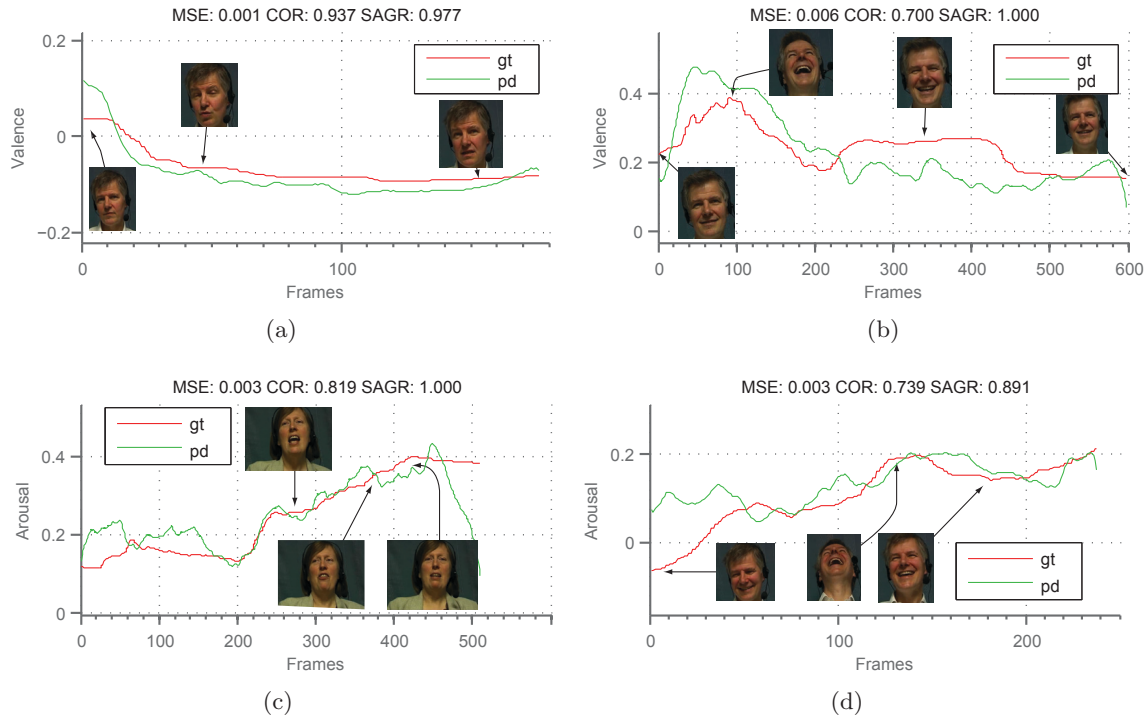


Figure 5.5: Example valence (5.5a, 5.5b) and arousal (5.5c, 5.5d) predictions obtained by output-associative fusion. (gt: ground truth, pd: prediction)

5.7 Conclusions

Affect sensing and recognition field has recently shifted its focus towards subtle, continuous, and context-specific interpretations of affective displays recorded in naturalistic and real-world settings, and towards combining multiple modalities for automatic analysis and recognition. The work presented in this chapter converges with this recent shift by (i) extracting audiovisual segments from databases annotated in dimensional affect space and automatically generating the ground truth, (ii) fusing facial expressions, shoulder and audio cues for dimensional and continuous prediction of emotions, (iii) experimenting with state-of-the-art learning techniques such as BLSTM-NNs and SVRs, and (iv) incorporating correlations between valence and arousal values via output-associative fusion to improve continuous prediction of emotions.

Based on the experimental results provided in Section 5.6 we are able to conclude the following:

- Arousal can be much better predicted than valence using audio cues. For valence dimension instead, visual cues (facial expressions and shoulder movements) appear to perform

better. This has also been confirmed by other related work on dimensional emotion recognition [151], [91], [251]. Whether such conclusions hold for different context and different data remains to be evaluated.

- Emotional expressions change over the course of time, and usually have start, peak, and end points (temporal dynamics). It appears that such temporal aspects (dynamics) are crucial in predicting both valence and arousal dimensions. A learning technique, such as the BLSTM-NNs, that can exploit these aspects, appears to outperform SVR (the static learning technique at hand).

- As confirmed by the psychological theory, valence and arousal are correlated. Such correlations appear to exist in our data where fusing predictions from both valence and arousal dimensions (via output-associative fusion) improves the results compared to using predictions from either valence or arousal dimension alone (both for feature-level and model-level fusion).

- In general, multi-modal data appear to be more useful for predicting valence than for predicting arousal. While arousal is better predicted by using acoustic features alone, valence is better predicted by using multi-cue and multi-modal data.

Overall, we conclude that compared to an average human annotator, the proposed system is well able to approximate the valence and arousal dimensions. More specifically, for valence dimension our output-associative fusion framework approximates the inter-annotator RMSE (≈ 0.141) and inter-annotator correlation (0.84) by obtaining a RMSE =0.15 and $COR \approx 0.8$ (see Table 5.2). It also achieves a higher SAGR ($\approx$ 0.91) than the inter-annotator SAGR (0.86).

Chapter 6

Output-Associative RVM Regression for Dimensional and Continuous Emotion Prediction

Contents

6.1 Introduction

Kernel methods such as Support Vector Machines (SVM), Relevance Vector Machines (RVM) and Gaussian Processes (GP) are amongst the most dominant techniques used in machine learning and computer vision. Many problems in these fields are inherently related to the prediction of multi-dimensional, inter-correlated structured outputs (e.g. pose normalization, pose estimation). While most machine learning techniques aim at capturing input relationships and patterns (e.g. extracted features), many problems expose an inherent dependency amongst the output dimensions (e.g. emotion dimensions). Not being able to learn such co-occurrences

can result in less robust and less accurate predictors, that will not be able to exploit specific output configurations manifested in the training data.

With these intrinsic motivations, we introduce the output-associative RVM (OA-RVM) regression, a framework that extends the traditional RVM regression by being able to learn temporal output correlations, while maintaining the advantage of a sparse formulation, fitting for large datasets. As we show by means of various experiments, OA-RVM appears to be advantageous against traditional RVM not only in terms of a variance-and-bias-based evaluation with Root Mean Squarred Error (RMSE, i.e., how much prediction and ground-truth values vary), but also with a structure-based evaluation with the correlation coefficient (COR, i.e., evaluating the covariance of the prediction with the ground truth), resulting in a more accurate and robust model. In order to evaluate whether the proposed technique's performance is *cue and modality invariant*, we focus on a highly challenging, yet a very suitable problem: dimensional and continuous emotion prediction from nonverbal *heterogeneous* cues (i.e., facial expressions, shoulder movements and audio cues).

Our motivation for the work presented in this chapter is three-fold. As in Chapter 5, we are primarily motivated by psychological evidence hinting that the V-A dimensions are inter-correlated [181, 5, 129, 138]. The proposed framework aims to enable the learning of such correlations and generate more substantiated predictions by embedding in the model an initial output estimation (using RVM) together with the original input features. Secondly, temporal dynamics play a significant role in emotion recognition [95, 285]. The proposed OA-RVM regression aims to capture the temporal dynamics by employing a temporal window (covering a set of past and future outputs) in order to accommodate temporal (output) patterns both in past and future context. Thirdly, dimensional and continuous prediction of emotions is a relatively unexplored area in the field of affective computing, and which prediction method is best suited to the task is still unknown. Therefore, as well as validating the proposed OA-RVM model with comprehensive experiments, we also compare it to traditional regression techniques such as RVM and Support Vector Regression (SVR).

The rest of this chapter is organized as follows. In Section 6.2, we mention some related work on output-associative regression which has been applied to different problems. In Section 6.3, we describe the proposed OA-RVM model, while also covering issues such as parametrisation and complexity. Subsequently, in Section 6.4, we describe the data and feature utilised in our experiments, while in Section 6.6, we present the experiments and discuss the results. Finally, we conclude the chapter in Section 6.7.

6.2 Related Work on Output-Associative Structured Regression

Although the idea of explicitly modelling the relationships amongst emotion dimensions in order to facilitate learning is has been unexplored before [174], various machine learning techniques exist that aim at capturing spatial output-associations. For example, Kernel Dependency Estimation (KDE) [269], utilises Kernel Principal Component Analysis and ridge regression for modelling output structure. KDE was later reformulated in [46], where a cost function was optimised directly thus disregarding the need for KPCA. KDE has been mostly applied to problems such as string matching and image reconstruction. Other attempts towards such solving such problems lie in extensions of Kernel Ridge Regression (KRR) and SVR (Support Vector Regression), which optimise an output-associative function [28]. Furthermore, in [29] the Twin Gaussian Process (TGP) is proposed, which employs GP priors for modelling input and output relations, while adopting the *Kullback-Leibler* divergence for optimisation. Both of these models have been applied to modelling human pose estimation. We choose to extend RVM as it is considered to be more efficient than traditional GP, and is known to provide a sparse solution. Note that other works on extending RVM have also been proposed, e.g. [165] proposed a robust RVM which models outlier noise while [243] proposed a multi-variate version of RVM.

Compared to the models presented in [28, 29] we offer a specific output temporal window parameter for fine-tuning our model. Furthermore, compared to [28], our OA-RVM regression framework offers a probabilistic formulation of the output-associative function by following the original RVM framework, and thus provides explicit modelling of the noise component.

6.3 The OA-RVM Framework

In this section we describe the proposed OA-RVM framework. The outline of the proposed method is presented in Fig. 6.1. The tracked / extracted features (from facial expressions, shoulder movements and audio) are fed into an initial (cue-specific) regressor, which in our case is chosen to be RVM (trained separately for each cue). An initial, noisy prediction is obtained by RVM. A temporal window v is applied on the multi-dimensional output of Valence and Arousal, thus constructing a set of new vectors which we call output features ($\mathbf{y_i^v}$). Both the input features $\mathbf{x_i}$ and the output features $\mathbf{y_i^v}$ are fed into the OA-RVM model which learns specific weights for each input and output feature vector. The final prediction is a linear combination of the kernel-projected input and output features.

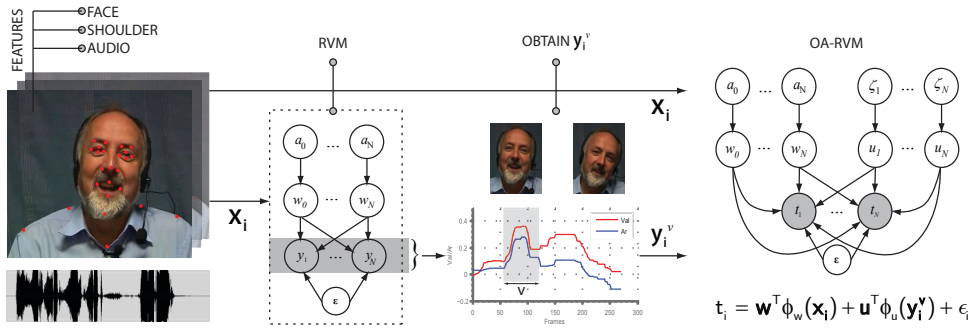


Figure 6.1: Outline of the proposed method. The tracked features (from facial expressions, shoulder movements and audio) are fed into an initial regressor (here, RVM) to obtain an initial prediction. A temporal window v is applied on the multi-dimensional output of Valence and Arousal, constructing the output feature vectors ($\mathbf{y_i^v}$). Both the input features $\mathbf{x_i}$ and the output features $\mathbf{y_i^v}$ are fed into the OA-RVM model which provides the final prediction.

Formally, we assume a (multidimensional) regression problem with N training examples, $(\mathbf{x_i}, \mathbf{t_i})$[1]. In the Bayesian framework applied in RVM (see also, Chapter 3), our goal is to learn the functional

$$t_i = \mathbf{w}^T \phi(\mathbf{x_i}) + \epsilon_i \tag{6.1}$$

where the ϵ_i are assumed to be independent Gaussian samples with zero mean and σ^2 variance, $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$. ϕ is a typically non-linear projection of the input features, $\mathbf{x_i}$. The method infers the set of weights $\mathbf{w}$ along with the noise estimation, given the training data.

In OA-RVM, we firstly increment Eq. 6.1 as follows

$$t_i = \mathbf{w}^T \phi_w(\mathbf{x_i}) + \mathbf{u}^T \phi_u(\mathbf{y_i^v}) + \epsilon_i \tag{6.2}$$

Where each $\mathbf{y_i^v}$ is a vector of multi-dimensional outputs over a temporal window of $[i-v, i+v]$[2] The $\mathbf{y_i^v}$ features are called the *output features*, while $\mathbf{x}$ are called the *input features*, henceforth. Note that the output features can be estimated by predicting the multi-dimensional ground truth using any (noisy and imperfect) prediction scheme. The goal now becomes learning not only the set of weights ($\mathbf{w}$) for the input features, but also the set of weights ($\mathbf{u}$) for the output features along with the noise estimate, (ϵ_i).

In this section we describe the Bayesian framework that our model is based on. Firstly, we consider $\mathbf{\Phi_w}$ ($N \times M_w$) to be the basis matrix attained by applying a selected kernel to

[1] We denote that $\mathbf{t_i}$ is a multidimensional vector containing all the values to be predicted for each frame (in our case, both valence and arousal). Nevertheless, the methods we apply are inherently single output methods. Thus, a different function is learnt for each output dimension (t_i).

[2] For frame based online application, we can limit the context to past input only, i.e. $[i-v, i]$. Furthermore, the output window regards *only* the output dimensions since we study the effect of output-covariances.

the input features $\mathbf{x}$, and $\mathbf{\Phi_u^v}$ (N×M_u) respectively, for the output features $\mathbf{y^v}$ (M_u and M_w, referring to the number of basis vectors). Then, by extending Eq. 6.2 we obtain:

$$\mathbf{t} = \mathbf{\Phi_w w} + \mathbf{\Phi_u^v u} + \boldsymbol{\epsilon} = \mathbf{\Phi_{wu} w_u} + \boldsymbol{\epsilon} \tag{6.3}$$

where $\mathbf{\Phi_{wu}} = [\mathbf{\Phi_w}|\mathbf{\Phi_u^v}]$ is the N×$(M_w + M_u)$ OA-RVM design matrix:

$$\mathbf{\Phi_{wu}} = \begin{bmatrix} K_w(\mathbf{x_1}, \mathbf{x_1}) & \dots & K_w(\mathbf{x_1}, \mathbf{x_n}) & K_u(\mathbf{y_1^v}, \mathbf{y_1^v}) & \dots & K_u(\mathbf{y_1^v}, \mathbf{y_n^v}) \\ \vdots & & \vdots & \vdots & & \vdots \\ K_w(\mathbf{x_n}, \mathbf{x_1}) & \dots & K_w(\mathbf{x_n}, \mathbf{x_n}) & K_u(\mathbf{y_n^v}, \mathbf{y_1^v}) & \dots & K_u(\mathbf{y_n^v}, \mathbf{y_n^v}) \end{bmatrix}$$

with K_w and K_u being the kernel applied to input and output features respectively. Typically, an extra unit column is appended to the kernel to account for the bias. Furthermore, $\mathbf{w_u} = [\mathbf{w_1} \dots \mathbf{w_{M_w}}|\mathbf{u_1} \dots \mathbf{u_{M_u}}]^T$ represents the concatenated vector of weights.

Thus, the complete data set likelihood is formulated as:

$$P(\mathbf{t}|\mathbf{w}, \mathbf{u}, \sigma^2) = \prod_{i=1}^{N} N(\mathbf{w}^T \phi_w(\mathbf{x_i}) + \mathbf{u}^T \phi_u(\mathbf{y_i^v}), \sigma^2)$$

$$= \prod_{i=1}^{N} N(\mathbf{w_u}^T [\phi_w(\mathbf{x_i})|\phi_u(\mathbf{y_i^v})], \sigma^2) \tag{6.4}$$

Following the Bayesian approach of RVM [246], we need to set the hyperpriors on our weights. Each set of weights $(\mathbf{w}, \mathbf{u})$ is assigned a Gaussian zero-mean prior to express preference over smaller weights, thus infer smoother, less complex functions and induce sparsity:

$$P(\mathbf{w}|\boldsymbol{\alpha}) = \prod_{i=0}^{M_u} \mathcal{N}(0, \alpha_i^{-1}) \tag{6.5}$$

$$P(\mathbf{u}|\boldsymbol{\zeta}) = \prod_{i=1}^{M_w} \mathcal{N}(0, \zeta_i^{-1}) \tag{6.6}$$

We have now introduced two vectors of hyperparameters, $\boldsymbol{\alpha}$ controlling the distribution of the weights $\mathbf{w}$ (as originally used in RVM), and $\boldsymbol{\zeta}$ controlling the distribution of the weights $\mathbf{u}$ (for our output features).

6.3.1 Inference

The goal is to infer the unknown parameters of our problem given the training data. The posterior is decomposed as:

$$P(\mathbf{w}, \mathbf{u}, \boldsymbol{\alpha}, \boldsymbol{\zeta}, \sigma^{\mathbf{2}}|\mathbf{t}) = \frac{P(\mathbf{t}|\mathbf{w}, \mathbf{u}, \boldsymbol{\alpha}, \boldsymbol{\zeta}, \sigma^{\mathbf{2}})P(\mathbf{w}, \mathbf{u}, \boldsymbol{\alpha}, \boldsymbol{\zeta}, \sigma^{\mathbf{2}})}{p(\mathbf{t})} \tag{6.7}$$

Ideally, given a new test data x_*, we would like to predict target t_*:

$$p(t_*|\mathbf{t}) =$$

$$\int P(\mathbf{t}_*|\mathbf{w}, \mathbf{u}, \boldsymbol{\alpha}, \boldsymbol{\zeta}, \sigma^{\mathbf{2}})P(\mathbf{w}, \mathbf{u}, \boldsymbol{\alpha}, \boldsymbol{\zeta}, \sigma^{\mathbf{2}}|\mathbf{t})d\mathbf{w}d\mathbf{u}d\boldsymbol{\alpha}d\boldsymbol{\zeta}d\sigma^{\mathbf{2}} \tag{6.8}$$

Unfortunately, the above equation is intractable, thus an approximation is needed. Therefore, similarly to the original RVM formulation [246], we decompose the posterior as follows:

$$P(\mathbf{w}, \mathbf{u}, \boldsymbol{\alpha}, \boldsymbol{\zeta}, \sigma^{\mathbf{2}}|\mathbf{t}) = P(\mathbf{w}, \mathbf{u}|\mathbf{t}, \boldsymbol{\alpha}, \boldsymbol{\zeta}, \sigma^{\mathbf{2}})P(\boldsymbol{\alpha}, \boldsymbol{\zeta}, \sigma^{2}|\mathbf{t}) \tag{6.9}$$

Using the Bayes theorem we obtain:

$$P(\mathbf{w}, \mathbf{u}|\mathbf{t}, \boldsymbol{\alpha}, \boldsymbol{\zeta}, \sigma^{\mathbf{2}}) = \frac{P(\mathbf{t}|\mathbf{w}, \mathbf{u}, \sigma^{\mathbf{2}})P(\mathbf{w}, \mathbf{u}|\boldsymbol{\alpha}, \boldsymbol{\zeta})}{P(\mathbf{t}|\boldsymbol{\alpha}, \boldsymbol{\zeta}, \sigma^{\mathbf{2}})} \tag{6.10}$$

This calculation is tractable, since all components are Gaussian distributions and it is well known that products and divisions of Gaussian distributions result also in Gaussian distributions. We will firstly examine the joint probability. By assuming independence, we obtain $P(\mathbf{w}, \mathbf{u}|\boldsymbol{\alpha}, \boldsymbol{\zeta})$, a zero-mean Gaussian distribution with a covariance matrix $\mathbf{A_Z} = \text{diag}(\alpha_1 \dots \alpha_{M_w}, \zeta_1 \dots \zeta_{M_u})$.

$$P(\mathbf{t}|\boldsymbol{\alpha}, \boldsymbol{\zeta}, \sigma^{\mathbf{2}}) = \int P(\mathbf{t}|\mathbf{w}, \mathbf{u}, \sigma^{\mathbf{2}})P(\mathbf{w}, \mathbf{u}|\boldsymbol{\alpha}, \boldsymbol{\zeta})d\mathbf{w}d\mathbf{u} \tag{6.11}$$

is a convolution of Gaussian and after replacing with $\mathbf{w_u}$, $\mathbf{A_z}$ and $\mathbf{\Phi_{wu}}$, it can be shown [246] to be a zero-mean Gaussian distribution with covariance matrix $\sigma^2\mathbf{I} + \Phi_{wu}\mathbf{A_Z^{-1}}\Phi_{wu}^T$.

Finally, Eq. 6.10 is considered to be a Gaussian distribution with a mean $\boldsymbol{\mu} = \sigma^2\Sigma\mathbf{\Phi}_{wu}^T\mathbf{t}$ and a covariance matrix $\mathbf{\Sigma} = (\mathbf{A_Z} + \sigma^2\mathbf{\Phi}_{wu}^T\mathbf{\Phi}_{wu})^{-1}$.

Returning to the second component $P(\boldsymbol{\alpha}, \boldsymbol{\zeta}, \sigma^2|\mathbf{t})$ of the posterior in Eq. 6.9, by following the Bayes rule, we find it to be proportional to:

$$P(\boldsymbol{\alpha}, \boldsymbol{\zeta}, \sigma^2|\mathbf{t}) \propto P(\mathbf{t}|\boldsymbol{\alpha}, \boldsymbol{\zeta}, \sigma^2)P(\boldsymbol{\alpha})P(\boldsymbol{\zeta})P(\sigma^2) \tag{6.12}$$

By assuming uniform uninformative hyperpriors [246], we need to maximize the marginal likelihood, $P(\mathbf{t}|\boldsymbol{\alpha}, \boldsymbol{\zeta}, \sigma^{\mathbf{2}})$ with respect to the hyperparameters. Again, we have a convolution of Gaussians (Eq. 6.11) which in turn generates another zero mean Gaussian distribution with covariance matrix $\sigma^2\mathbf{I} + \boldsymbol{\Phi}_{wu}{\mathbf{A}_\mathbf{z}}^{-1}\boldsymbol{\Phi}_{wu}^T$. The maximization of the marginal likelihood can be performed by expectation maximization as described in [246] or the faster marginal maximization algorithm proposed in [249]. The most probable values (MP) are selected by the chosen optimization procedure ([246, 249]), while we adopt an approximation of $P(\boldsymbol{\alpha}, \boldsymbol{\zeta}, \sigma^2|\mathbf{t})$ in Eq. 6.9 by replacing the distribution with a delta function at its mode.

6.3.2 Prediction

Given a new (multi-dimensional) input data $\mathbf{x}_*, \mathbf{y}_*^\mathbf{v}$, we want to calculate t_* given the training data. By considering $\boldsymbol{\alpha}_\mathbf{z} = [a_1 \dots a_{M_w}, \zeta_1 \dots \zeta_{M_u}]$ and using Eq. 6.8 and Eq. 6.10 we obtain:

$$P(t_*|\mathbf{t}, \boldsymbol{\alpha}_{\mathbf{z}MP}, \sigma^2_{MP}) =$$

$$\int P(t_*|\mathbf{w_u}, \sigma^2_{MP}) P(\mathbf{w_u}|\mathbf{t}, \boldsymbol{\alpha}_{\mathbf{z}MP}, \sigma^2_{MP}) d\mathbf{w_u} \tag{6.13}$$

Again, this is a convolution of Gaussians and it can be shown that

$$P(t_*|\mathbf{t}, \boldsymbol{\alpha}_{\mathbf{z}MP}, \sigma^2_{MP}) \sim N(t_*|\sigma^2_*) \tag{6.14}$$

where

$$t_* = \boldsymbol{\mu}_{wu}^T [\phi_w(\mathbf{x}_*)|\phi_u(\mathbf{y}_*^\mathbf{v})] \tag{6.15}$$

$$\sigma^2_* = \sigma^2_{MP} + [\phi_w(\mathbf{x}_*)|\phi_u(\mathbf{y}_*^\mathbf{v})]^T \boldsymbol{\Sigma} [\phi_w(\mathbf{x}_*)|\phi_u(\mathbf{y}_*^\mathbf{v})] \tag{6.16}$$

with t_* being the test point prediction, and σ^2_* being the prediction variance (relating to *confidence* in the obtained prediction). The parameter vector $\boldsymbol{\mu}_{wu}$ contains the weights for the input and output relevance vectors, i.e. $\boldsymbol{\mu}_{wu} = [\boldsymbol{\mu}_w|\boldsymbol{\mu}_u]$. The basis matrix for a new set of test points should now contain both the distances from the new test input features $\mathbf{x}_*$ to all the input feature relevance vectors, as well as the test output feature $\mathbf{y}_*^\mathbf{v}$ distances to the output feature relevance vectors. The graphical model of OA-RVM with respect to the original RVM can be seen in Fig. 6.2. An overview of the OA-RVM training and prediction procedures is presented in Alg. 1.

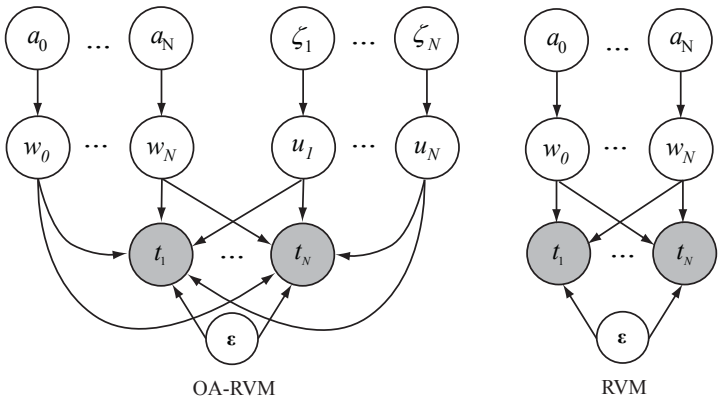


Figure 6.2: Graphical model comparison of RVM and OA-RVM. Shaded nodes are observed variables.

Algorithm 1 OA-RVM algorithm

Training. Data: $(\mathbf{x_i}, \mathbf{t_i})$, i=1,..., N
1. Obtain output features $\mathbf{y_i^v}$
2. Construct basis matrix $\mathbf{\Phi_{wu}} = [\mathbf{\Phi_w}|\mathbf{\Phi_u^v}]$
2a. Apply kernel K_w for obtaining $\mathbf{\Phi_w}$ for input features $\mathbf{x}$
2b. Apply kernel K_u for obtaining $\mathbf{\Phi_u^v}$ for output features $\mathbf{y^v}$
3. Marginal Likelihood Maximization
3a. Determine hyperparameters $(\boldsymbol{\alpha}, \boldsymbol{\zeta}, \sigma^2)$
3b. $\boldsymbol{\mu} = \sigma^2 \Sigma \mathbf{\Phi}_{wu}^T \mathbf{t}$, $\mathbf{\Sigma} = (\mathbf{A_Z} + \sigma^2 \mathbf{\Phi}_{wu}^T \mathbf{\Phi}_{wu})^{-1}$
Prediction for test point $\mathbf{x}_*$:
1. Obtain output features $\mathbf{y_*^v}$
2. Predict and estimate variance:
2a. $t_* = \boldsymbol{\mu}_{wu}^T [\phi_w(\mathbf{x}_*)|\phi_u(\mathbf{y_*^v})]$
2b. $\sigma_*^2 = \sigma_{MP}^2 + [\phi_w(\mathbf{x}_*)|\phi_u(\mathbf{y_*^v})]^T \mathbf{\Sigma} [\phi_w(\mathbf{x}_*)|\phi_u(\mathbf{y_*^v})]$

6.3.3 Window Size

The output feature window length v for OA-RVM is treated as a regular parameter in the framework. Therefore, many heuristics and validation techniques can be employed in order to define the parameter for a given training set. The most direct method would be to perform cross-validation (i.e. similarly to SVM) in order to determine the optimal value for the specific error metric employed. Another way is to compare the maximised marginal likelihood of each model trained with a specific window size (i.e. a Maximum Likelihood test). Assuming we have a set V of windows to be evaluated, for each $v_i \in V$ the marginal likelihood $L_{v_i} \sim N(0, \sigma^2\mathbf{I} + [\mathbf{\Phi_w}|\mathbf{\Phi_u^{v_i}}]\mathbf{A_z}^{-1}[\mathbf{\Phi_w}|\mathbf{\Phi_u^{v_i}}]^T)$ is maximized. The window size providing the maximum likelihood can then be selected, i.e. $v = \arg\max_{v_i} L_{v_i}$.

6.3.4 A Generalised View

In this section we aim to provide a more general perspective of the proposed framework while comparing it to other static regression frameworks (e.g. SVM and RVM).

In a typical static regression framework (e.g. SVM and RVM), we consider only the current input to participate in the prediction, i.e.

$$P(t_i|\mathbf{x_1} \dots \mathbf{x_i} \dots \mathbf{x_N}) = P(t_i|\mathbf{x_i})$$

In the proposed framework, each prediction not only depends on the current input but also on the output features, which essentially represent a *temporal* noisy version of the targets to be estimated:

$$P(t_i|\mathbf{x_1} \dots \mathbf{x_i} \dots \mathbf{x_N}) = P(t_i|\mathbf{x_i}, \mathbf{y_i^v})$$

The output features $\mathbf{y_i^v}$ represent a noisy prediction of the targets over time (a pre-defined temporal window). Therefore,

$$P(t_i|\mathbf{x_1} \dots \mathbf{x_i} \dots \mathbf{x_N}) = P(t_i|\mathbf{x_i}, \hat{\mathbf{t}}_{i-v}, \dots, \hat{\mathbf{t}}_i, \dots, \hat{\mathbf{t}}_{i+v})$$

where each $\hat{\mathbf{t}}_i$ is the noisy prediction of $\mathbf{t_i}$ at input datum i. The prediction is thus conditioned both on the current input frame, as well as the noisy prediction of the multi-dimensional targets over the specified temporal window.

Conditioning on the intermediate noisy predictions can be considered as a form of *ensemble learning*, specifically of *stacked generalisation* [276, 31] with continuous labels. A specific stacked generalisation algorithm could also be investigated for training OA-RVM to obtain insight on its benefits for method generalisation.

6.3.5 Complexity

The optimisation algorithm of RVM generally involves the optimisation of a non-convex function. The inversion of an $M \times M$ matrix is required, where M is the number of basis vectors in the model, thus inducing $O(M^3)$ computational complexity. In OA-RVM, without loss of generality, we assume that we have $2M$ basis vectors: A dimensionality of M for the input features and an additional M for the output features. Thus, the complexity is $O((2M)^3) = O(M^3)$. Furthermore, the output features in OA-RVM are obtained by utilising the original RVM algorithm. If for a d-dimensional output problem, the complexity of the original RVM algorithm is $O(dC)$, then for OA-RVM the complexity would be $2O(dC)$ which is still $O(dC)$. In conclusion, the theoretical complexity of OA-RVM is of the same order as RVM. Nevertheless, in

practise OA-RVM has a higher computational complexity than RVM, since it involves executing the original RVM algorithm as well as OA-RVM, which implies an augmented kernel with twice the number of candidate basis vectors compared to RVM.

6.4 Dataset and Feature Extraction

As a proof of concept, the proposed OA-RVM regression is applied to the highly challenging problem of *dimensional and continuous prediction* of emotions from heterogeneous nonverbal cues, namely facial expressions, shoulder movements and audio cues. Our aim is to explore how the behavior of the OA-RVM model changes (in terms of prediction accuracy and spatio-temporal structure) depending on the expressive cue / modality employed.

6.4.1 Data Set

For experimental validation we use the Sensitive Artificial Listener (SAL) Database [64]. As described in Chapter 2, SAL contains audio-visual, naturalistic affective conversational data taking place between a participant and an avatar (operated by a human). Similarly to Chapter 5, as our aim is to achieve continuous emotion prediction, we could only take advantage of the amount of data which was annotated in the *valence-arousal dimensional affect space*. This corresponds to a portion of the database that contains data from 4 subjects (subjects 1 and 2 are female, and subjects 3 and 4 are male) and their respective annotations (provided by 3-4 annotators). Based on the annotations provided, we used a set of automatic segmentation and ground truth generation algorithms [173] that generates segments of positive/negative emotional displays. More specifically, we generated segments capturing transitions to an emotional state and back (e.g., going from non-positive to positive and back to non-positive). Henceforth, we refer to these classes as positive for the transition to a positive emotional state, and negative for the transition to a negative emotional state. In total, we used 61 positive and 73 negative segments, and approximately 30,000 video frames.

6.4.2 Facial Expressions

For extracting facial expression features, we employ the Patras - Pantic particle filtering tracking scheme [190] for tracking the facial feature movements displayed during the naturalistic interactions. We track the corners of the eyebrows (4 points), the eyes (8 points), the nose (3 points), the mouth (4 points) and the chin (1 point). For each video segment containing n frames, the tracker results in a feature set with dimensions $n * 20 * 2$. We register each set of points in a given frame to the corresponding coordinate system centred at the fixed point of

the face (the average of the inner eye points and the tip of the nose). We thus end up with a simple translation applied to every point in every frame (also using the fixed point itself as a feature). Fig. 2.7(a) shows examples from the data set employed together with the tracking of the facial feature points.

6.4.3 Shoulder Movements

The motion of the shoulders is captured by tracking 2 points on each shoulder and one stable point on the torso, usually just below the neck (see Fig. 2.7(b)). We initialize the tracked points in the first frame of each sequence manually, while the standard Auxiliary Particle Filtering (APF) [199] is subsequently used to track the shoulder points. This scheme is less complex and faster compared to the Patras - Pantic particle filtering tracking scheme, it does not require learning the model of prior probabilities of the relative positions of the shoulder points, while resulting in sufficiently high accuracy. For each video segment containing n frames, the tracker results in a feature set with dimensions $n * 5 * 2$.

6.4.4 Acoustic Features

Utilised acoustic features include Mel-frequency Cepstrum Coefficients (MFCC, MFCC-Delta) [115] and prosody features (the energy of the signal, the Root Mean Squared Energy and the pitch obtained by using a Praat pitch estimator [191]).

We used 6 cepstrum coefficients, thus obtaining 6 MFCC and 6 MFCC-Delta features for each audio frame. We have essentially extracted the typical set of features used by other works (e.g., [196]) for automatic affect recognition. Along with pitch, energy and RMS energy, we obtained a set of features with dimensionality 15 (per audio frame).

6.5 Why Output-Association for Continuous Emotion Prediction?

In this section, we demonstrate how the proposed OA-RVM regression framework is efficiently applicable to the problem of automatic emotion prediction in a continuous dimensional space. We focus our analysis and discussion on Fig. 6.3. The figure illustrates how employing the original RVM and the proposed OA-RVM provides continuous prediction of valence and arousal dimensions for one training sequence (consisting of 315 frames).

The predictions generated by RVM are shown in Fig. 6.3(a,b) while the OA-RVM generated

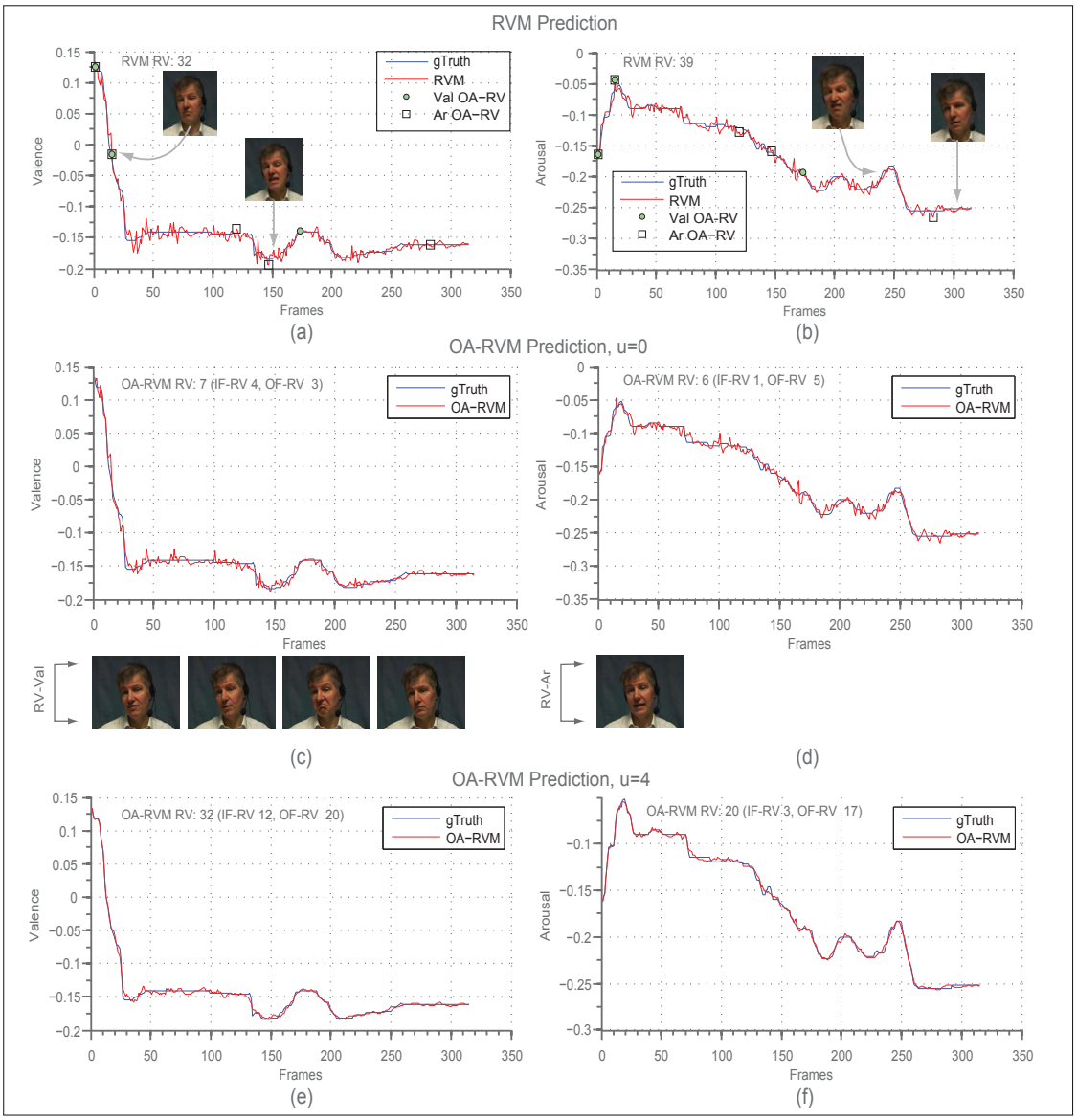


Figure 6.3: Illustration of how employing the original RVM and the proposed OA-RVM provide continuous prediction of valence and arousal dimensions for one training sequence (315 frames). (a,b) RVM prediction with RVs used for OA-RVM, (c,d) OA-RVM prediction with a window of $v = 0$ and IF-RV frames, and (e,f) OA-RVM with prediction with a window of $v = 4$.

.

predictions with a window of $v = 0$ and $v = 4$ are shown in Fig. 6.3(c,d) and Fig. 6.3(e,f), respectively. The ground truth for both the valence and the arousal dimensions is shown in all figures as *gTruth*, for comparison purposes. The generated predictions for valence appear on the left column of Fig. 6.3, while the generated predictions for arousal appear on the right. The window of $v = 0$ is meant to represent the most sparse results, while a window of $v = 4$ is deemed sufficient for a sequence of 315 frames as it embeds 9 temporal steps (frames) in terms of past (4 frames), present (current frame) and future (4 frames) context.

In this particular sequence, the subject appears to be displaying negatively *valenced* emotions (e.g., sadness, disappointment), with a decreasing arousal over time (towards a more passive emotional state). In the figure we observe how the RVM framework generates predictions (depicted with RVM line) by using 32 relevance vectors (RVs) for valence (Fig. 6.3a) and 39 RVs for arousal (Fig. 6.3b). Fig. 6.3(c,d) then illustrates how the proposed OA-RVM framework generates predictions for the sequence at hand, for valence and arousal, with a temporal window of $v = 0$. Note how OA-RVM is able to learn a *smoother* and *more accurate* model by using 7 RVs for valence and 6 RVs for arousal, respectively.

As specified in Eq. 6.2, OA-RVM depends on both the input features ($\mathbf{x}$, depicted as *IF* in the figure) and the output features ($\mathbf{y^v}$, depicted as *OF* in the figure). To illustrate the behavior of the framework, we decompose the relevance vectors (RVs) selected by OA-RVM into the RVs centered around the input features (RV-IF) and the RVs centred around the output features (RV-OF).

For the valence dimension, the 7 RVs used for the OA-RVM model can be decomposed into 4 RVs corresponding to input features (the relevant frames shown in Fig. 6.3c) and 3 RVs corresponding to output features (shown in Fig. 6.3(a,b) as *Val OA-RV*). A similar analysis is performed for the arousal dimension. For the sequence at hand, in Fig. 6.3d we can see that 6 RVs are needed by OA-RVM. Note how for this prediction, only one input feature RV is used by OA-RVM. The remaining 5 RVs centered around the output features are depicted in Fig. 6.3(a,b) as *Ar OA-RV*. An interesting observation is that, in frame 1 and frame 15 the arousal begins to decrease, and is accompanied by a change of sign in the valence dimension. The OA-RVM framework is able to capture this in its valence and arousal prediction via two common RVs centred around the output-features in frame 1 and frame 15.

To conclude this section, in Fig. 6.3(e,f), we show the results of applying OA-RVM with a temporal window of $v = 4$ (Eq. 6.2). Note how the learned OA-RVM model provides a nearly perfect fit by using no more RVs than the original RVM model (from which the output

features are generated). Yet, both the MSE and COR metrics are improved. Although the complexity of the model is observed to increase with an increase in the window size, overall, the OA-RVM model appears to generalise to new data very well while avoiding overfitting.

6.6 Experimental Evaluation

Experimental Setting

We apply the proposed OA-RVM regression to the highly challenging problem of *dimensional and continuous prediction* of emotions from heterogeneous nonverbal cues, namely facial expressions, shoulder movements and audio cues. Our aim is to conduct comprehensive experiments in order to explore how the behavior of the OA-RVM model changes (in terms of prediction accuracy and spatio-temporal structure) depending on the expressive cue / modality employed.

We use the traditional RVM as the baseline for our comparisons with OA-RVM. We also use SVR as it is one of the most widely adopted regression techniques in the field. The kernel used for the construction of the basis matrices is a Gaussian, $K(x, x_i) = exp\left\{(-(\mathbf{x} - \mathbf{x_i})^2)/r^2\right\}$ where r stands for the width of the function. The window parameter v in the output-associative functional we employ (Eq. 6.1) is generally varied in the range $[0 - 30]$ and determined by cross-validation. It should be noted that for the probabilistic regression methods (RVM, OA-RVM), the hyperparameters are determined by optimizing the likelihood function (by using fast marginal likelihood maximization algorithm proposed in [249]). We use RVM to obtain the initial (noisy) output estimation (i.e., the output features) for OA-RVM. For SVR we apply cross-validation employing an ϵ-insensitive loss function.

In our current setting, we assume that the segments contained in our data set have been coarsely classified into either positive or negative, prior to the prediction (regression) procedure[3]. The classification stage is beyond the scope of this chapter, and can be achieved by applying an accurate (coarse) classifier, e.g. [172], as the basis for the current framework. This assumption is motivated by the fact that we would like to focus on the prediction results in detail, and study them in isolation for each class (e.g., which dimension is easier to predict for which class). Based on the aforementioned assumptions, we conduct experiments using *subject-independent* cross-validation, where we train the model using data from three subjects and evaluate the trained model using the data from the subject left out for evaluation. Results are averaged across four-fold subject-independent cross-validation.

[3]Note that each sequence usually contains some portion of both positively / negatively *valenced* frames.

Note that subject-independent evaluation using this database is considered highly challenging [151] as annotated data is only available for a small number of subjects. More specifically, during training, the model is able to learn only a limited (subject-specific) subspace of the human affective variability. Moreover, performing regression in a continuous space (rather than classification into a predetermined set of labels) poses additional challenges.

As evaluation metrics we use both the root mean squared error (RMSE) and the correlation (COR) between the prediction and the ground truth values. RMSE evaluates the prediction by taking into account the squared error of the prediction from the ground truth. As discussed in Chapter 5, the RMSE, which represents the bias error and variance of the prediction, can be misleading with regards to how realistic the prediction of a regression technique can be. The correlation coefficient (COR) provides an evaluation of the linear relationship between the prediction and the ground truth, and subsequently, an evaluation of whether the model has managed to capture the linear structural patterns inhibited in the data at hand.

Experimental Results and Analysis

In this section, we will discuss the results of the proposed OA-RVM model, focusing on prediction accuracy as evaluated by the root mean squared error (RMSE), presented in Table 6.1, and the correlation coefficient (COR) presented in Table 6.2.

Table 6.1: Evaluating SVM, RVM and OA-RVM for the task of predicting arousal and valence from face, shoulder and audio cues. Results are averaged across four-fold subject-independent cross-validation. The evaluation is based on the Root Mean Squared Error (RMSE).

		Valence			Arousal		
Class	Cue	SVM	RVM	OA-RVM	SVM	RVM	OA-RVM
POS	Face	0.200	0.166	**0.160**	0.157	0.166	**0.147**
	Shoulders	0.257	0.177	**0.171**	0.164	0.146	**0.132**
	Audio	0.176	0.179	**0.171**	0.146	0.144	**0.130**
NEG	Face	0.150	0.940	**0.088**	0.365	0.374	**0.342**
	Shoulders	0.110	0.103	**0.097**	0.355	0.371	**0.354**
	Audio	0.101	0.102	**0.097**	0.339	0.339	**0.300**

Firstly, we observe that for both emotion dimensions and classes, OA-RVM outperforms RVM and SVM in terms of both COR and RMSE. The improvement is especially noticeable in terms of COR rather than RMSE. This can be justified by the fact that the goal of OA-RVM is to enforce common, temporal output patterns, thus increasing the covariance of the prediction with the ground truth. The prediction results provided by SVR and RVM are

Table 6.2: Evaluating SVM, RVM and OA-RVM for predicting arousal and valence from face, shoulder and audio cues. Results are averaged across four-fold subject-independent cross-validation. Errors obtained by evaluating the correlation coefficient (COR) of the prediction to the ground truth.

		Valence			Arousal		
Class	Cue	SVM	RVM	OA-RVM	SVM	RVM	OA-RVM
POS	Face	0.28	0.30	**0.43**	0.09	0.09	**0.16**
	Shoulders	0.01	0.16	**0.32**	0.12	0.19	**0.30**
	Audio	0.02	0.03	**0.19**	0.04	0.07	**0.21**
NEG	Face	0.14	0.20	**0.27**	0.13	0.18	**0.27**
	Shoulders	0.14	0.28	**0.29**	0.09	0.09	**0.22**
	Audio	0.01	0.05	**0.10**	0.23	0.23	**0.38**

fairly similar, with RVM in general achieving better correlation with the ground truth. In what follows, we discuss the results for OA-RVM.

Focusing on the RMSE results of each class in isolation, we denote that for the positive class arousal appears to be easier to predict than valence. Nevertheless, for the same class, the COR achieved is higher for valence, showing that the structure of the valence dimension is modelled more accurately. When analysing the results obtained for the negative class we observe that valence prediction is better than arousal prediction. In fact, considering the RMSE metric, arousal prediction for the negative class appears to be the most challenging case for OA-RVM prediction framework.

Let us now compare the results obtained by employing different sets of nonverbal cues. When utilising the facial expression cues, the correlation between the prediction and the ground truth appears to be equivalent for both emotion dimensions. In general, correlation obtained for the negative class appears to be highly dependent on the set of cues employed.

Related work on dimensional emotion recognition reported that arousal can be much better predicted than valence using audio cues [151], [91], [251]. Results obtained from our experiments are in line with such findings, showing that audio cues appear to provide the best prediction results (in terms of RMSE) for the arousal dimension. When considering the COR metric and the negative class, audio cues *again* appear to provide the best prediction results (0.38) compared to facial expression (0.27) and shoulder cues (0.22). For the positive class, while the audio cues still provide better correlation compared to using the facial expression cues, the shoulder cues appear very capable of capturing the arousal structure (and perform better than the audio cues).

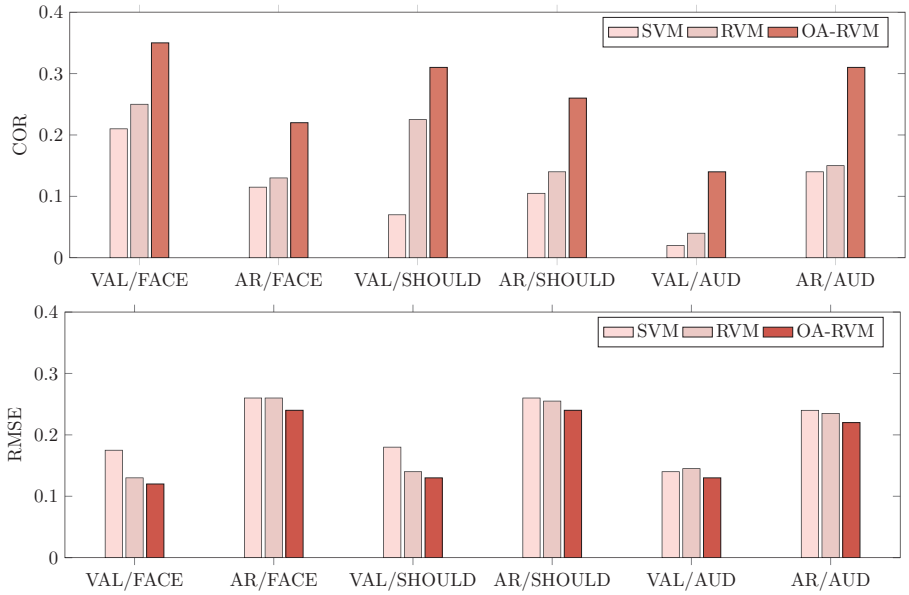


Figure 6.4: Comparing the prediction correlation (COR) and root mean squared error (RMSE) for Valence (VAL) and Arousal (AR), when utilising facial expressions (FACE), shoulder movement (SHOULD) and audio cues (AUD).

It is well known that the facial expression cues are very informative for predicting valence. Our RMSE-based results confirm this, utilizing the facial expression cues provides better prediction results for the valence dimension. The shoulder cues also appear to be better at capturing useful information regarding the valence dimension compared to the audio cues.

When evaluating the *valence prediction models* in terms of the correlation metric, the models trained using the visual cues in general appear to perform better than the models trained using the audio cues (see Table 6.2). Additionally, for the negative class, the prediction models trained on the shoulder cues appear to slightly outperform the models trained on the facial expression cues.

In Fig. 6.4, we illustrate the average results for both classes evaluated in terms of RMSE and COR. Overall, we observe that regardless of the set of cues utilized or dimensions predicted, there is a significant increase in terms of correlation when applying OA-RVM. As denoted earlier, compared to OA-RVM and RVM, SVM provides the lowest correlation. Additionally, it can be seen that prediction models trained with facial expressions provide the lowest RMSE for the valence dimension, and the prediction models trained using the audio cues provide the lowest RMSE for the arousal dimension.

In Fig. 6.5 we also provide an illustrative comparison between the predictions generated by

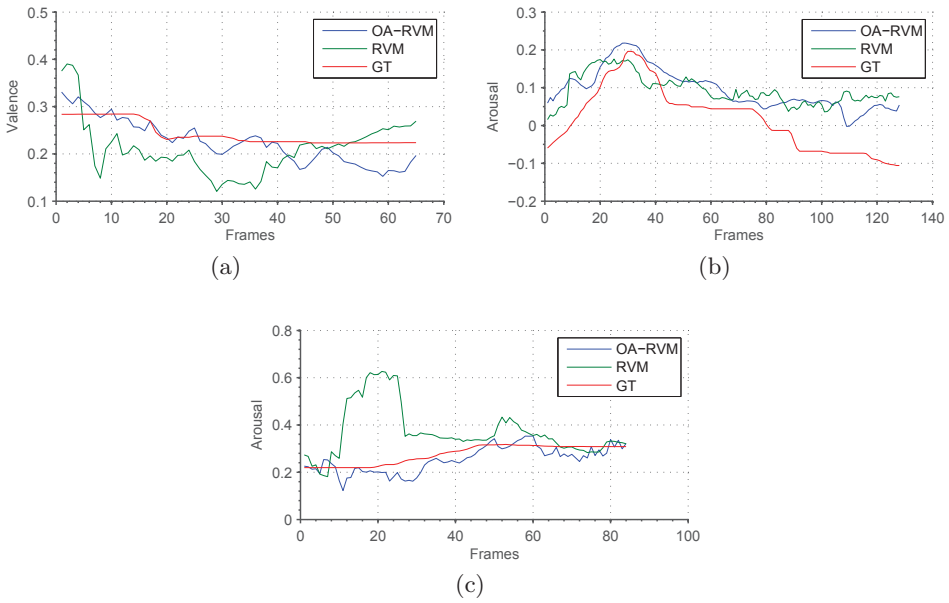


Figure 6.5: An illustrative comparison between the predictions generated by OA-RVM and RVM, on test data, with respect to the ground truth, utilizing different cues: (a) facial expressions, (b) shoulder movements, and (c) audio cues.

OA-RVM and RVM, on test data, with respect to the ground truth (utilizing different cues). Overall, naturalistic emotional expressions are highly subject-dependent [95]. However, from our experiments we conclude that automatic, subject-independent, dimensional and continuous prediction of emotions becomes feasible by utilising input and output associations as well as temporal context.

Psychological research findings suggest that there exist gender-related differences in expressing emotions (e.g., women appear to be more facially expressive than men [124]). In our experiments we found no consistent differentiations between male and female subjects. However, the data set employed is relatively small in order to draw generalising conclusions regarding gender-related differences.

6.7 Conclusions and Discussion

In this chapter, we proposed a novel Output-Associative Relevance Vector Machine (OA-RVM) regression framework that augments traditional RVM by being able to learn *non-linear input-output dependencies*. Instead of depending solely on the input patterns, OA-RVM models output structure and covariances within a predefined temporal window, thus capturing past

and future context. We successfully applied the proposed framework for dimensional and continuous prediction of emotions from heterogeneous nonverbal cues (facial expressions, shoulder movement and audio cues) and demonstrated its advantages and efficiency over a comprehensive set of experiments using subject-independent cross-validation. Our experimental results show that:

- OA-RVM outperforms both RVM and SVR in terms of prediction accuracy (RMSE) and prediction structure (COR), regardless of the set of cues utilized or emotion dimensions predicted. Employing a temporal (output) window, which induces the learning of past and future context, contributes significantly to the prediction accuracy. The size of the optimal temporal window may vary depending on the task and the data at hand.
- Although there is an inherent, subject-dependent characteristic attributed to naturalistic emotional expressions; automatic, subject-independent, dimensional and continuous prediction of emotions is possible by utilising input and output associations, and temporal context.

Chapter 7

Correlated-Spaces Regression for learning continuous emotion dimensions

Contents

7.1 Introduction

Motivated by the challenges arising from adopting continuous emotion dimensions, the focus of this chapter can be separated into three parts. Firstly, in Section 7.3.1 and 7.3.2 we provide empirical, quantitative results on several important questions related to the correlations of emotion dimensions. Secondly, in Section 7.3.3, we study the Level of Interest (LOI) as a continuous dimension, and evaluate the relationship of interest to other emotion dimensions.

Finally, in Section 7.4 we present a regression algorithm which correlates both labels and multi-modal features by projecting them on a common space, eliciting an elegant framework for multi-modal fusion, dimensionality reduction and output-correlations learning. Finally, conclusions are drawn in Section 7.5. In the remainder of the introduction, we discuss the chapter organisation in more detail.

Analysing emotion dimension correlations. The occurrence of inter-correlations amongst emotion dimensions such as valence and arousal has been well-supported by various research in psychology [129], and has recently been explored in affective computing in terms of valence and arousal (Chapter 6, [176]). Nevertheless, to the best of our knowledge, none of the previous work studies (i) correlation between emotion dimensions in isolation, (i.e. without including features), and (ii) the correlations of emotion dimensions to *basic emotions* such as joy and sadness. Furthermore, most works only employ valence and arousal without addressing dimensions such as power and expectation. We address all of these points in this work. Firstly, by using a set $\mathbf{R}_s$ of 5 dimensions (Valence, Arousal, Power, Expectation and Intensity) [157], in our first experiment (Section 7.3.1), we essentially pose the problem of predicting dimension k given the rest. We also perform experiments using facial/acoustic features for comparison. Interestingly enough, we show that the correlation of the $k-1$ other dimensions to dimension k is *higher* than the correlation of audio/face features to dimension k.

In our second experiment (Section 7.3.2), we attempt to answer an interesting question which has not been explored so far: *how correlated are emotion dimensions to basic emotions?* Intuitively, the correlation should be high, since in theory there is a (rather abstract and relatively ambiguous) mapping from these dimensions to basic emotions (e.g., high valence, positive arousal can point to joy, excitement etc.). To verify this intuition empirically, we use a set of basic emotions $\mathbf{L}_s$ (e.g., anger, happiness). Using the set of dimensions $\mathbf{R}_s$, we evaluate how correlated the emotion dimensions are to basic emotions, in comparison to facial points and audio cues. Our findings are in line with the previous experiment: Emotion dimensions are positively correlated with the intensity of basic emotions, exhibiting higher correlations than facial/acoustic features.

Continuous Interest & Emotion Dimensions. Evidence from the field of psychology

points to various correlations between emotion dimensions and interest [130]. Nevertheless, this has remained unexplored in the field of affective computing and machine learning, mostly due to the fact that interest has been regarded as a discrete emotion rather than a latent dimension. In this chapter, we model interest as a continuous dimension. In more detail, in Section 7.3.3, we provide, to the best of our knowledge, the first[1] empirical experimental evidence on continuous annotations which show that interest is highly correlated with specific emotion dimensions such as arousal, valence and intensity. Furthermore, our analysis reveals that although we use a disjoint set of annotators for interest, correlations between interest and other emotion dimensions are still high, thus motivating the utilisation of models exploiting output-correlations for detecting interest (c.f., Chapters 5, 6, [13, 176]).

Exploiting emotion dimension correlations. An important contribution lies in the introduction of the Correlated-Spaces Regression (CSR), a principled, novel framework based on canonical correlation analysis, which elegantly combines multi-modal fusion, the learning of output-correlations and supervised dimensionality reduction. Presented in Section 7.4, the proposed algorithm, heavily motivated by conclusions drawn from our empirical study, is shown to increase the accuracy of both single-cue and fused experiments and up to a point, "heal" the relatively weak correlation of facial/acoustic features to the emotion dimensions[2].

7.2 Data & Feature Extraction

Semaine Database. For evaluation, we employ the SEMAINE database [157], which contains a set of audio-visual recordings of subjects interacting with operators. As described in Chapter 2, each operator assumes a certain personality, i.e. happy, gloomy, angry and pragmatic, with a goal of inducing spontaneous emotions during a naturalistic conversation. We use a portion of the database running approximately 85 minutes, which has been annotated for the emotion dimensions at hand by 5 raters, from which we use the averaged annotation[3]. Furthermore, following the procedure in the next section, we obtained interest annotations

[1] At time of publication of [177].

[2] Regarding dimensionality reduction for regression, c.f. [119].

[3] For the basic emotion experiments, we use only the subset of this data which was annotated in terms of basic emotions.

from 8 annotators. For extracting facial expression features, we employ an Active Appearance Model (AAM) based tracker [182], designed for simultaneous tracking of 3D head pose, lips, eyebrows, eyelids and irises in video sequences. For each video frame, we obtain 113 2D-points, resulting in an 226 dimensional feature vector. To compensate for translation variations, we center the coordinate system to the fixed point of the face (average of inner eyes and nose), while for scaling we normalise by dividing with the inter-ocular distance. Regarding acoustic features, we utilise MFCC and MFCC-Delta coefficients along with prosody features (energy, RMS Energy and pitch). We used 13 cepstrum coefficients for each audio frame, essentially employing the typical set of features used for automatic affect recognition [285]. We obtain a feature vector with dimensionality $d = 29$, obtaining a frame-rate equivalent to 100-fps. To match the video fps, the acoustic features used are vertically concatenated for each pair of consecutive frames, thus obtaining 58 dimensional feature vectors. For feature-level fusion, the vectors are concatenated, resulting to 284 dimensions.

Obtaining Interest Annotations. In this section, we detail the process which we followed in order to obtain continuous interest annotations. Firstly, the instructions given to the annotators were based on earlier work [227], and have been readjusted in order to fit to a continuous scale and enriched in order to correspond to the conversational setting of the SEMAINE database. They are as follows:

- *Interest Rating in* $[-1, -0.5)$: the subject is *disinterested* in the conversation, can be mostly passive or appear bored, does not follow the conversation and possibly wants to stop the session.
- *Interest Rating in* $[-0.5, 0)$: the subject appears passive, replies to the interaction partner, possibly with hesitation, just because he/she has to reply (unmotivated). The subject appears *indifferent*.
- *Interest Rating approx.* 0: the subject seems to follow the conversation with the interaction partner, but it can not be recognized if he/she is interested. The subject is *neutral*.
- *Interest Rating in* $(0, 0.5]$: The subject seems eager to discuss with the interaction partner, and interested in getting involved in the conversation. The subject is *interested*.
- *Interest Rating in* $(0.5, 1]$: The subject seems pleased to participate in the conversation,

can show some signs of *enthusiasm*, is expressive in terms of (positive) emotions (e.g., laughing at a joke, curious to discuss a topic).

7.3 Analysis of Emotion Dimensions and Interest

In this section we present several experiments evaluating the correlations of emotion dimensions. For regression, we employ the Relevance Vector Machine (RVM [246]), which given the input-output pair $(\mathbf{x}_i, \mathbf{y}_i)$ models the function $\mathbf{y}_i = \mathbf{w}^T\phi(\mathbf{x_i}) + \epsilon_i, \epsilon_i \sim \mathcal{N}(0, \sigma^2)$ with $\phi(\mathbf{x}_i, \mathbf{x}_j) = \exp\left\{-\frac{||\mathbf{x}_i - \mathbf{x}_j||}{l}\right\}$ being the RBF kernel. Using the extracted features and annotations (Section 7.2) we perform cross-validation. For evaluation, we use the mean-squared error (MSE) to measure bias error and the correlation coefficient (COR) to measure the correlation deviation. We mostly refer to COR, since (i) it is most commonly used in related work [229], and (ii) the MSE bias errors are relatively very small.

7.3.1 Inter-Correlations and Multimedia

In this section we pose the problem of predicting an emotion dimension given a set of annotated dimensions. Let us assume we have a set of ρ annotations $\mathbf{R} = \{\mathbf{r}_1, \ldots, \mathbf{r}_\rho\}$ with $\mathbf{r}_i \in \mathbb{R}^{1\times T}$. In this experiment, we assume that $\mathbf{R}$ consists of dimensions valence, arousal, power, expectation and intensity, i.e. $\rho = 5$. Our problem can then be defined as

$$f : \mathbf{R}_{\setminus k} \rightarrow \hat{\mathbf{r}}_k, \forall k \in \{1, \ldots, \rho\} \tag{7.1}$$

where $\mathbf{R}_{\setminus k}$ denotes the entire set of annotations excluding dimension k and $\hat{\mathbf{r}}_k$ the estimated values of dimension k. The performance of the learnt functions is then compared against the performance obtained when using facial expressions and audio cues as features, in order to obtain a comparative measure of performance. By this experiment, we essentially ask the following question: *Which signal is most correlated with a specific emotion dimension k, the features extracted from audio/video cues or the annotations for the rest of the dimensions, $\mathbf{R}_{\setminus k}$?* Results are presented in Table 7.1 and Fig. 7.2. It is very interesting to observe that by using all the emotion dimensions except the one being tested provides better results for *all* dimensions at hand. This important observation empirically confirms that each and every emotion dimension has higher correlation with the rest of the dimensions than with the audio/face

Table 7.1: Results for predicting each emotion dimension, using the other four dimensions as features ($\mathbf{R}_s$), compared to using facial features (F), acoustic features (A) and the feature-level fusion of face and audio (F+A). Results shown are based on the Mean Squared Error (MSE) and the correlation coefficient (COR).

	Valence		**Arousal**		**Power**		**Expectation**		**Intensity**	
	MSE	COR	MSE	COR	MSE	COR	MSE	COR	MSE	COR
$\mathbf{R}_{s\setminus k}$	0.074	**0.28**	0.051	**0.47**	0.088	**0.28**	0.037	**0.15**	0.067	**0.30**
Face	0.088	0.14	0.061	0.41	0.131	0.06	0.024	0.02	0.066	0.17
Audio	0.072	0.14	0.050	0.44	0.082	0.05	0.018	0.01	0.042	0.26
F+A	0.880	0.16	0.055	0.44	0.080	0.06	0.020	0.02	0.058	0.20

features. It is also interesting to observe that for the arousal and the intensity dimensions, the audio cues appear to perform better than the facial features in terms of correlation coefficient, a conclusion that confirms previous findings (c.f., Chapter 5 and [174]).

7.3.2 Correlations to Basic Emotions

Another question we address in this work refers to the correlations amongst the dimensional emotion descriptions, as perceived by Russel [216] and a set of emotions which are of discrete nature (e.g., basic emotions). Although emotion dimensions can be inherently more expressive in comparison to discrete emotions such as joy and sadness, no explicit mapping between the two descriptions has been established. One would of course assume that e.g., negative valence with negative arousal maps to sadness or boredom, nevertheless this is more of an abstract and relatively ambiguous correspondence. In this section we evaluate the correlations of emotion dimensions when learning to predict emotions such as anger, happiness, sadness, surprise etc. In more detail, given the set $\mathbf{R}$, as defined in Section 7.3.1 (consisting of dimensions valence, arousal, power, expectation and intensity) we aim to predict a specific emotion belonging in the set $\mathbf{L} = \{\mathbf{l}_1, \ldots, \mathbf{l}_\nu\}$, i.e.

$$f : \mathbf{R} \rightarrow \hat{\mathbf{L}}_k, \forall k \in \{1, \ldots, \nu\} \tag{7.2}$$

Results are presented in Tab, 7.2 and Fig. 7.2, where we also use face/acoustic features for comparison. The first conclusion is that the emotion dimensions (namely valence, arousal, power, expectation and intensity) are highly correlated with the discrete emotions we study. Similarly to the results regarding the previous experiment, the dimension to discrete-emotion

Table 7.2: Predicting each basic emotion using the five emotion dimensions as features ($\mathbf{R}_{s\setminus k}$), compared to using facial features (F) and acoustic features (A). Results shown are based on the Mean Squared Error (MSE) and the correlation coefficient (COR).

COR	Anger	Happiness	Sadness	Contempt	Amusement
$\mathbf{R}_s$	**0.74**	**0.48**	**0.67**	**0.33**	**0.49**
F	0.06	0.11	0.13	0.05	0.06
A	0.02	0.10	0.10	0.11	0.02
MSE	**Anger**	**Happiness**	**Sadness**	**Contempt**	**Amusement**
$\mathbf{R}_s$	**0.07**	**0.10**	**0.06**	**0.02**	**0.07**
F	0.21	0.21	0.26	0.34	0.15
A	0.17	0.17	0.10	0.21	0.09

correlation is quite higher compared to face or acoustic features. The most correlated discrete emotion to emotion dimensions appears to be anger.

7.3.3 Interest and Emotion Dimensions

In this section, we attempt to empirically evaluate the correlation of interest with other emotion dimensions. The question is of high interest for many algorithms which aim to model output-structure (e.g., Chapters 5 and 6, [13]). This has been partly demonstrated for various emotion dimensions in the previous section. In this case we examine the problem from a different perspective. The interest annotations differ from the annotations provided with SEMAINE by (i) the set of annotators are *disjoint* from the annotators for SEMAINE, and (ii) the annotation tool employed for interest is joystick-based, (with a neutral position of 0, i.e. when no force is applied on the joystick), while for SEMAINE, a mouse-based tool was used (FeelTrace [157]).

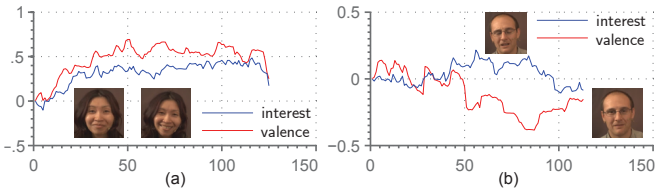


Figure 7.1: Examples from SEMAINE where (a) interest is positively correlated with valence, since the subject is in a joyful mood, (b) interest is negatively correlated with valence since the subject is angry/sad but interested in the conversation.

Firstly, we study the correlations of other emotion dimensions included in SEMAINE to the obtained interest annotations. By analysing the entire annotation set based on the correlation coefficient, we find that interest seems to be highly correlated firstly with arousal (.74), and secondly with valence (.49) and intensity (.48). We note that these findings are in accordance to previous work on evaluating the dependencies between interest, valence and arousal [130]. Plots comparing valence and interest annotations can be seen in Fig. 7.1.

Secondly, we perform experiments to evaluate the correlations between emotion dimensions and interest based on prediction accuracy. In what follows, we denote $\boldsymbol{\mathcal{S}}$ as the set of emotion dimensions (valence, arousal, power, intensity and expectation), and $\boldsymbol{\mathcal{I}}$ as the interest annotation. For each emotion dimension k in $\boldsymbol{\mathcal{S}}$, we learn the mapping $f : \boldsymbol{\mathcal{S}}_{\setminus k} \rightarrow k$, where $\boldsymbol{\mathcal{S}}_{\setminus k}$ is the set of all emotion dimensions in $\boldsymbol{\mathcal{S}}$ except k. We repeat the experiment with $\boldsymbol{\mathcal{SI}} = \boldsymbol{\mathcal{S}} \cup \boldsymbol{\mathcal{I}}$ in place of $\boldsymbol{\mathcal{S}}$, i.e. we also use interest along with emotion dimensions. Results are presented in Table 7.3. As can be seen, the correlation (COR) for most emotion dimensions increases when also using interest as a feature. As expected, the most significant increase occurs for arousal. Interestingly, this experimentally validates that although the annotations have been obtained via different tools and a disjoint set of annotators, still the obtained signals exhibit linear and non-linear correlations.

Table 7.3: Results for each emotion dimension, using (i) other emotion dimensions as features ($\boldsymbol{\mathcal{S}}_{\setminus k}$), and (ii) other emotion dimensions and interest dimension as features ($\boldsymbol{\mathcal{SI}}_{\setminus k}$). Results shown are based on the Mean Squared Error (MSE) and the correlation coefficient (COR).

	Valence		**Arousal**		**Power**		**Expectation**		**Intensity**	
	MSE	**COR**	**MSE**	**COR**	**MSE**	**COR**	**MSE**	**COR**	**MSE**	**COR**
$\boldsymbol{\mathcal{S}}_{\setminus k}$	0.074	0.28	0.051	0.47	**0.088**	**0.28**	**0.037**	0.15	0.067	0.30
$\boldsymbol{\mathcal{SI}}_{\setminus k}$	**0.063**	**0.30**	**0.052**	**0.56**	**0.088**	0.23	0.039	**0.16**	**0.052**	**0.330**

7.4 Correlated-Spaces Regression

Inspired by the results described in previous sections, we demonstrate a method which exploits output-correlations, while performing multi-modal fusion and dimensionality reduction. Note that the latter experiments also motivate the idea of dimensionality reduction on this problem:

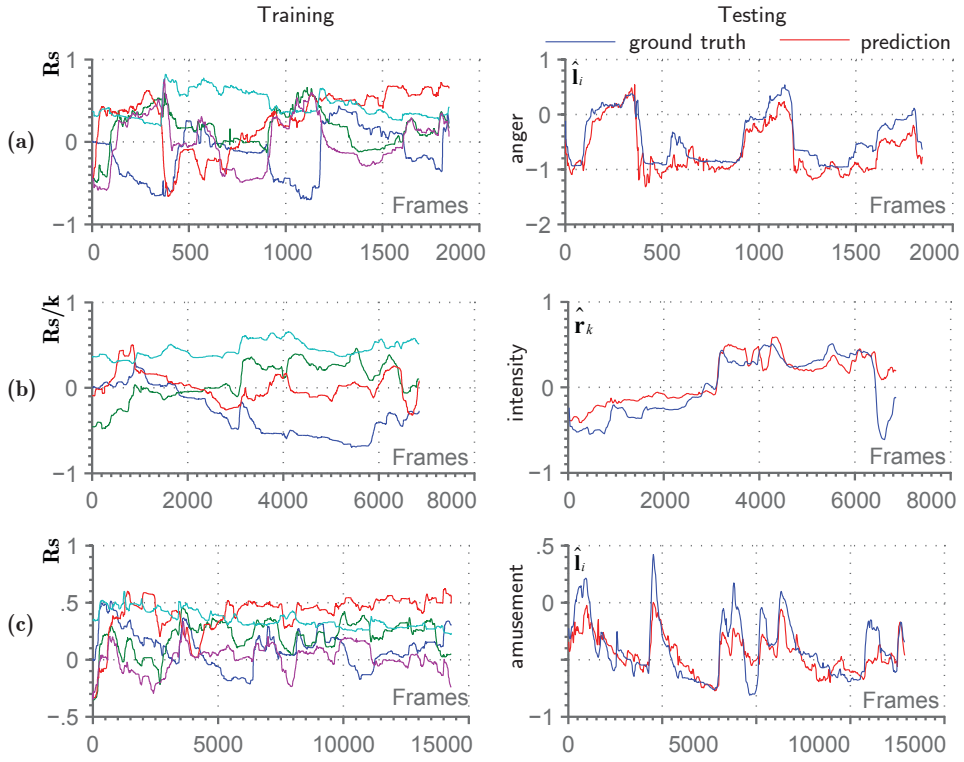


Figure 7.2: (a,c) Using emotion dimensions ($\mathbf{R}_s$) for predicting basic emotions, (b) using $k-1$ emotion dimensions ($\mathbf{R}_{s\backslash k}$) for predicting dimension k.

In the experiments in Section 7.3.1, $\mathbf{R}_{\backslash k}$ consists of 4-dimensional feature vectors and attains better performance than, i.e. the 226-dimensional facial expression vectors. We show how by exploiting feature-label, inter-feature and inter-label correlations we can significantly improve the results.

Let us assume that for a training sequence s, we have a set of annotations for emotion dimensions $\mathbf{R}_s$, containing the five dimensions used in Section 7.3.1, along with a given set of features, $\mathbf{F}_{j,s}, j = \{1, \ldots, \mu\}$ which can contain e.g., video or/and audio cues. Canonical Correlation Analysis (CCA) enables the discovery of projections of the features onto a space where

they are maximally correlated. We reformulate the problem to match our context as follows

$$\begin{aligned} &\arg\min_{\mathbf{V}_{F_s},\mathbf{V}_R} ||\mathbf{F}_s\mathbf{V}_{F_s} - \mathbf{R}_s\mathbf{V}_R||_2^F \\ &s.t.\ \mathbf{F}_s\mathbf{V}_{F_s}\mathbf{V}_{F_s}^T\mathbf{F}_s^T = \mathbf{R}_s\mathbf{V}_R\mathbf{V}_R^T\mathbf{R}_s^T = \mathbf{I} \\ &\mathbf{F}_s = [\mathbf{F}_{1,s},\ldots,\mathbf{F}_{j,s}],\ \mathbf{V}_{F_s}^T = [\mathbf{V}_{F_{1,s}}^T,\ldots,\mathbf{V}_{F_{\mu,s}}^T]^T, \end{aligned} \tag{7.3}$$

where $\mathbf{I}$ is the identity matrix. Therefore, by applying CCA on *both* the labels and the features, we are in a sense employing supervision on the feature projections, i.e. performing supervised component analysis. This is due to the fact that the labels and features are projected into a common space where they maximally correlate. In fact, for problems where labels are discrete classes, it has been shown that applying CCA on both features and binary labels collapses to applying Linear Discriminant Analysis [11], where $\mathbf{F}_s\mathbf{V}_F$ are the discriminant projections. Furthermore, as an implication of the orthogonality constraints of the problem statement in Eq. 7.3, the projected label space will be uncorrelated, thus enabling regessors to learn output-correlations which exist in the label space. Finally, due to the block-matrix formulation we learn correlated features from *all* feature sets, i.e. we perform multi-modal supervised fusion. Our model is described in Alg. 3, and visually depicted in Fig. 7.3. During training, the

Algorithm 2 Correlated-Spaces Regression

Data: Train=$(\mathbf{R}_s, \mathbf{F}_{1,s},\ldots,\mathbf{F}_{\mu,s})$ Test=$(\mathbf{F}_{1,t},\ldots,\mathbf{F}_{\mu,t})$

Result: $\hat{\mathbf{R}}_t$

train

Set $[\mathbf{V}_R, \overbrace{\mathbf{V}_{F_1},\ldots,\mathbf{V}_{F_\mu}}^{\mathbf{V}_{F_s}}]$ to the leading eigenvectors of

$$\begin{bmatrix} \mathbf{0} & \mathbf{F}_s\mathbf{R}_s^T \\ \mathbf{R}_s\mathbf{F}_s^T & \mathbf{0} \end{bmatrix} \begin{bmatrix} \mathbf{V}_{F_s} \\ \mathbf{V}_R \end{bmatrix} = \begin{bmatrix} \mathbf{F}_s\mathbf{F}_s^T & \mathbf{0} \\ \mathbf{0} & \mathbf{R}_s\mathbf{R}_s^T \end{bmatrix} \begin{bmatrix} \mathbf{V}_{F_s} \\ \mathbf{V}_R \end{bmatrix} \Lambda$$

(*Problem defined in Eq. 7.3*)

$\mathbf{F}_{i,s}^c = \mathbf{F}_{i,s}\mathbf{V}_{F_i}, \forall i \in \{1,\ldots,\mu\}$ $f : \mathbf{F}_{1:\mu,s}^c \rightarrow \mathbf{R}_s\mathbf{V}_R$

test

$\mathbf{F}_{i,t}^c = \mathbf{F}_{i,t}\mathbf{V}_{F_i}, \forall i \in \{1,\ldots,\mu\}$ $\hat{\mathbf{R}}_t^c \leftarrow f(\mathbf{F}_{1:\mu,t}^c)$ $\hat{\mathbf{R}}_t = \hat{\mathbf{R}}_t^c\mathbf{V}_R^{-1}$

projection vectors for the continuous label space $\mathbf{V}_R$ and the feature sets employed $\mathbf{F}_{1:\mu}$ are obtained. Using these projection matrices, the training features $\mathbf{F}_{1:\mu,s}$ and labels $\mathbf{R}_s$ are projected onto the space where they maximally correlate, obtaining the matrices $\mathbf{F}_{1:\mu,s}^c$ and $\mathbf{R}_s^c$.

Table 7.4: Results for predicting each emotion dimension using Correlated-Spaces Regression (CSR) utilising facial features (F^{CSR}), acoustic features (A^{CSR}) and the fusion of face and audio ($\{\mathrm{F+A}\}^{CSR}$) using CSR, utilising the Mean Squared Error (MSE) and the correlation coefficient (COR).

	Valence		**Arousal**		**Power**		**Expectation**		**Intensity**	
	MSE	COR	MSE	COR	MSE	COR	MSE	COR	MSE	COR
$\mathbf{F}^{CSR}$	0.070	0.20	**0.046**	**0.46**	0.080	0.11	**0.020**	0.06	**0.044**	**0.29**
$\mathbf{A}^{CSR}$	0.070	0.15	0.510	0.45	0.075	**0.11**	0.022	0.02	0.040	**0.29**
$\{\mathbf{F}, \mathbf{A}\}^{CSR}$	**0.056**	**0.21**	0.050	**0.46**	**0.063**	**0.12**	**0.020**	**0.07**	**0.044**	**0.29**

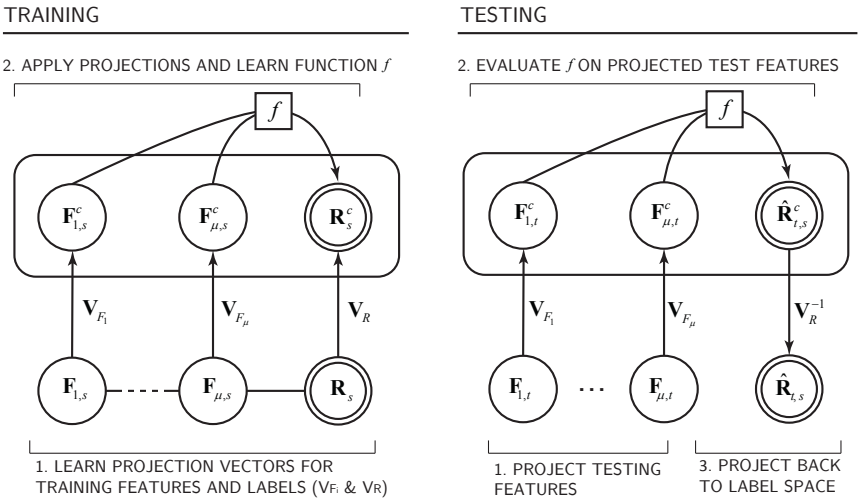


Figure 7.3: Correlated-Spaces Regression model, following Algorithm 3.

The regressor is subsequently optimised on this space

$$f : \mathbf{F}^c_{1:\mu,s} \rightarrow \mathbf{R}^c_s \tag{7.4}$$

For testing, we obtain a set of features $\mathbf{F}_{1:\mu,t}$, which we project as $\mathbf{F}^c_{i,t} = \mathbf{F}_{i,t}\mathbf{V}_{F_i}$. The learnt function f is evaluated on $\mathbf{F}^c_{i,t}$, obtaining the predictions $\hat{\mathbf{R}}^c_t$, which are then projected back to the annotation space. Results with our method are presented in Table 7.4. As can be clearly seen, our method performs much better than using simply the raw features or performing feature-level fusion, as seen in Table 7.1. In fact, it is interesting to observe that in some dimensions, our method achieves comparable correlation to using all the other annotations/labels as features ($\mathbf{R}_s$, Section 7.3.1). Essentially this means that the model manages to

capture output-correlations and in addition propagate this information during dimensionality reduction onto the projected features.

7.5 Conclusions

In this work, we performed a thorough investigation on the inter-correlation of emotion dimensions and their correlation to basic emotions. We have shown that there are more dominant correlations within emotion dimensions rather than to face or acoustic features. Furthermore, we also introduced the level of interest as a continuous dimension, and evaluated the correlations of the Level of Interest to emotion dimensions, finding that interest is mostly correlated with arousal and secondly with valence. Most importantly, we presented CSR, a CCA-based algorithm which learns output-correlations while performing multi-modal fusion and supervised dimensionality reduction. Our algorithm increases the accuracy both in terms of multi-modal fusion and single-cue regression, successfully learning output structure and maximising input-output correlations. Our algorithm can be straight-forwardly applied to any learning problem with a set of feature modalities and multi-dimensional output vectors.

Part II

Component Analysis for Affective Behaviour

Chapter 8

Introduction

The first part of this thesis dealt mostly with the problem of analysing continuous dimensional emotion annotations. Several conclusions of previous chapters influence the direction which this path takes. For example, the conclusion of Chapter 7 points to the finding that emotion dimensions appear better correlated with each other, rather than to observations such as facial features and audio cues. Beyond motivating the utilisation of emotion dimension relationships for learning (as presented in the previous part), these results also motivate the utilisation of component analysis and dimensionality reduction, since as it appears, the high dimensional observations seem to convey redundant information which is not so well correlated with the annotations. This also justifies why low-dimensional features, such as the shoulder movement feature set utilised in the previous part (consisting of 10 dimensions) as well as the audio features (15 dimensions) perform comparably, in most cases, to facial expression features (with a dimensionality of 80 or more). Beyond the motivation for lower dimensionality representations, this also points to an issue which has not been adequately dealt with, both in other related work, as well as in the previous chapters. As discussed in the introductory chapter, the annotations obtained for continuous emotion dimensions are performed on-line, and are thus vulnerable to temporal lags and discrepancies which depend on the response time of each annotator. If the fusion of annotations is performed without taking into account the various temporal discrepancies arising in the annotations, the resulting annotation will be misaligned to the corresponding samples of the ground truth. In essence, this affects the correlation

of the two signals e.g., in episodes with large variance. Note that simple averaging will not eliminate such lags, even if in the unrealistic case where they are constant in time, since lags depending on annotator response time are always positive shifts in time. Motivated by this, a large portion of this part is dedicated to the *fusion of multiple, continuous annotations*, where as we show in Chapter 9, when temporal discrepancies are eliminated, the features become much better correlated to the annotations.

Technically, the second part of this thesis is focused on Component Analysis (CA). As defined in Chapter 3, Component Analysis (CA) is a set of statistical methods aiming to factorise a signal into components which are relevant for a particular task at hand. CA is a particularly fitting paradigm for dealing with the multiple challenges arising in affect sensing. In what follows, we propose a set of novel CA models, mostly focusing on probabilistic and robust formulations, with which we are able to deal with many emerging challenges in automatic behaviour understanding via elegant and principled novel methods. Examples of applications include the fusion of multiple annotations, the robust fusion of temporal sequences, the temporal alignment of human behaviour as well as the utilisation of probabilistic feature extraction in terms of face visualisation and analysis. We summarise the work presented in this part in what follows, by following a coarse categorisation of the models presented into shared-space models (which aim at discovering a shared space underlying multiple observations) and probabilistic component analysis models aimed at analysing a single observation set.

Shared-Space Component Analysis (Chapters 9, 10)

We firstly introduce two novel methodologies based on a shared-space formulation. The aim of such models is to discover the common, underlying signal shared by many observation sets (shared space) while isolating uninteresting characteristics exhibited by each observation set separately (private space). The inferred shared space is particularly important for tasks such as fusion of multiple modalities as well as the temporal alignment of sequences. Any prior information regarding this spaces is incorporated into this models via priors in the probabilistic case and matrix norms in the deterministic case, e.g., smoothing the shared space via a Linear Dynamic System prior or modelling gross non-Gaussian noise by utilising robust norms. In

what follows, we introduce two novel Shared-Space Component Analysis models.

(i) Firstly, in Chapter 9 we propose the Dynamic Probabilistic Canonical Correlation Analysis (DPCCA). The main motivation behind this model is the *fusion of multiple continuous annotations*, which as discussed in the thesis introduction (Chapter 1), is one of the major challenges arising in the analysis of continuous dimensional emotions. Inspired by the concept of learning private-shared spaces, the proposed model is able to learn the common signal which underlies all annotations, while isolating annotator-specific characteristics which are attributed to bias and noise. By imposing Markov dependencies on the latent spaces, DPCCA is further able to model the temporal dynamics of the annotations, and further smooth out various errors arising during the annotation process. Finally, in order to "heal" the various temporal discrepancies which manifest in the annotations, DPCCA is further integrated with a temporal alignment process which is applied on the derived, clean shared space, resulting to the inferred "ground truth". The incorporation of the temporal warping leads to the DPCCA with Time Warpings model (DPCTW). Although most component analysis methods are inherently unsupervised (i.e., no label information is used, just observations), in Chapter 9 we subsequently introduce various supervised variants of DPCCA, both in a discriminative and supervised manner. Supervision can be particularly useful for fusing noisy annotations, since the observations can be used in order to impose supervision, as they are essentially the only objective reference to the sequence at hand. We show that the resulting family of models (i) can be used as a unifying framework for solving the problems of temporal alignment and fusion of multiple annotations in time, (ii) can automatically rank and filter annotations based on latent posteriors or other model statistics, and (iii) that by incorporating dynamics, modelling annotation-specific biases, noise estimation, time warping and supervision, while DPCTW outperforms state-of-the-art methods for both the aggregation of multiple, yet imperfect expert annotations as well as the alignment of affective behavior.

(ii) In Chapter 10, we introduce a robust, shared-space component analysis method. Based on Canonical Correlation Analysis (CCA), the proposed Robust CCA (RCCA) is able to

handle gross errors in the data. Such errors are often in abundance in data acquired under real-world conditions, due to occlusions, errors in localization and tracking and other forms of data corruption. Furthermore, gross errors rarely follow a Gaussian distribution, which is the de-facto assumption in the vast majority of machine learning methods. RCCA assumes that the given observations can be separated into a matrix of low-rank components shared by all observation sets, while any noise terms can be isolated into a sparse, private component specific to each observation set, while simultaneously maximising the correlation of the observations in the error-free space. We further increment RCCA with temporal warpings, enabling the temporal alignment of high-dimensional observation sequences with gross noise and corruptions. In terms of experiments, we evaluate RCCA in many, challenging scenarios such as (i) the robust audio-visual fusion for the prediction of the level of interest, (ii) robust fusion for heterogeneous face recognition, as well as (iii) the temporal alignment of facial action units and human walking sequences. We note that for (ii), the fusion scenario is more challenging, in the sense that only one of the fused modalities is available during testing. RCCA outperforms other CCA variants as well as state-of-the-art methods for temporal alignment.

A Unified Framework for Probabilistic Component Analysis (Chapter 11)

Although CA has received great attention by many researchers over the past decades, much fewer works exist on probabilistic CA. Furthermore, in terms of deterministic component analysis, various frameworks have been introduced which aim to unify many CA techniques under unifying frameworks, thus both enabling the better understanding of such methods as well providing novel methods to the community. Nevertheless, no such unifying probabilistic framework has been proposed thus far. In Chapter 11, we introduce the first unifying probabilistic component analysis framework which unifies all CA methods where the corresponding deterministic formulation leads to a trace optimisation problem without domain constraints for the parameters. In more detail, we unify methods such as Principal Component Analysis (PCA), Linear Discriminant Analysis (LDA), Locality Preserving Projections (LPP) and Slow Feature Analysis (LPP), some of which have no probabilistic equivalent in literature so far. The framework is based on modelling the latent variables as Markov Random Fields (MRFs),

while each component analysis method arises when utilising a specific MRF prior. We show that the methods derived via our framework recover projections which are co-directional to the deterministic solutions in the Maximum Likelihood case (ML), while we propose a novel Expectation Maximisation (EM) framework for component analysis. We generalise the proposed methodologies to arbitrary connectivities via parametrizable MRF products, thus facilitating the generation of novel component analysis techniques. We evaluate the proposed models on problems such as level of interest detection, face recognition as well as face recognition and visualisation of high-dimensional data. The methods derived via the proposed probabilistic framework well-outperform related probabilistic and deterministic techniques.

Chapter 9

Dynamic Probabilistic CCA for Analysis of Affective Behaviour and Fusion of Continuous Annotations

Contents

9.1 Introduction

In this chapter, we introduce a novel, probabilistic dynamic model which is tailored to the problem of fusing multiple, continuous annotations. As mentioned in the introduction of this

thesis, the annotation process itself is a tedious, expensive and highly error prone process, which in turn greatly affects the level of difficulty in terms of processing such annotations and deriving the "ground truth" that will be utilised in predictive analysis scenarios, that is, in order to train machine learning models.

We remind the reader that the annotation process in terms of continuous emotion dimensions is both continuous in space and time. That is, experts annotate in real-time audiovisual sequences of spontaneous emotion expressions, in terms of emotion dimensions such as valence (ranging from unpleasant to pleasant) and arousal (ranging from relaxed to aroused). This leads to a series of problems which have been detailed in Chapter 1 such as rendering the annotation subject to individual human judgement (i.e. varying perceptions of the intensity of an emotional state), varying temporal lags exhibited by annotators due to person-specific response times[1], as well as various types of noise introduced by the input device or the annotation procedure. The latter issues arise not only due to human factors (such as annotator skill and expertise as well as annotator characteristics such as age, fatigue and stress) but also to the fuzziness of the meaning associated with various labels related to human behaviour.

The only information which can be utilised in order to improve the quality of the derived ground truth can be obtained by (i) exploiting the fact that there exist *multiple annotations*, which in turn can provide e.g., common truths but in isolation may provide biased and uninteresting information, and (ii) by utilising any extracted features from the sequence being annotated, e.g., facial expressions or audio information, in order to aid the derivation of the ground truth. Both of these points are addressed within the proposed DPCCA model. In more detail, in this chapter we propose DPCCA, a probabilistic method following a private-shared space formulation which models latent dynamics via Markov dependencies. DPCCA is able to (i) isolate any annotator-specific bias in the private space, (ii) nullify any temporal discrepancies present in the annotations by time warping (iii) utilise any extracted features for the ground truth derivation, (iv) model latent dynamics of annotations, (v) provide a ranking of the annotations in terms of uncertainty, and in conclusion, is able to infer a clean version

[1] i.e., each annotator firstly perceives the emotional state observed, and subsequently applies a force on the input device, i.e. move the mouse or joystick; this can not be entirely synchronised to the video stream and to the precise frame at which the emotional state is manifested.

of the "ground truth", as a representation of the noise-free, shared information conveyed by all annotations. As can be seen, the aim of DPCCA is to overcome all challenges arising in fusing continuous dimensional emotion annotations from multiple experts, as also discussed in Chapter 1.

The rest of the chapter is organised as follows. In Section 9.2, we perform a summarising comparison on related work, as discussed in Chapters 2 and 3. In Section 9.3, we describe PCCA and present our extension to multiple sequences. In Sec.9.4, we introduce our proposed Dynamic PCCA, which we subsequently extend with latent space time-warping (DPCTW) as described in Section 9.5. In Section 9.6, we introduce two supervised variants of DPCTW which incorporate inputs in a generative (Section 9.6.1) and discriminative (Section 9.6.2) manner, while in Section 9.7 we present an algorithm based on the proposed family of models which ranks and filters annotators. In Section 9.8, we present various experiments on both synthetic (Section 9.8.1) and real (Section 9.8.2, 9.8.3) experimental data, emphasising the advantages of the proposed methods on both the fusion of multiple annotations and sequence alignment. Finally, conclusions are drawn in Section 9.9.

9.2 Contributions and Related Work

The usually employed technique in terms of fusing multiple annotations in affect sensing is based on simply averaging the annotations (or taking the majority value) [99], thus assuming that the average annotation approximates the true annotation which is conveyed by the annotators. As discussed in Chapter 2, simply averaging is suboptimal, as (i) we assume that all annotators are likely capable without modelling their precision (since averaging is the expected value of each annotation weighted with equal probability), and (ii) we propagate noise and temporal discrepancies in the generated ground truth. Since annotators are usually "laggy" (i.e. they exhibit a *positive* temporal delay), this essentially means that the annotation will always exhibit some lag compared to the annotated sequence.

A state-of-the-art approach in fusing multiple continuous annotations that *can* be applied to emotion descriptions is proposed by Raykar et al. [209]. In this work, each noisy annotation

is considered to be generated by a Gaussian distribution with the mean being the true label and the variance representing the annotation noise.

A main drawback of [209] lies in the assumption that temporal correspondences of samples are known. One way to find such arbitrary temporal correspondences is via time warping. A state-of-the-art approach for time warping, Canonical Time Warping (CTW) [298], combines Dynamic Time Warping (DTW) and Canonical Correlation Analysis (CCA) with the aim of aligning a pair of sequences of both different duration and different dimensionality. CTW accomplishes this by simultaneously finding the most correlated features and samples among the two sequences, both in feature space and time. This task is reminiscent of the goal of fusing expert annotations. However, CTW does not directly yield the prototypical sequence, which is considered as a common, denoised and fused version of multiple experts' annotations. As a consequence, this renders neither of the two state-of-the-art methods applicable to our setting.

The latter observation precisely motivates our work; inspired by Probabilistic Canonical Correlation Analysis (PCCA) [121], we initially present the first generalisation of PCCA to learning temporal dependencies in the shared/individual spaces (Dynamic PCCA, DPCCA). By further augmenting DPCCA with time warping, the resulting model (Dynamic PCCA with Time Warpings, DPCTW) can be seen as a unifying framework, concisely applied to both problems. The individual contributions of this work can be summarised as follows:

- In comparison to state-of-the-art approaches in both fusion of multiple annotations and sequence alignment, our model bears several advantages. We assume that the "true" annotation/sequence lies in a shared latent space. E.g., in the problem of fusing multiple emotion annotations, we know that the experts have a common training in annotation. Nevertheless, each carries a set of individual factors which can be assumed to be uninteresting (e.g., annotator/sequence specific bias). In the proposed model, individual factors are accounted for within an annotator-specific latent space, thus effectively preventing the contamination of the shared space by individual factors. Most importantly, we introduce latent-space dynamics which model temporal dependencies in both com-

mon and individual signals. Furthermore, due to the probabilistic and dynamic nature of the model, each annotator/sequence's uncertainty can be estimated for each *sample*, rather than for each sequence.

- In contrast to current work on fusing multiple annotations, we propose a novel framework able to handle temporal tasks. In addition to introducing dynamics, we also employ temporal alignment in order to eliminate temporal discrepancies amongst the annotations.
- We present an elegant extension of DTW-based sequence alignment techniques (e.g., Canonical Time Warping, CTW) to a probabilistic multiple-sequence setting. We accomplish this by treating the problem in a generative probabilistic setting, both in the static (multiset PCCA) and dynamic case (Dynamic PCCA).

9.3 Multiset Probabilistic CCA

We consider the probabilistic interpretation of CCA, introduced by Bach & Jordan [11] and generalised by Klami & Kaski [121][2]. In this section, we present an extended version of PCCA [121] (multiset PCCA[3]) which is able to handle any arbitrary number of sets. We consider a collection of datasets $\mathcal{D} = \{\mathbf{X}_1, \mathbf{X}_2, ..., \mathbf{X}_N\}$, with each $\mathbf{X}_i \in \mathbb{R}^{D_i \times T}$ where D_i is the dimensionality and T the number of instances. By adopting the generative model for PCCA, the observation sample n of set $\mathbf{X}_i \in \mathcal{D}$ is assumed to be generated as

$$\mathbf{x}_{i,n} = f(\mathbf{z}_n|\mathbf{W}_i) + g(\mathbf{z}_{i,n}|\mathbf{B}_i) + \epsilon_i, \tag{9.1}$$

where $\mathbf{Z}_i = [\mathbf{z}_{i,1}, \ldots, \mathbf{z}_{i,T}] \in \mathbb{R}^{d_i \times T}$ and $\mathbf{Z} = [\mathbf{z}_1, \ldots, \mathbf{z}_T] \in \mathbb{R}^{d \times T}$ are the *independent* latent variables that capture the set-specific individual characteristics and the shared signal amongst all observation sets, respectively. $f(.)$ and $g(.)$ are functions that transform each of the latent signals $\mathbf{Z}$ and $\mathbf{Z}_i$ into the observation space. They are parametrised by $\mathbf{W}_i$ and $\mathbf{B}_i$, while the noise for each set is represented by ϵ_i, with $\epsilon_i \perp \epsilon_j, i \neq j$. Similarly to [121], $\mathbf{z}_n$, $\mathbf{z}_{i,n}$ and ϵ_i are considered to be independent (both over the set and the sequence) and normally distributed:

$$\mathbf{z}_n, \mathbf{z}_{i,n} \sim \mathcal{N}(0, \mathbf{I}), \epsilon_i \sim \mathcal{N}(0, \sigma_n^2 \mathbf{I}). \tag{9.2}$$

[2] [121] is also related to Tucker's inter-battery factor analysis [253, 32]

[3] In what follows we refer to multiset PCCA as PCCA.

By considering f and g to be linear functions we have $f(\mathbf{z}_n|\mathbf{W}_i) = \mathbf{W}_i\mathbf{z}_n$ and $g(\mathbf{z}_{i,n}|\mathbf{B}_i) = \mathbf{B}_i\mathbf{z}_{i,n}$, transforming the model presented in Eq. 9.1, to

$$\mathbf{x}_{i,n} = \mathbf{W}_i\mathbf{z}_n + \mathbf{B}_i\mathbf{z}_{i,n} + \epsilon_i. \tag{9.3}$$

Learning the multiset PCCA can be accomplished by generalising the EM algorithm presented in [121], applied to two or more sets. Firstly, $P(\mathcal{D}|\mathbf{Z}, \mathbf{Z}_1, \ldots, \mathbf{Z}_N)$ is marginalised over set-specific factors $\mathbf{Z}_1, \ldots, \mathbf{Z}_N$ and optimised on each $\mathbf{W}_i$. This leads to the generative model $P(\mathbf{x}_{i,n}|\mathbf{z}_n) \sim \mathcal{N}(\mathbf{W}_i\mathbf{z}_n, \mathbf{\Psi}_i)$, where $\mathbf{\Psi}_i = \mathbf{B}_i\mathbf{B}_i^T + \sigma_i^2\mathbf{I}$. Subsequently, $P(\mathcal{D}|\mathbf{Z}, \mathbf{Z}_1, \ldots, \mathbf{Z}_N)$ is marginalised over the common factor $\mathbf{Z}$ and then optimised on each $\mathbf{B}_i$ and σ_i. When generalising the algorithm for more than two sets, we also have to consider how to (i) obtain the expectation of the latent space and (ii) provide stable variance updates for all sets.

Two quantities are of interest regarding the latent space estimation. The first is the common latent space given one set, $\mathbf{Z}|\mathbf{X}_i$. In the classical CCA this is analogous to finding the canonical variables [121]. We estimate the posterior of the shared latent variable $\mathbf{Z}$ as follows:

$$\begin{aligned} P(\mathbf{z}_n|\mathbf{x}_{i,n}) &\sim \mathcal{N}(\boldsymbol{\gamma}_i\mathbf{x}_{i,n}, \mathbf{I} - \gamma_i\mathbf{W}_i), \\ \boldsymbol{\gamma}_i &= \mathbf{W}_i^T(\mathbf{W}_i\mathbf{W}_i^T + \mathbf{\Psi}_i)^{-1}. \end{aligned} \tag{9.4}$$

The latent space given the n-th sample from *all* sets in $\mathcal{D}$, which provides a better estimate of the shared signal manifested in all observation sets is estimated as

$$\begin{aligned} P(\mathbf{z}_n|\mathbf{x}_{1:N,n}) &\sim \mathcal{N}(\boldsymbol{\gamma}\mathbf{x}_{1:N,n}, \mathbf{I} - \boldsymbol{\gamma}\mathbf{W}), \\ \boldsymbol{\gamma} &= \mathbf{W}^T(\mathbf{W}\mathbf{W}^T + \mathbf{\Psi})^{-1}, \end{aligned} \tag{9.5}$$

while the matrices $\mathbf{W}$, $\mathbf{\Psi}$ and $\mathbf{X}_n$ are defined as $\mathbf{W}^T = [\mathbf{W}_1^T, \mathbf{W}_2^T, \ldots, \mathbf{W}_n^T]$, $\mathbf{\Psi}$ as the block diagonal matrix of $\mathbf{\Psi}_{i=1:N}$ [4] and $\mathbf{x}_{1:N,n}^T = [\mathbf{x}_{1,n}^T, \mathbf{x}_{2,n}^T, \ldots, \mathbf{x}_{1:N,n}^T]$. Finally, the variance is recovered on the full model, $x_{i,n} \sim \mathcal{N}(\mathbf{W}_i\mathbf{z}_n + \mathbf{B}_i\mathbf{z}_{i,n}, \sigma_i^2\mathbf{I})$, as

$$\begin{aligned} \sigma_i^2 =& tr(\mathbf{S} - \mathbf{X}\mathbb{E}[\mathbf{Z}^T|\mathbf{X}]\mathbf{C}^T \\ &-\mathbf{C}\mathbb{E}[\mathbf{Z}|\mathbf{X}]\mathbf{X}^T - \mathbf{C}\mathbb{E}[\mathbf{Z}\mathbf{Z}^T|\mathbf{X}]\mathbf{C}^T)_i\frac{T}{D_i}, \end{aligned} \tag{9.6}$$

[4] For brevity of notation, we use $1:N$ to indicate elements $[1, \ldots, N]$, e.g., $\mathbf{X}_{1:N} \equiv [\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_N]$

where $\mathbf{S}$ is the sample covariance matrix, $\mathbf{B}$ is the block diagonal matrix of $\mathbf{B}_{i=1:N}$, $\mathbf{C} = [\mathbf{W}, \mathbf{B}]$, while the subscript i in Eq. 9.6 refers to the i-th block of the full covariance matrix. Finally, we note that the computational complexity of PCCA for each iteration is similar to deterministic CCA (cubic in the dimensionality of the datasets and linear in the number of samples). PCCA though also recovers the private space.

9.4 Dynamic Probabilistic CCA (DPCCA)

The PCCA model described in Section 9.3 exhibits several advantages when compared to the classical formulation of CCA, mainly by providing a probabilistic estimation of a latent space shared by an arbitrary collection of datasets along with explicit noise and private space estimation. Nevertheless, static models are unable to learn temporal dependencies which are very likely to exist when dealing with real-life problems. In fact, dynamics are deemed essential for successfully performing tasks such as emotion recognition, AU detection etc. [285].

Motivated by the former observation, we propose a dynamic generalisation of the static PCCA model introduced in the previous section, where we now treat each $\mathbf{X}_i$ as a temporal sequence. For simplicity of presentation, we introduce a linear model[5] where Markovian dependencies are learnt in the latent spaces $\mathbf{Z}$ and $\mathbf{Z}_i$. In other words, the variable $\mathbf{Z}$ models the temporal, shared signal amongst all observation sequences, while $\mathbf{Z}_i$ captures the temporal, individual characteristics of each sequence. It is easy to observe that such a model fits perfectly with the problem of fusing multiple annotations, as it does not only capture the temporal shared signal of all annotations, but also models the unwanted, annotator-specific factors over time. Essentially, instead of directly applying the doubly independent priors to $\mathbf{Z}$ as in Eq. 9.2, we now use the following:

$$p(\mathbf{z}_t|\mathbf{z}_{t-1}) \sim \mathcal{N}(\mathbf{A}_z\mathbf{z}_{t-1}, \mathbf{V}_Z), \tag{9.7}$$

$$p(\mathbf{z}_{i,t}|\mathbf{z}_{i,t-1}) \sim \mathcal{N}(\mathbf{A}_{z_i}\mathbf{z}_{i,t-1}, \mathbf{V}_{Z_i}), n = 1, \ldots, N, \tag{9.8}$$

[5] A non-linear DPCCA model can be derived similarly to [118, 83].

where the transition matrices $\mathbf{A}_z$ and $\mathbf{A}_{z_i}$ model the latent space dynamics for the shared and sequence-specific space respectively. Thus, idiosyncratic characteristics of dynamic nature appearing in a single sequence can be accurately estimated and prevented from contaminating the estimation of the shared signal.

The resulting model bears similarities with traditional Linear Dynamic System (LDS) models (e.g. [212]) and the so-called Factorial Dynamic Models, c.f. [82]. Along with Eq. 9.7,9.8 and noting Eq. 9.3, the dynamic, generative model for DPCCA[6] can be described as

$$\mathbf{x}_{i,t} = \mathbf{W}_{i,t}\mathbf{z}_t + \mathbf{B}_i\mathbf{z}_{i,t} + \epsilon_i, \epsilon_i \sim \mathcal{N}(0, \sigma_i^2\mathbf{I}), \tag{9.9}$$

where the subscripts i and t refer to the i-th observation sequence timestep t respectively.

9.4.1 Inference

To perform inference, we reduce the DPCCA model to a LDS[7]. This can be accomplished by defining a joint space $\hat{\mathbf{Z}}^T = [\mathbf{Z}^T, \mathbf{Z}_1^T, \ldots, \mathbf{Z}_N^T]$, $\hat{\mathbf{Z}} \in \mathbb{R}^{\hat{d} \times T}$ where $\hat{d} = d + \sum_i^N d_i$ with parameters $\boldsymbol{\theta} = \{\mathbf{A}, \mathbf{W}, \mathbf{B}, \mathbf{V}_{\hat{z}}, \hat{\boldsymbol{\Sigma}}\}$. Dynamics in this joint space are described as $\mathbf{X}_t = [\mathbf{W}, \mathbf{B}]\hat{\mathbf{Z}}_t + \boldsymbol{\epsilon}, \hat{\mathbf{Z}}_t = \mathbf{A}\hat{\mathbf{Z}}_{t-1} + \mathbf{u}$, where the noise processes $\boldsymbol{\epsilon}$ and $\mathbf{u}$ are defined as

$$\boldsymbol{\epsilon} \sim \mathcal{N}\left(0, \underbrace{\begin{bmatrix} \sigma_1^2\mathbf{I} & & \\ & \ddots & \\ & & \sigma_N^2\mathbf{I} \end{bmatrix}}_{\hat{\boldsymbol{\Sigma}}}\right), \tag{9.10}$$

$$\mathbf{u} \sim \mathcal{N}\left(0, \underbrace{\begin{bmatrix} \mathbf{V}_z & & & \\ & \mathbf{V}_{z_1} & & \\ & & \ddots & \\ & & & \mathbf{V}_{z_N} \end{bmatrix}}_{\mathbf{V}_{\hat{z}}}\right), \tag{9.11}$$

[6] The model of Raykar et al. [209] can be considered as a special case of (D)PCCA by setting $\mathbf{W} = \mathbf{I}$, $\mathbf{B} = \mathbf{0}$ (and disregarding dynamics).

[7] For more details on LDS, please see [212] and [26], Chapter 13.

where $\mathbf{V}_z \in \mathbb{R}^{d\times T}$ and $\mathbf{V}_{z_i} \in \mathbb{R}^{d_i\times T}$. The other matrices used above are defined as $\mathbf{X}^T = [\mathbf{X}_1^T, \ldots, \mathbf{X}_N^T]$, $\mathbf{W}^T = [\mathbf{W}_1^T, \ldots, \mathbf{W}_N^T]$, $\mathbf{B}$ as the block diagonal matrix of $[\mathbf{B}_1, \ldots, \mathbf{B}_N]$ and $\mathbf{A}$ as the block diagonal matrix of $[\mathbf{A}_z, \mathbf{A}_{z_1}, \ldots, \mathbf{A}_{z_N}]$. Similarly to LDS, the joint log-likelihood function of DPCCA is defined as

$$\begin{aligned} lnP(\mathbf{X}, \mathbf{Z}|\theta) =& lnP(\hat{\mathbf{z}}_1|\mu, V) + \sum_{t=2}^{T} lnP(\hat{\mathbf{z}}_t|\hat{\mathbf{z}}_{t-1}, \mathbf{A}, \mathbf{V}_{\hat{z}}) \\ &+ \sum_{t=1}^{T} lnP(\mathbf{x}_t|\hat{\mathbf{z}}_t, \mathbf{W}, \mathbf{B}, \hat{\mathbf{\Sigma}}). \end{aligned} \tag{9.12}$$

In order estimate the latent spaces, we apply the Rauch-Tung-Striebel (RTS) smoother on $\hat{\mathbf{Z}}$ (the algorithm can be found in [212], A.3). In this way, we obtain $\mathbb{E}[\hat{\mathbf{z}}_t|\mathbf{X}^T]$, $V[\hat{\mathbf{z}}_t|\mathbf{X}^T]$ and $V[\hat{\mathbf{z}}_t\hat{\mathbf{z}}_{t-1}|\mathbf{X}^T]$[8].

9.4.2 Parameter Estimation

The parameter estimation of the M-step has to be derived specifically for this factorised model. We consider the expectation of the joint model log-likelihood (Eq. 9.12) wrt. posterior and obtain the partial derivatives of each parameter for finding the stationary points. Note the $\mathbf{W}$ and $\mathbf{B}$ matrices appear in the likelihood as:

$$\begin{aligned} \mathbb{E}_{\hat{z}}[lnP(\mathbf{X}, \hat{\mathbf{Z}})] = -\frac{T}{2} ln|\hat{\mathbf{\Sigma}}| - \mathbb{E}_{\hat{z}}\Bigg[\sum_{t=1}^{T} (\mathbf{x}_t - [\mathbf{W}, \mathbf{B}]\hat{\mathbf{z}}_t)^T \\ \hat{\mathbf{\Sigma}}^{-1} (\mathbf{x}_t - [\mathbf{W}, \mathbf{B}]\hat{\mathbf{z}}_t) \Bigg] + \ldots. \end{aligned} \tag{9.13}$$

Since they are composed of individual $\mathbf{W}_i$ and $\mathbf{B}_i$ matrices (which are parameters for each sequence i), we calculate the partial derivatives $\partial\mathbf{W}_i$ and $\partial\mathbf{B}_i$ in Eq. 9.13. Subsequently, by setting to zero and re-arranging, we obtain the update equations for each $\mathbf{W}_i^*$ and $\mathbf{B}_i^*$:

$$\mathbf{W}_i^* = \left(\sum_{t=1}^{T} \mathbf{x}_{i,t}\mathbb{E}[\mathbf{z}_{i,t}] - \mathbf{B}_i^*\mathbb{E}[\mathbf{z}_{i,t}\mathbf{z}_t^T] \right) \left(\sum_{t=1}^{T} \mathbb{E}[\mathbf{z}_t\mathbf{z}_t^T] \right)^{-1} \tag{9.14}$$

[8] We note that the complexity of RTS is cubic in the dimension of the state space. Thus, when estimating high dimensional latent spaces, computational or numerical issues may arise (due to the inversion of large matrices). If any of the above is a concern, the complexity of RTS can be reduced to quadratic [260], while inference can be performed more efficiently similarly to [82].

$$\mathbf{B}_i^* = \left(\sum_{t=1}^{T} \mathbf{x}_{i,t} \mathbb{E}[\mathbf{z}_t^T] - \mathbf{W}_i^* \mathbb{E}[\mathbf{z}_t \mathbf{z}_{i,t}^T] \right) \left(\sum_{t=1}^{T} \mathbb{E}[\mathbf{z}_{i,t} \mathbf{z}_{i,t}^T] \right)^{-1} \tag{9.15}$$

Note that the weights are *coupled* and thus the optimal solution should be found iteratively. As can be seen, in contrast to PCCA, in DPCCA the individual factors of each sequence are explicitly estimated instead of being marginalised out. Similarly, the transition weight updates for the individual factors $\mathbf{Z}_i$ are as follows:

$$\mathbf{A}_{z,i}^* = \left(\sum_{t=2}^{T} E[\mathbf{z}_{i,t} \mathbf{z}_{i,t-1}^T] \right) \left(\sum_{t=2}^{T} E[\mathbf{z}_{i,t-1} \mathbf{z}_{i,t-1}^T] \right)^{-1} \tag{9.16}$$

where by removing the subscript i we obtain the updates for $\mathbf{A}_z$, corresponding to the shared latent space $\mathbf{Z}$. Finally, the noise updates $\mathbf{V}_{\hat{Z}}$ and $\hat{\mathbf{\Sigma}}$ are estimated similarly to LDS [212].

9.5 DPCCA with Time Warpings

Both PCCA and DPCCA exhibit several advantages in comparison to the classical formulation of CCA. Mainly, as we have shown, (D)PCCA can inherently handle more than two sequences, building upon the multiset nature of PCCA. This is in contrast to the classical formulation of CCA, which due to the pairwise nature of the correlation operator is limited to two sequences[9]. This is crucial for the problems at hand since both methods yield an accurate estimation of the underlying signals of *all* observation sequences, free of individual factors and noise. However, both PCCA and DPCCA carry the assumption that the temporal correspondences between samples of different sequences are *known*, i.e. that the annotation of expert i at time t directly corresponds to the annotation of expert j at the same time. Nevertheless, this assumption is often violated since different experts exhibit different time lags in annotating the same process. Motivated by the latter, we extend the DPCCA model to account for this *misalignment* of data samples by introducing a latent warping process into DPCCA, in a manner similar to [298]. In what follows, we firstly describe some basic background on time-warping and subsequently proceed to define our model.

[9] The recently proposed multiset-CCA [101] can handle multiple sequences but requires maximising over sums of pairwise operations.

9.5.1 Time Warping

The basics of time warping have been described in Chapter 3. Nevertheless, in order to make this chapter self-complete, we briefly summarise related work in what follows. Dynamic Time Warping (DTW) [205] is an algorithm for optimally aligning two sequences of possibly different lengths. Given sequences $\mathbf{X} \in \mathbb{R}^{D \times T_x}$ and $\mathbf{Y} \in \mathbb{R}^{D \times T_y}$, DTW aligns the samples of each sequence by minimising the sum-of-squares cost, i.e. $||\mathbf{X}\mathbf{\Delta}_x - \mathbf{Y}\mathbf{\Delta}_y||_F^2$, where $\mathbf{\Delta}_x \in \mathbb{R}^{T_x \times T_\Delta}$ and $\mathbf{\Delta}_y \in \mathbb{R}^{T_y \times T_\Delta}$ are binary selection matrices, with T_Δ the aligned, common length. In this way, the warping matrices $\mathbf{\Delta}$ effectively re-map the samples of each sequence. Although the number of possible alignments is exponential in $T_x T_y$, employing dynamic programming can recover the optimal path in $\mathcal{O}(T_x T_y)$. Furthermore, the solution must satisfy the boundary, continuity and monotonicity constraints, effectively restricting the space of $\mathbf{\Delta}_x$, $\mathbf{\Delta}_y$ [205]. An important limitation of DTW is the inability to align signals of different dimensionality. Motivated by the former, CTW [298] combines CCA and DTW, thus alowing the alignment of signals of different dimensionality by projecting into a common space via CCA. The optimisation function now becomes $||\mathbf{V}_x^T \mathbf{X}\mathbf{\Delta}_x - \mathbf{V}_y^T \mathbf{Y}\mathbf{\Delta}_y||_F^2$, where $\mathbf{X} \in \mathbb{R}^{D_x \times T_x}$, $\mathbf{Y} \in \mathbb{R}^{D_y \times T_x}$, and $\mathbf{V}_x, \mathbf{V}_y$ are the projection operators (matrices).

9.5.2 DPCTW Model

We define DPCTW based on the graphical model presented in Fig. 9.1. Given a set $\mathcal{D}$ of N sequences of varying duration, with each sequence $\mathbf{X}_i = [\mathbf{x}_{i,1}, \ldots, \mathbf{x}_{i,T_i}] \in \mathbb{R}^{D_i \times T_i}$, we postulate the latent common Markov process $\mathbf{Z} = \{\mathbf{z}_1, \ldots, \mathbf{z}_t\}$. Firstly, $\mathbf{Z}$ is warped using the warping operator $\mathbf{\Delta}_i$, resulting in the warped latent sequence $\boldsymbol{\zeta}_i$. Subsequently, each $\boldsymbol{\zeta}_i$ generates each observation sequence $\mathbf{X}_i$, also considering the annotator/sequence bias $\mathbf{Z}_i$ and the observation noise σ_i^2. We note that we do not impose parametric models for warping processes. Inference in this general model can be prohibitively expensive, in particular because of the need to handle the unknown alignments. We instead propose to handle the inference in two steps: (i) fix the alignments $\mathbf{\Delta}_i$ and find the latent $\mathbf{Z}$ and $\mathbf{Z}_i$'s, and (ii) given the estimated $\mathbf{Z}, \mathbf{Z}_i$ find

the optimal warpings $\mathbf{\Delta}_i$. For this, we propose to optimise the following objective function:

$$\mathcal{L}_{\text{(D)PCTW}} = \sum_{i}^{N} \sum_{j,j\neq i}^{N} \frac{||\mathbb{E}[\mathbf{Z}|\mathbf{X}_i]\mathbf{\Delta}_i - \mathbb{E}[\mathbf{Z}|\mathbf{X}_j]\mathbf{\Delta}_j||_F^2}{N(N-1)} \tag{9.17}$$

where when using PCCA, $\mathbb{E}[\mathbf{Z}|\mathbf{X}_i] = \mathbf{W}_i^T(\mathbf{W}_i\mathbf{W}_i^T + \mathbf{\Psi}_i)^{-1}\mathbf{X}_i$ (Eq. 9.4). For DPCCA, $\mathbb{E}[\mathbf{Z}|\mathbf{X}_i]$ is inferred via RTS smoothing (Section 9.4). A summary of the full algorithm is presented in Algorithm 3.

At this point, it is important to clarify that our model is flexible enough to be straightforwardly used with varying warping techniques. For example, the Gauss-Newton warping proposed in [296] can be used as the underlying warping process for DPCCA, by replacing the projected data $\mathbf{V}_i^T\mathbf{X}_i$ with $\mathbb{E}[\mathbf{Z}|\mathbf{X}_i]$ in the optimisation function. Algorithmically, this only changes the warping process (line 3, Algorithm 3). Finally, we note that since our model iterates between estimating the latent spaces with (D)PCCA and warping, the computational complexity of time warping is additive to the cost of each iteration. In case of the DTW alignment for two sequences, this incurs an extra cost of $O(T_xT_y)$. In case of more than two sequences, we utilise a DTW-based algorithm, which is a variant of the so-called Guide Tree Progressive Alignment, since the complexity of dynamic programming increases exponentially with the number of sequences. Similar algorithms are used in state-of-the-art sequence alignment software in biology, e.g., Clustar [131]. The complexity of the employed algorithm is $O(N^2T_{max}^2)$ where T_{max} is the maximum (aligned) sequence length and N the number of sequences. More efficient implementations can also be used by employing various constraints [205].

9.6 Features for Annotator Fusion

In the previous sections, we considered the observed data to consist only of the given annotations, $\mathcal{D} = \{\mathbf{X}_1, \ldots, \mathbf{X}_N\}$. Nevertheless, in many problems one can extract additional observed information, which we can consider as a form of *complementary input* (e.g., visual or acoustic features). In fact, in problems where annotations are subjective and no objective ground truth is available for any portion of the data, such input can be considered as the only objective reference to the annotation/sequence at hand. Thus, incorporating it into the model can significantly aid the determination of the ground truth.

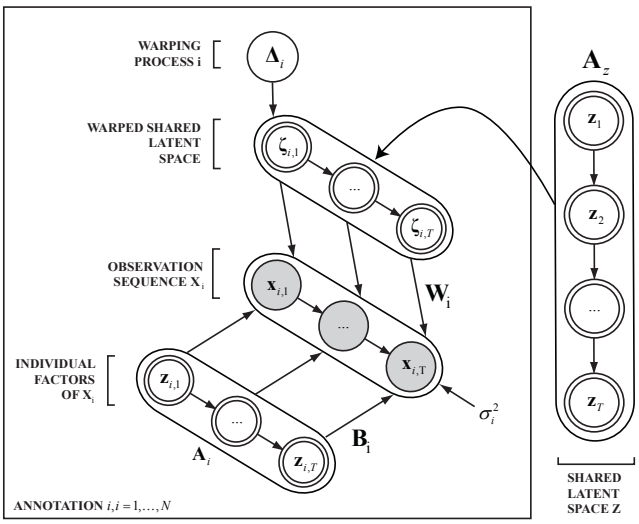


Figure 9.1: Graphical model of DPCTW. Shaded nodes represent the observations. By ignoring the temporal dependencies, we obtain the PCTW model.

Motivated by the latter argument, we propose two models which augment DPCCA/DPCTW with inputs. Since the family of component analysis techniques we study are typically unsupervised, incorporating inputs leads to a form of supervised learning. Such models can find a wide variety of applications since they are able to exploit label information in addition to observations. A suitable example lies in dimensional affect analysis, where it has been shown that specific emotion dimensions correlate better with specific cues, (e.g., valence with facial features, arousal with acoustic features (Chapter 5, [174, 99]). Thus, one can know a-priori which features to use for specific annotations.

Throughout this discussion, we assume that a set of complementary input or features $\mathbf{Y} = \{\mathbf{Y}_1, \ldots, \mathbf{Y}_\nu\}$ is available, where $\mathbf{Y}_j \in \mathbb{R}^{D_{y_j} \times T_{y_j}}$. While discussing extensions of DPCCA, we assume that all sequences have equal length. When incorporating time warping, sequences can have different lengths.

Algorithm 3 Dynamic Probabilistic CCA with Time Warpings (DPCTW)

Data: $\mathcal{D} = \mathbf{X}_1, \ldots, \mathbf{X}_N,\ \mathbf{X}^T = [\mathbf{X}_1^T, \ldots, \mathbf{X}_N^T]$
Result: $P(\mathbf{Z}|\mathbf{X}_1, \ldots \mathbf{X}_N), P(\mathbf{Z}|\mathbf{X}_i), \boldsymbol{\Delta}_i, \sigma_i^2, i = 1:N$
repeat
  Obtain alignment matrices $(\boldsymbol{\Delta}_1, \ldots, \boldsymbol{\Delta}_N)$ by optimising Eq. 9.17 on $\mathbb{E}[\mathbf{Z}|\mathbf{X}_1^T], \ldots, \mathbb{E}[\mathbf{Z}|\mathbf{X}_N^T]$
  $\mathbf{X}_\Delta^T = [(\mathbf{X}_1\boldsymbol{\Delta}_1)^T, \ldots, (\mathbf{X}_N\boldsymbol{\Delta}_N)^T]$
  repeat
    Estimate $\mathbb{E}[\hat{\mathbf{z}}_t|\mathbf{X}_\Delta^T]$, $V[\hat{\mathbf{z}}_t|\mathbf{X}_\Delta^T]$ and $V[\hat{\mathbf{z}}_t\hat{\mathbf{z}}_{t-1}|\mathbf{X}_\Delta^T]$ via RTS **for** $i = 1, \ldots, N$ **do**
      repeat
        Update $\mathbf{W}_i^*$ according to Eq. 9.14 Update $\mathbf{B}_i^*$ according to Eq. 9.15
      until $\mathbf{W}_i$, $\mathbf{B}_i$ *converge*
      Update $\mathbf{A}_i^*$ according to Eq. 9.16
    end
    Update $\mathbf{A}^*, \mathbf{V}_{\hat{Z}}^*, \hat{\boldsymbol{\Sigma}}^*$ according to Section 9.4.2
  until *DPCCA converges*
  for $i = 1, \ldots, N$ **do**
    $\boldsymbol{\theta}_i = \left\{ \begin{bmatrix} \mathbf{A}_z & 0 \\ 0 & \mathbf{A}_i \end{bmatrix}, \mathbf{W}_i, \mathbf{B}_i, \begin{bmatrix} \mathbf{V}_\mathbf{Z} & 0 \\ 0 & \mathbf{V}_i \end{bmatrix}, \sigma_i^2\mathbf{I} \right\}$
    Estimate $\mathbb{E}[\hat{\mathbf{z}}_t|\mathbf{X}_i^T]$, $V[\hat{\mathbf{z}}_t|\mathbf{X}_i^T]$ and $V[\hat{\mathbf{z}}_t\hat{\mathbf{z}}_{t-1}|\mathbf{X}_i^T]$ via RTS on θ_i.
  end
until $\mathcal{L}_{DPCTW}$ *converges*
* Since $\mathbb{E}[\hat{\mathbf{z}}_t|\mathbf{X}_i^T]$ is unkown in the first iteration, use $\mathbf{X}_i$ instead.

9.6.1 Supervised-Generative DPCCA (SG-DPCCA)

We firstly consider the model where we simply augment the observation model with a set of features $\mathbf{Y}_j$. In this case, the generative model for DPCCA (Eq. 9.9) is:

$$\mathbf{x}_{i,t} = \mathbf{W}_{i,t}\mathbf{z}_t + \mathbf{B}_i\mathbf{z}_{i,t} + \epsilon_i, \tag{9.18}$$

$$\mathbf{y}_{j,t} = h_{j,s}(\mathbf{z}_t|\mathbf{W}_{j,t}) + h_{j,p}(\mathbf{z}_{j,t}|\mathbf{B}_j) + \epsilon_j, \tag{9.19}$$

where $i = \{1, \ldots, N\}$ and $j = \{N+1, \ldots, N+\nu+1\}$. The arbitrary functions h map the shared space to the feature space in a generative manner, while $\epsilon_j \sim \mathcal{N}(0, \sigma_j^2\mathbf{I})$. The latent priors are still defined as in Eq. 9.7,9.8. By assuming that h is linear, we can group the parameters $\mathbf{W} = [\mathbf{W}_1, \ldots, \mathbf{W}_N, \ldots, \mathbf{W}_{N+\nu}]$, $\mathbf{B}$ as the block diagonal of $([\mathbf{B}_1, \ldots, \mathbf{B}_N, \ldots, \mathbf{B}_{N+\nu}])$ and $\hat{\boldsymbol{\Sigma}}$ as the block diagonal of $([\sigma^2\mathbf{I}_1, \ldots, \sigma^2\mathbf{I}_N, \ldots, \sigma^2\mathbf{I}_{N+\nu}])$. Inference is subsequently applied as described in Section 9.4.

This model, which we dub SG-DPCCA, in effect captures a common shared space of both annotations $\mathbf{X}$ and available features $\mathbf{Y}$ for each sequence. In our generative scenario, the shared space generates both features and annotations. By further setting $h_{j,p}$ to zero, one can force the representation of the entire feature space $\mathbf{Y}_j$ onto the shared space, thus imposing stronger constraints on the shared space given each annotation $\mathbf{Z}|\mathbf{X}_i$. As we will show, this model can help identify unwanted annotations by simply analysing the posteriors of the shared latent space. We note that the additional form of supervision imposed by the input on the model is reminiscent of SPCA for PCA [284]. The discriminative ability added by the inputs (or labels) also relates DPCCA to LDA [11]. The graphical model of SG-DPCCA is illustrated in Fig. 9.2(b).

SG-DPCCA can be easily extended to handle time-warping as described in Section 9.5 for DPCCA (SG-DPCTW). The main difference is that now one would have to introduce one more warping function for each set of features, resulting in a set of $N+\nu$ functions. Denoting the complete data/input set as $\mathcal{D}^o = \{\mathbf{X}_1, \ldots, \mathbf{X}_N, \mathbf{Y}_1, \ldots, \mathbf{Y}_\nu\}$, the objective function for obtaining the time warping functions Δ_i for SG-DPCTW can be defined as:

$$\mathcal{L}_{SDPCTW^o} = \sum_{i}^{N+\nu} \sum_{j,j\neq i}^{N+\nu} \frac{||\mathbb{E}[\mathbf{Z}|\mathcal{D}_i^o]\mathbf{\Delta}_i - \mathbb{E}[\mathbf{Z}|\mathcal{D}_j^o]\mathbf{\Delta}_j||_F^2}{(N+\nu)(N+\nu-1)}. \tag{9.20}$$

9.6.2 Supervised-Discriminative DPCCA (SD-DPCCA)

The second model augments the DPCCA model by regressing on the given features. In this case, the posterior of the shared space (Eq. 9.7) is formulated as

$$p(\mathbf{z}_t|\mathbf{z}_{t-1}, \mathbf{Y}_{1:\nu}, \mathbf{A}, \mathbf{V}_{\hat{z}}) \sim$$
$$\mathcal{N}(\mathbf{A}_z\mathbf{z}_{t-1} + \sum_{j=1}^{\nu} h_j(\mathbf{Y}_j|\mathbf{F}_j), \mathbf{V}_z), \tag{9.21}$$

where each function h_j performs regression on the features $\mathbf{Y}_j$, while $\mathbf{F}_j \in \mathbb{R}^{d\times D_{y_j}}$ are the loadings for the features (where the latent dimensionality is d). This is similar to how input is modelled in a standard LDS [83]. To find the parameters, we maximise the complete-data likelihood (Eq. 9.12), where we replace the second term referring to the latent probability with Eq. 9.21,

$$\sum_{t=2}^{T} lnP(\hat{\mathbf{z}}_t|\hat{\mathbf{z}}_{t-1}, \mathbf{Y}_{1:\nu}, \mathbf{A}, \mathbf{V}_{\hat{z}}). \tag{9.22}$$

In this variation, the shared space at step t is generated from the previous latent state $\mathbf{z}_{t-1}$ as well as the features at step $t-1, \sum_{j=1}^{\nu} \mathbf{y}_{j,t-1}$ (Fig. 9.2(c)). We dub this model SD-DPCCA. Without loss of generality we assume h is linear, i.e. $h_{j,s} = \mathbf{W}_{j,t}\mathbf{z}_t$, while we model the feature signal only in the shared space, i.e. $h_{j,p} = 0$. Finding the saddle points of the derivatives with respect to the parameters yields the following updates for the matrices $\mathbf{A}_z$ and $\mathbf{F}_j, \forall j = 1, \ldots, \nu$:

$$\mathbf{A}_z^* = \left(\sum_{t=2}^{T} E[\mathbf{z}_t \mathbf{z}_{t-1}^T] - \sum_{j=1}^{\nu} \mathbf{F}_j^* \mathbf{y}_{j,t} \right) \left(\sum_{t=2}^{T} E[\mathbf{z}_{t-1} \mathbf{z}_{t-1}^T] \right)^{-1}, \tag{9.23}$$

$$\mathbf{F}_j^* = \left(\mathbb{E}[\mathbf{z}_t] - \mathbf{A}_z^* \mathbb{E}[\mathbf{z}_{t-1}] - \sum_{i=1, i \neq j}^{\nu} \mathbf{F}_i^* \mathbf{Y}_i \right) \mathbf{Y}_j^{-1}. \tag{9.24}$$

Note that as with the loadings on the shared/individual spaces ($\mathbf{W}$ and $\mathbf{B}$), the optimisation of $\mathbf{A}_z$ and $\mathbf{F}_j$ matrices should again be determined recursively. Finally, the estimation of $\mathbf{V_Z}$ also changes accordingly:

$$\begin{aligned} \mathbf{V}_\mathbf{z}^* = \; & \tfrac{1}{T-1} \textstyle\sum_{t=2}^{T} (\mathbb{E}[\mathbf{z}_t \mathbf{z}_t^T] - \mathbb{E}[\mathbf{z}_t \mathbf{z}_{t-1}^T] \mathbf{A}_z^{*T} \\ & - \mathbf{A}_z^* \mathbb{E}[\mathbf{z}_{t-1} \mathbf{z}_t^T] + \mathbf{A}_z^* \mathbb{E}[\mathbf{z}_{t-1} \mathbf{z}_{t-1}^T] \mathbf{A}_z^{*T} \\ & + \textstyle\sum_{j=1}^{\nu} (\mathbf{A}_z^* \mathbb{E}[\mathbf{z}_{t-1}] \mathbf{Y}_j^{*T} \mathbf{F}_j^{*T} + \mathbf{F}_j^* \mathbf{Y}_j \mathbb{E}[\mathbf{z}_{t-1}^T] \mathbf{A}_z^{*T} \\ & + \mathbf{F}_j^* \mathbf{Y}_j \textstyle\sum_{i=1, i \neq j}^{\nu} \mathbf{Y}_i^T \mathbf{F}_i^{*T} - \mathbb{E}[\mathbf{z}_t] \mathbf{Y}_j^T \mathbf{F}_j^{*T} \\ & - \mathbf{F}_j^* \mathbf{Y}_j \mathbb{E}[\mathbf{z}_t^T])). \end{aligned} \tag{9.25}$$

SD-DPCCA can be straight-forwardly extended with time-warping as with DPCCA in Section 9.5, resulting in SD-DPCTW. Another alignment step is required before performing the recursive updates mentioned above in order to find the correct training/testing pairs for $\mathbf{z}_t$ and $\mathbf{Y}$. Assuming the warping matrices are $\mathbf{\Delta}_z$ and $\mathbf{\Delta}_y$, then in Eq. 9.23 $\mathbf{z}$ is replaced with $\mathbf{\Delta}_z\mathbf{z}$ and $\mathbf{y}$ with $\mathbf{\Delta}_y\mathbf{y}$. The influence of features $\mathbf{Y}$ on the shared latent space $\mathbf{Z}$ in SD-DPCCA and SG-DPCCA is visualised in Fig. 9.2.

9.6.3 Varying Dimensionality

Typically, we would expect the dimensionality of a set of annotations to be the same. Nevertheless in certain problems, especially when using input features as in SG-DPCCA (Section

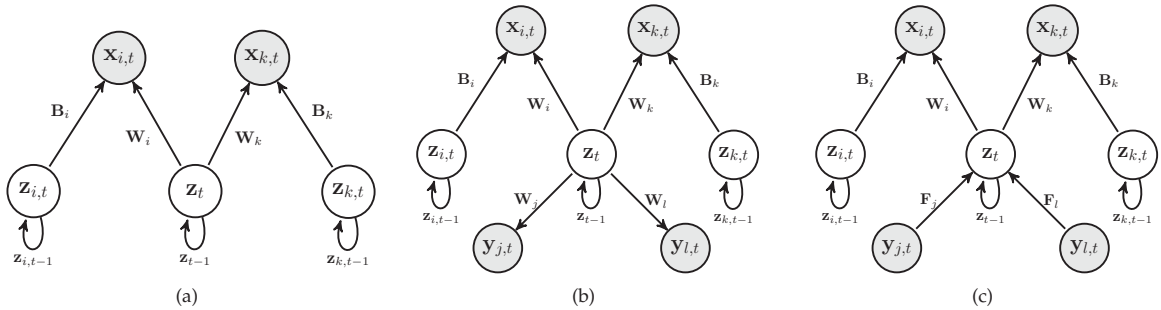


Figure 9.2: Comparing the model structure of DPCCA (a) to SG-DPCCA (b) and SD-DPCCA (c). Notice that the shared space $\mathbf{z}$ generates both observations and features in SG-DPCCA, while in SD-DPCCA, the shared space at time t is generated by regressing from the features $\mathbf{y}$ and the previous shared space state $\mathbf{z}_{t-1}$.

9.6.1), this is not the case. Therefore, in case the observations/input features are of varying dimensionalities, one can scale the third term of the likelihood (Eq. 9.12) in order to balance the influence of each sequence during learning regardless of its dimensionality:

$$\sum_{t=1}^{T}\left(\sum_{j=1}^{\nu}\frac{1}{D_{y_j}}ln\left(P(\mathbf{y}_{t,j}|\hat{\mathbf{z}}_t,\mathbf{W}_j,\mathbf{B}_j,\sigma_j^2)\right)+\sum_{j=1}^{N}\frac{1}{D_i}ln\left(P(\mathbf{x}_{t,j}|\hat{\mathbf{z}}_t,\mathbf{W}_j,\mathbf{B}_j,\sigma_i^2)\right)\right). \tag{9.26}$$

9.7 Ranking and Filtering Annotations

In this section, we will refer to the issue of ranking and filtering available annotations. Since in general, we consider that there is no "ground truth" available, it is not an easy task to infer which annotators should be discarded and which kept. A straightforward option would be to keep the set of annotators which exhibit a decent level of agreement with each other. Nevertheless, this naive criterion will not suffice in case where e.g., all the annotations exhibit moderate correlation, or where sets of annotations are clustered in groups which are intra-correlated but not inter-correlated.

The question that naturally arises is how to rank and evaluate the annotators when there

is no ground truth available and their inter-correlation is not helpful. We remind that DP-CCA maximises the correlation of the annotations in the shared space $\mathbf{Z}$, by removing bias, temporal discrepancies and other nuisances from each annotation. It would therefore be reasonable to expect the latent *posteriors* for each annotation $(\mathbf{Z}|\mathbf{X}_i)$, to be as close as possible. Furthermore, the closer the posterior given each annotation $(\mathbf{Z}|\mathbf{X}_i)$ to the posterior given all sequences $(\mathbf{Z}|\mathcal{D})$, the higher the ranking of the annotator should be, since the closer it is, the larger the portion of the shared information is contained in the annotators signal.

The aforementioned procedure can detect *spammers*, i.e. annotators who do not even pay attention at the sequence they are annotating and *adversarial* or *malicious* annotators that provide erroneous annotations due to e.g., a conflict of interests and can rank the confidence that should be assigned to the rest of the annotators. Nevertheless, it does not account for the case where multiple clusters of annotators are intra-correlated but not inter-correlated. In this case, it is most probable that the best-correlated group will prevail in the ground truth determination. Yet, this does not mean that the best-correlated group is the correct one. In this case, we propose using a set of inputs (e.g., tracking facial points), which can essentially represent the "gold standard". The assumption underlying this proposal is that the correct sequence features should maximally correlate with the correct annotations of the sequence. This can be straightforwardly performed with SG-DPCCA, where we attain $\mathbf{Z}|\mathbf{Y}$ (shared space given input) and compare to $\mathbf{Z}|\mathbf{X}_i$ (shared space given annotation i).

The comparison of latent posteriors is further motivated by R.J. Aumann's agreement theorem [10]: "If two people are Bayesian rationalists with common priors, and if they have common knowledge of their individual posteriors, then their posteriors must be equal". Since our model maintains the notion of "common knowledge" in the estimation of the shared space, it follows from Aumann's theorem that the individual posteriors $\mathbf{Z}|\mathbf{X}_i$ of each annotation i should be as close as possible. This is a sensible assumption, since one would expect that if all bias, temporal discrepancies and other nuisances are removed from annotations, then there is no rationale for the posteriors of the shared space to differ.

A simple algorithm for filtering/ranking annotations (utilising spectral clustering [233]) can be found in Algorithm 4. The goal of the algorithm is to find two clusters, C_x and C_o,

containing (i) the set of annotations which are correlated with the ground truth, and (ii) the set of "outlier" annotations, respectively. Firstly, DPCCA/DPCTW is applied. Subsequently, a similarity/distance matrix is constructed based on the posterior distances of each annotation $\mathbf{Z}|\mathbf{X}_i$ along with the features $\mathbf{Z}|\mathbf{Y}$. By performing spectral clustering, one can keep the cluster to which $\mathbf{Z}|\mathbf{Y}$ belongs (C_x) and disregard the rest of the annotations belonging in C_o. The ranking of the annotators is computed implicitly via the distance matrix, as it is the relative distance of each $\mathbf{Z}|\mathbf{X}_i$ to $\mathbf{Z}|\mathbf{Y}$. In other words, the feature posterior is used here as the "ground truth". Depending on the application (or in case features are not available), one can use the posterior given all annotations, $\mathbf{Z}|\mathbf{X}_1, \ldots, \mathbf{X}_N$ instead of $\mathbf{Z}|\mathbf{Y}$. Examples of distances/metrics that can be used include the alignment error (see Section 9.5) or the KL divergence between normal distributions (which can be made symmetric by employing e.g., the Jensen-Shannon divergence, i.e. $D_{JS}(P||Q) = \frac{1}{2}D_{KL}(P||Q) + \frac{1}{2}D_{KL}(Q||P)$).

Algorithm 4 Ranking and filtering annotators

Data: $\mathbf{X}_1, \ldots, \mathbf{X}_N, \mathbf{Y}$
Result: Rank of each $\mathbf{X}_i$, C_c
begin
Apply SG-DPCTW/SG-DPCCA($\mathbf{X_1}, \ldots, \mathbf{X_N}, \mathbf{Y}$) Obtain $P(\mathbf{Z}|\mathbf{Y})$, $P(\mathbf{Z}|\mathbf{X}_i), i = 1, \ldots, N$ Compute Distance Matrix $\mathbf{S}$ of $[P(\mathbf{Z}|\mathbf{X}_1), \ldots, P(\mathbf{Z}|\mathbf{X}_N), P(\mathbf{Z}|\mathbf{Y})]$ Normalise $\mathbf{S}$, $\mathbf{L} \leftarrow \mathbf{I} - \mathbf{D}^{-\frac{1}{2}} S \mathbf{D}^{-\frac{1}{2}}$ $\{C_x, C_o\} \leftarrow$ Spectral Clustering($\mathbf{L}$) Keep C_x where $P(\mathbf{Z}|\mathbf{Y}) \in C_x$ Rank each $X_i \in C_x$ based on distance of $P(\mathbf{Z}|\mathbf{X}_i)$ to $P(\mathbf{Z}|\mathbf{Y})$
end
In case $\mathbf{Y}$ is not available, replace $P(\mathbf{Z}|\mathbf{Y})$ with $P(\mathbf{Z}|\mathbf{X}_{1:N})$.

We note that in case of irrelevant or malicious annotations, we assume that the corresponding signals will be moved to the private space and will not interfere with the time warping. Nevertheless, in order to ensure this, one can impose constraints on the warping process. This is easily done by modifying the DTW by imposing e.g., slope or global constraints such as the Itakura Parallelogram or the Sakoe-Chiba band, in order to constraint the warping path while also decreasing the complexity (c.f., Chap. 5, of [205]). Furthermore, other heuristics can be applied, e.g. firstly filter out the most irrelevant annotations by applying SG-DPCCA without time warping, or threshold the warping objective directly (Eq. 9.17).

9.8 Experimental Evaluation

In order to evaluate the proposed models, in this section, we present a set of experiments on both synthetic (Section 9.8.1) and real (Section 9.8.2 & 9.8.3) data.

9.8.1 Synthetic Data

For synthetic experiments, we employ a setting similar to [298]. A set of 2D spirals are generated as $\mathbf{X}_i = \mathbf{U}_i^T \tilde{\mathbf{Z}} \mathbf{M}_i^T + \mathbf{N}$, where $\tilde{\mathbf{Z}} \in \mathbb{R}^{2\times T}$ is the true latent signal which generates the $\mathbf{X}_i$, while the $\mathbf{U}_i \in \mathbb{R}^{2\times 2}$ and $\mathbf{M}_i \in \mathbb{R}^{T_i \times m}$ matrices impose random spatial and temporal warping. The signal is furthermore perturbed by additive noise via the matrix $\mathbf{N} \in \mathbb{R}^{2\times T}$. Each $\mathbf{N}(i,j) = e \times b$, where $e \sim \mathcal{N}(0,1)$ and b follows a Bernoulli distribution with $P(b=1) = 1$ for Gaussian and $P(b=1) = 0.4$ for spike noise. The length of the synthetic sequences varies, but is approximately 200.

This experiment can be interpreted as both of the problems we are examining. Viewed as a sequence alignment problem the goal is to recover the alignment of each noisy $\mathbf{X}_i$, where in this case the true alignment is known. Considering the problem of fusing multiple annotations, the latent signal $\tilde{\mathbf{Z}}$ represents the true annotation while the individual $\mathbf{X}_i$ form the set of noisy annotations containing annotation-specific characteristics. The goal is to recover the true latent signal (in DPCCA terms, $\mathbb{E}[\mathbf{Z}|\mathbf{X}_1, \ldots, \mathbf{X}_N]$).

The error metric we used computes the distance from the ground truth alignment ($\tilde{\mathbf{\Delta}}$) to the alignment recovered by each algorithm ($\mathbf{\Delta}$) [296], and is defined as:

$$\begin{aligned} \text{error} &= \frac{\text{dist}(\mathbf{\Pi}, \tilde{\mathbf{\Pi}}) + \text{dist}(\tilde{\mathbf{\Pi}}, \mathbf{\Pi})}{T_\Delta + \tilde{T}_\Delta}, \\ \text{dist}(\mathbf{\Pi}_1, \mathbf{\Pi}_2) &= \sum_{i=1}^{T_\Delta^1} \min(\{||\pi_1^{(i)} - \pi_2^{(j)}||\})_{j=1}^{T_\Delta^2}), \end{aligned} \tag{9.27}$$

where $\mathbf{\Pi}_i \in \mathbb{R}^{T_\Delta^i \times N}$ contains the indices corresponding to the binary selection matrices $\mathbf{\Delta}_i$, as defined in Section 9.5.1 (and [296]), while $\pi^{(j)}$ refers to the j-th row of $\mathbf{\Pi}$. For qualitative evaluation, in Fig. 9.3, we present an example of applying (D)PCTW on 5 sequences. As can be seen, DPCTW is able to recover the true, de-noised, latent signal which generated the

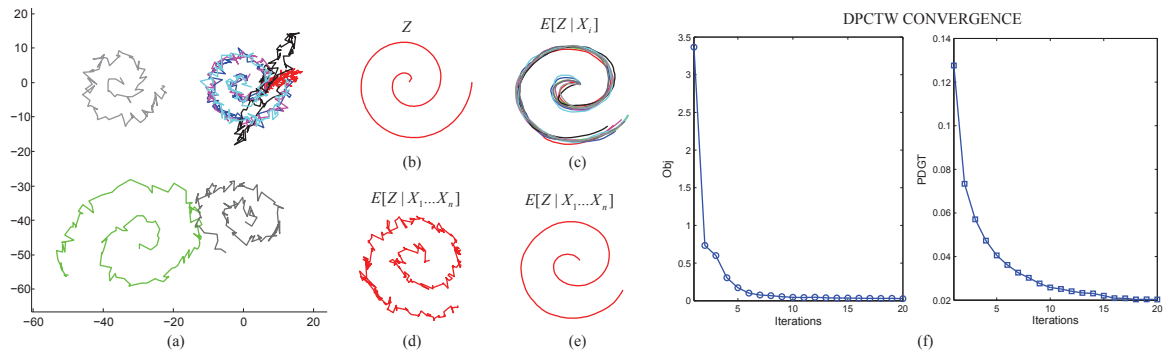


Figure 9.3: Noisy synthetic experiment. (a) Initial, noisy time series. (b) True latent signal from which the noisy, transformed spirals where attained in (a). (c) The alignment achieved by DPCTW. The shared latent space recovered by (d) PCTW and (e) DPCTW. (f) Convergence of DPCTW in terms of the objective (Obj) (Eq. 9.17) and the path difference between the estimated alignment and the true alignment path (PDGT).

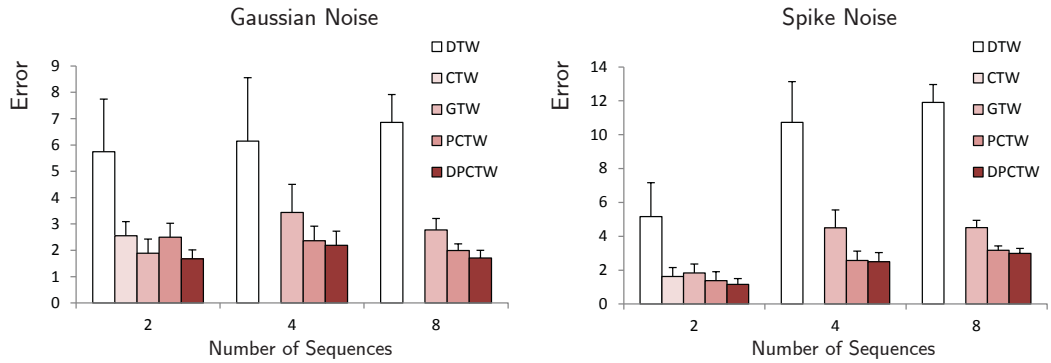


Figure 9.4: Synthetic experiment comparing the alignment attained by DTW, CTW, GTW, PCTW and DPCTW on spirals with spiked and Gaussian noise.

noisy observations (Fig. 9.3(e)), while also aligning the noisy sequences (Fig. 9.3(c)). Due to the temporal modelling of DPCTW, the recovered latent space is almost identical to the true signal $\tilde{\mathbf{Z}}$ (Fig. 9.3(b)). PCTW on the other hand is unable to entirely remove the noise (Fig. 9.3(d)). Fig. 9.4 shows further results comparing related methods. CTW and GTW perform comparably for two sequences, both outperforming DTW. In general, PCTW seems to perform better than CTW, while DPCTW provides better alignment than other methods compared.

9.8.2 Real Data I: Fusing Multiple Annotations

In order to evaluate (D)PCTW in case of real data, we employ the SEMAINE database [158]. The database contains a set of audio-visual recordings of subjects interacting with operators. Each operator assumes a certain personality - happy, gloomy, angry and pragmatic - with a goal of inducing spontaneous emotions by the subject during a naturalistic conversation. We use a portion of the database containing recordings of 6 different subjects, from over 40 different recording sessions, with a maximum length of 6000 frames per segment. As the database was annotated in terms of emotion dimensions by a set of experts (varying from 2 to 8), no single ground truth is provided along with the recordings. Thus, by considering $\mathbf{X}$ to be the set of annotations and applying (D)PCTW, we obtain $\mathbb{E}[\mathbf{Z}|\mathcal{D}] \in \mathbb{R}^{1\times T}$ (given all *warped* annotations)[10], which represents the shared latent space with annotator-specific factors and noise removed. We assume that $\mathbb{E}[\mathbf{Z}|\mathcal{D}]$ represents the ground truth. An example of this procedure for (D)PCTW can be found in Fig. 9.5. As can be seen, DPCTW provides a smooth, aligned estimate, eliminating temporal discrepancies, spike-noise and annotator bias. In this experiment, we evaluate the proposed models on four emotion dimensions: valence, arousal, power, and anticipation (expectation).

To obtain features for evaluating the ground truth, we track the facial expressions of each subject via a particle filtering tracking scheme [190]. The tracked points include the corners of the eyebrows (4 points), the eyes (8 points), the nose (3 points), the mouth (4 points) and the chin (1 point), resulting in 20 2D points for each frame.

For evaluation, we consider a training sequence $\mathbf{X}$, for which the set of annotations $\mathcal{A}_x = \{\mathbf{a}_1, \ldots, \mathbf{a}_R\}$ is known. From this set ($\mathcal{A}_x$), we derive the ground truth $\mathcal{GT}_{\mathbf{X}}$ - for (D)PCTW, $\mathcal{GT}_{\mathbf{X}} = \mathbb{E}[\mathbf{Z}|\mathcal{A}_x]$. Using the tracked points $\mathcal{P}_{\mathbf{X}}$ for the sequence, we train a regressor to learn the function $f_x : \mathcal{P}_{\mathbf{X}} \rightarrow \mathcal{GT}_{\mathbf{X}}$. In (D)PCTW, $\mathcal{P}_x$ is firstly aligned with $\mathcal{GT}_x$ as they are not necessarily of equal length. Subsequently given a testing sequence $\mathbf{Y}$ with tracked points $\mathcal{P}_y$, using f_x we predict each emotion dimension ($f_x(\mathcal{P}_y)$). The procedure for deriving the ground truth is then applied on the annotations of sequence $\mathbf{Y}$, and the resulting $\mathcal{GT}_y$ is evaluated

[10] We note that latent (D)PCTW posteriors used, e.g. $\mathbf{Z}|\mathbf{X}_i$ are obtained on time-warped observations, e.g. $\mathbf{Z}|\mathbf{X}_i\mathbf{\Delta_i}$ (See Alg. 3)

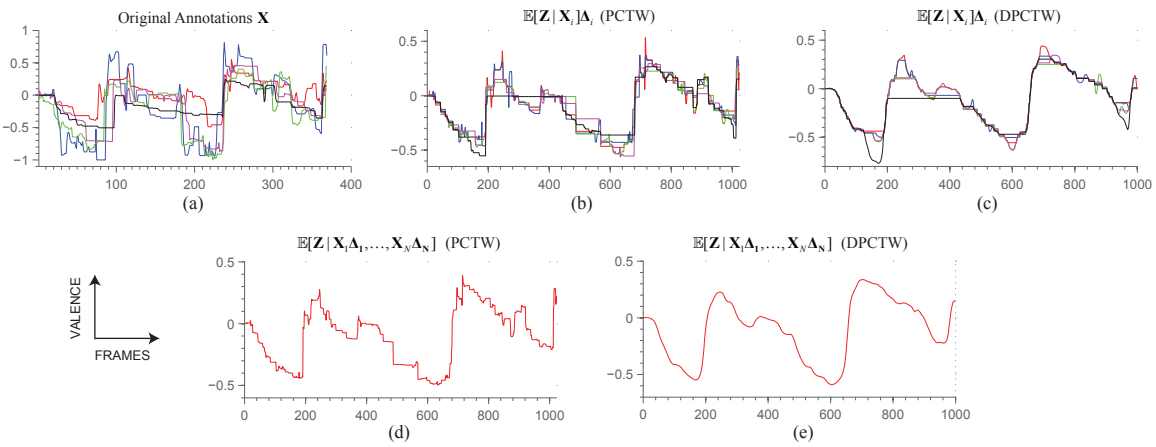


Figure 9.5: Applying (D)PCTW to continuous emotion annotations. (a) Original valence annotations from 5 experts. (b,c) Alignment obtained by PCTW and DPCTW respectively, (d,e) Shared space obtained by PCTW and DPCTW respectively, which can be considered as the "derived ground truth".

against $f_x(\mathcal{P}_y)$. The correlation coefficient of the $\mathcal{GT}_y$ and $f_x(\mathcal{P}_y)$ (after the two signals are temporally aligned) is then used as the evaluation metric for *all* compared methods.

The reasoning behind this experiment is that the "best" estimation of the ground truth (i.e. the gold standard) should maximally correlate with the corresponding input features - thus enabling any regressor to learn the mapping function more accurately.

We also perform experiments with the supervised variants of DPCTW, i.e. SG-DPCTW and SD-DPCTW. In this case, a set of features $\mathbf{Y}$ is used for inferring the ground truth, $\mathbf{Z}|\mathcal{D}$. Since we already used the facial trackings for evaluation, in order to avoid biasing our results[11], we use features from the audio domain. In particular, we extract a set of acoustic features consisting of 6 mel-frequency Cepstrum Coefficients (MFCC), 6 MFCC-Delta coefficients along with prosody features (signal energy, root mean squared energy and pitch), resulting in a 15 dimensional feature vector. The acoustic features are used to derive the ground truth with our supervised models, exactly acting an objective reference to our sequence. In this way, we

[11] Since we use the facial points for *evaluating* the derived ground truth, if we had also used them for *deriving* the ground truth we would bias the evaluation procedure.

impose a further constraint on the latent space: it should also explain the audio cues and not only the annotations, given that the two sets are correlated. Subsequently, the procedure described above for unsupervised evaluation with facial trackings is employed.

For regression, we employ RVM [246] with a Gaussian kernel. We perform both session-dependent experiments, where the validation was performed on each session separately, and session-independent experiments where different sessions were used for training/testing. In this way, we validate the derived ground truth generalisation ability (i) when the set of annotators is the same and (ii) when the set of annotators may differ.

Session-dependent and session-independent results are presented in Tables 9.1 and 9.2. We firstly discuss the unsupervised methods. As can be seen, taking a simple annotator average (A-AVG) gives the worse results (as expected), with a very high standard deviation and weak correlation. The model of Raykar et al. [209] provides better results, which can be justified by the variance estimation for each annotator. Modelling annotator bias and noise with (D)PCCA further improves the results. It is important to note that incorporating alignment is significant for deriving the ground truth; this is reasonable since when the annotations are misaligned, shared information may be modelled as individual factors or vice-versa. Thus, PCTW improves the results further while DPCTW provides the best results, confirming our assumption that combining dynamics, temporal alignment, modelling noise and individual-annotator bias leads to a more objective ground truth. Finally, regarding supervised models SG-DPCTW and SD-DPCTW, we can observe that the inclusion of acoustic features in the ground truth generation improves the results, with SG-DPCTW providing better correlated results than SD-DPCTW. This is reasonable since in SG-DPCTW the features $\mathbf{Y}$ are explicitly generated from the shared space, thus imposing a form of strict supervision, in comparison to SD-DPCTW where the inputs essentially elicit the shared space.

Ranking Annotations

We perform the ranking of annotations as proposed in Algorithm 4 to a set of emotion dimension annotations from the SEMAINE database.

In Fig. 9.6(a), we illustrate an example where an irrelevant structured annotation (sinusoid),

Table 9.1: Comparison of ground truth evaluation based on the correlation coefficient (COR), on session dependent experiments. The standard deviation over all results is denoted by σ.

	SD-DPCTW		SG-DPCTW		DPCTW		PCTW		DPCCA		PCCA		RAYKAR [209]		A-AVG	
	COR	σ	COR	σ	COR	σ	COR	σ	COR	σ	COR	σ	COR	σ	COR	σ
Valence	**0.78**	*0.18*	**0.78**	*0.17*	**0.77**	*0.18*	0.70	*0.18*	0.64	*0.21*	0.63	*0.20*	0.61	*0.20*	0.54	*0.36*
Arousal	0.75	*0.18*	**0.77**	*0.19*	**0.75**	*0.22*	0.64	*0.22*	0.63	*0.23*	0.63	*0.26*	0.60	*0.25*	0.42	*0.41*
Power	0.78	*0.13*	**0.85**	*0.10*	0.77	*0.16*	0.76	*0.10*	0.68	*0.16*	0.67	*0.18*	0.62	*0.22*	0.42	*0.36*
Expectation	0.82	*0.09*	**0.83**	*0.10*	0.78	*0.11*	0.75	*0.16*	0.68	*0.16*	0.74	*0.17*	0.62	*0.20*	0.48	*0.40*

Table 9.2: Comparison of ground truth evaluation based on the correlation coefficient (COR), on session independent experiments. The standard deviation over all results is denoted by σ.

	SD-DPCTW		SG-DPCTW		DPCTW		PCTW		DPCCA		PCCA		RAYKAR [209]		A-AVG	
	COR	σ	COR	σ	COR	σ	COR	σ	COR	σ	COR	σ	COR	σ	COR	σ
Valence	**0.73**	*0.19*	**0.73**	*0.19*	**0.72**	*0.22*	0.66	*0.24*	0.62	*0.28*	0.58	*0.23*	0.57	*0.27*	0.53	*0.33*
Arousal	**0.74**	*0.15*	**0.74**	*0.17*	**0.71**	*0.20*	0.61	*0.23*	0.59	*0.23*	0.52	*0.28*	0.50	*0.29*	0.33	*0.40*
Power	0.72	*0.28*	**0.75**	*0.24*	**0.72**	*0.34*	0.70	*0.19*	0.60	*0.26*	0.58	*0.27*	0.57	*0.27*	*0.39*	*0.31*
Expectation	0.76	*0.21*	**0.76**	*0.15*	**0.73**	*0.20*	0.70	*0.18*	0.63	*0.20*	0.64	*0.25*	0.63	*0.22*	0.44	*0.39*

has been added to a set of five true annotations. Obviously the sinusoid can be considered a spammer annotation since essentially, it is independent of the actual sequence at hand. In the figure we can see that (i) the derived ground truth is not affected by the spammer annotation, (ii) the spammer annotation is completely captured in the private space, and (iii) that the spammer annotation is detected in the distance matrix of $\mathbb{E}[\mathbf{Z}|\mathbf{X}_i]$ and $\mathbb{E}[\mathbf{Z}|\mathbf{X}]$.

In Fig. 9.6(b), we present an example where a set of 5 annotations has been used along with 8 spammers. The spammers consist of random Gaussian distributions along with structured periodical signals (i.e. sinusoids). We can see that it is difficult to discriminate the spammers by analysing the distance matrix of $\mathbf{X}$ since they do maintain some correlation with the true annotations. By applying Algorithm 4, we obtain the distance matrix of the latent posteriors $\mathbf{Z}|\mathbf{X}_i$ and $\mathbf{Z}|\mathcal{D}$. In this case, we can clearly detect the cluster of annotators which we should keep. By applying spectral clustering, the spammer annotations are isolated in a single cluster, while the shared space along with the true annotations fall into the other cluster. This is also obvious by observing the inferred weight vector ($\mathbf{W}$), which is near-zero for sequences 6-14, implying that the shared signal is ignored when reconstructing the specific annotation (i.e. the reconstruction is entirely from the private space). Finally, this is also obvious by calculating the KL divergence comparing each individual posterior $\mathbf{Z}|X_i$ to the shared space

posterior given all annotations $\mathbf{Z}|\mathcal{D}$, where sequences 6-14 have a high distance while 1-5 have a distance which is very close to zero.

In Fig. 9.6(c), we present another example where in this case, we joined two sets of annotations which were recorded for two distinct sequences (annotators 1-6 for sequence A and annotators 7-12 for sequence B). In the distance matrix taken on the observations $\mathbf{X}$, we can see how the two clusters of annotators are already discriminable, with the second cluster, consisting of annotations for sequence B, appearing more correlated. We use the facial trackings for sequence A (tracked as described in this section) as the features $\mathbf{Y}$, and then apply Algorithm 4. As can be seen in the distance matrix of $[\mathbf{Z}|\mathbf{X}_i, \mathbf{Z}|\mathbf{Y}]$, (i) the two clusters of annotators have been clearly separated, and (ii) the posterior of features $\mathbf{Z}|\mathbf{Y}$ clearly is much closer to annotations 1-6, which are the true annotations of sequence A.

9.8.3 Real Data II: Action Unit Alignment

In this experiment we aim to evaluate the performance of (D)PCTW for the temporal alignment of facial expressions. Such applications can be useful for methods which require pre-aligned data, e.g. AAM (Active Appearance Models). For this experiment, we use a portion of the MMI database which contains more than 300 videos, ranging from 100 to 200 frames. Each video is annotated (per frame) in terms of the temporal phases of each Action Unit (AU) manifested by the subject being recorded, namely neutral, onset, apex and offset. For this experiment, we track the facial expressions of each subject capturing 20 2D points, as in Section 9.8.2.

Given a set of videos where the same AU is activated by the subjects, the goal is to temporally align the phases of each AU activation across *all* videos containing that AU, where the facial points are used as features. In the context of DPCTW, each $\mathbf{X}_i$ is the facial points of video i containing the same AU, while $\mathbf{Z}|\mathbf{X}_i$ is now the common latent space given video i, the size of which is determined by cross-validation, and is constant over all experiments for a specific noise level.

In Fig. 9.7 we present results based on the number of misaligned frames for AU alignment, on all action unit temporal phases (neutral, onset, apex, offset) for AU 12 (smile), on a set of 50 pairs of videos from MMI. For this experiment, we used the facial features relating to the lower

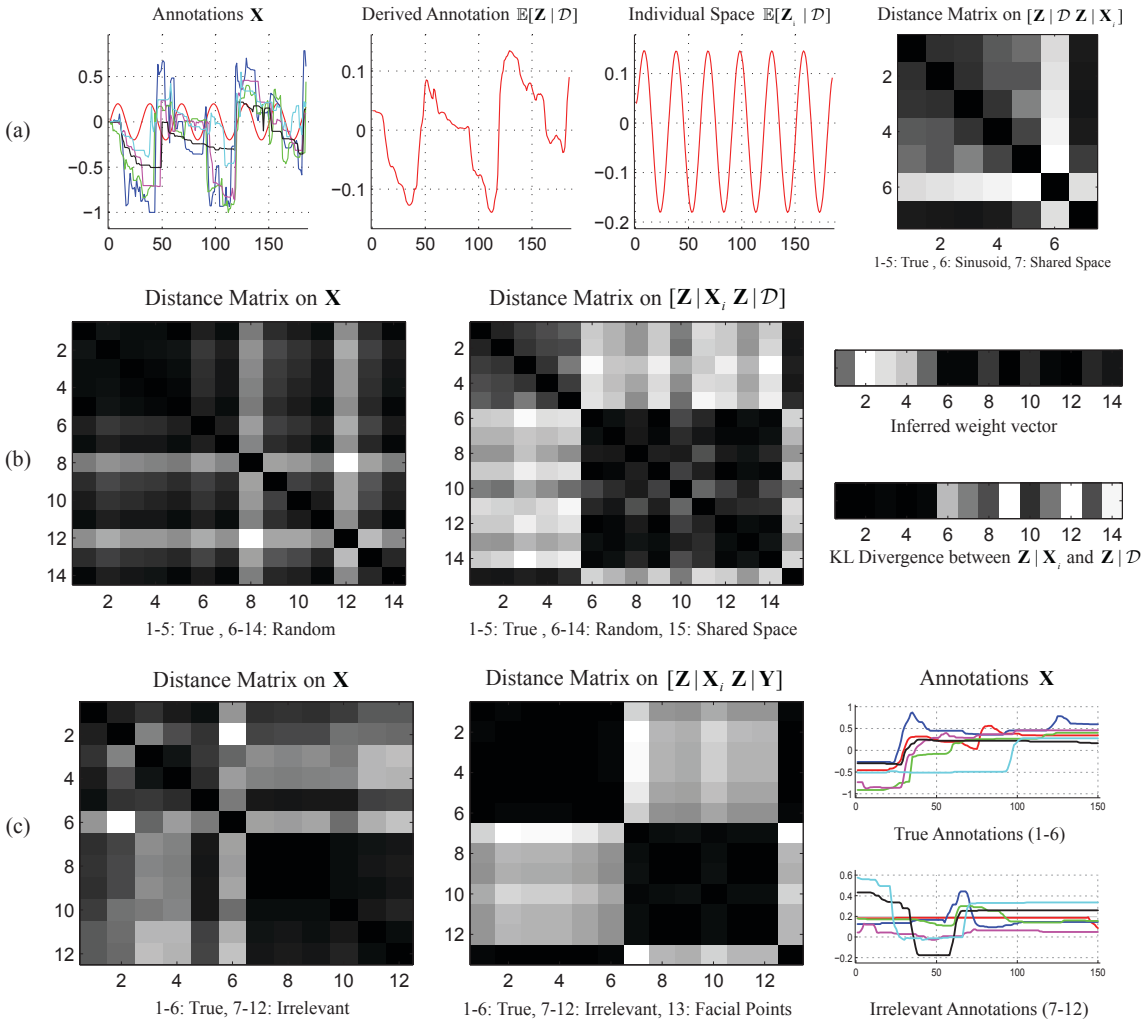


Figure 9.6: Annotation filtering and ranking (black - low, white - high). (a) Experiment with a structured false annotation (sinusoid). The shared space is not affected by the false annotation, which is isolated in the individual space. (b) Experiment with 5 true and 9 spammer (random) annotations. (c) Experiment with 6 true annotations, 7 irrelevant but correlated annotations (belonging to a different sequence). The facial points $\mathbf{Y}$, corresponding to the 6 true annotations, were used for supervision (with SG-DPCCA).

face, which consist of 11 2D points. The features were perturbed with sparse spike noise in order to simulate the misdetection of points with detection-based trackers, in order to evaluate the robustness of the proposed techniques. Values were drawn from the normal distribution $\mathcal{N}(0,1)$ and added (uniformly) to 5% of the length of each video. We gradually increased the number of features perturbed by noise from 0 to 4. To evaluate the accuracy of each algorithm, we use a robust, normalised metric. In more detail, let us say that we have two videos, with features $\mathbf{X}_1$ and $\mathbf{X}_2$, and AU annotations $\mathcal{A}_1$ and $\mathcal{A}_2$. Based on the features, the algorithm at hand recovers the alignment matrices Δ_1 and Δ_2. By applying the alignment matrices on the AU annotations ($\mathcal{A}_1\Delta_1$ and $\mathcal{A}_2\Delta_2$), we know to which temporal phase of the AU each aligned frame of each video corresponds to. Therefore, for a given temporal phase (e.g., neutral), we have a set of frame indices which are assigned to the specific temporal phase in video 1, Ph_1 and video 2, Ph_2. The accuracy is then estimated as $\frac{Ph_1 \cap Ph_2}{Ph_1 \cup Ph_2}$. This essentially corresponds to the ratio of correctly aligned frames to the total duration of the temporal phase accross the aligned videos.

As can be seen in the average results in Fig. 9.7, the best performance is clearly obtained by DPCTW. It is also interesting to highlight the accuracy of DPCTW on detecting the apex, which essentially is the peak of the expression. This can be attributed to the modelling of dynamics, not only in the shared latent space of all facial point sequences but also in the domain of the individual characteristics of each sequence (in this case identifying and removing the added temporal spiked noise). PCTW peforms better on average compared than CTW and GTW, while the latter two methods perform similarly. It is interesting to note that GTW seems to overpeform CTW and PCTW for aligning the apex of the expression for higher noise levels. Furthermore, we point-out that the Gauss-Newton warping used in GTW is likely to perform better for longer sequences. Example frames from videos showing the unaligned and DPCTW-aligned videos are shown in Fig. 9.8 .

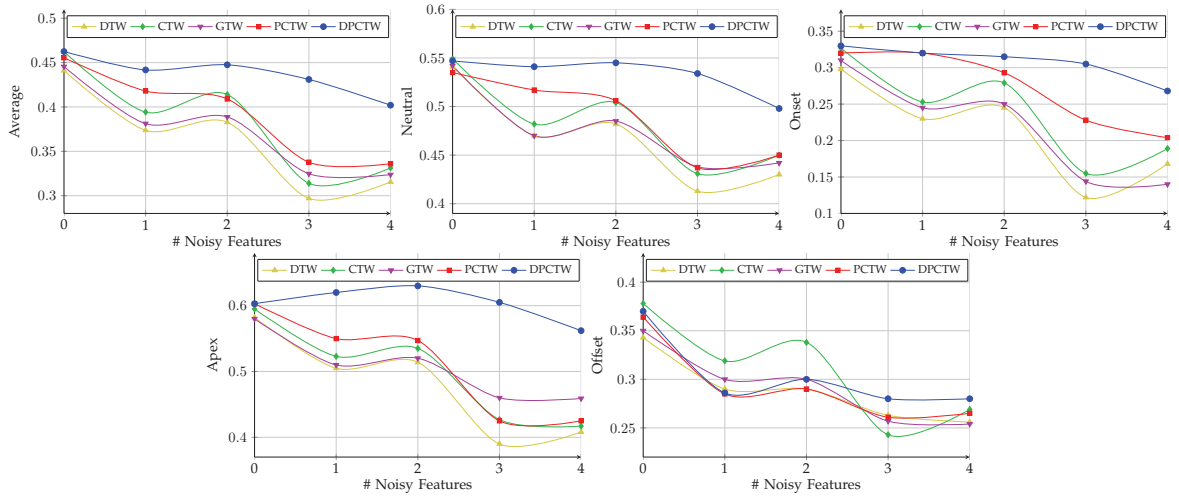


Figure 9.7: Accuracy of DTW, CTW, GTW, PCTW and DPCTW on the problem of action unit alignment under spiked noise added to an increasing number of features for AU = 12 (smile).

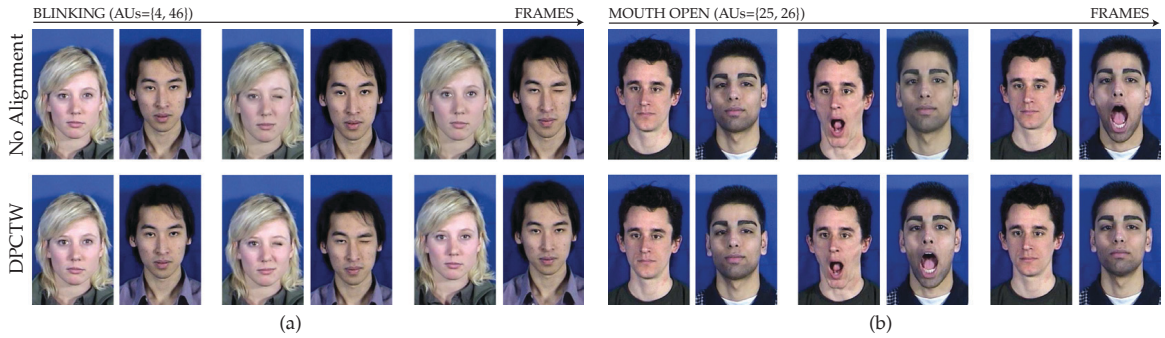


Figure 9.8: Example stills from a set of videos from the MMI database, comparing the original videos to the aligned videos obtained via DPCTW under spiked noise on 4 2D points. (a) Blinking, AUs 4 and 46. (b) Mouth open, AUs 25 and 27.

9.9 Conclusions

In this work, we presented DPCCA, a novel, dynamic and probabilistic model based on the multiset probabilistic interpretation of CCA. By integrating DPCCA with time warping, we proposed DPCTW, which can be interpreted as a unifying framework for solving the problems of (i) fusing multiple imperfect annotations and (ii) aligning temporal sequences. Furthermore, we extended DPCCA/DPCTW to a supervised scenario, where one can exploit inputs and observations, both in a discriminative and generative framework. We show that the family of probabilistic models which we present is this chapter is able to rank and filter annotators merely by utilising inferred model statistics. Finally, our experiments show that DPCTW features such as temporal alignment, learning dynamics, identifying individual annotator/sequence factors and incorporating inputs are critical for robust performance of fusion in challenging affective behaviour analysis tasks.

Chapter 10

Robust Canonical Correlation Analysis with Time Warpings

Contents

10.1 Introduction

In this chapter, we present a robust variant of Canonical Correlation Analysis (CCA) which we coin Robust Canonical Correlation Analysis (RCCA). The main advantage of this method lies in being able to model non-Gaussian, sparse noise which commonly occurs in real-world scenarios and applications, unlike methods based on traditional Canonical Correlation Analysis (CCA). Via RCCA, we decompose the observed sequences into a low-dimensional, low-rank component and a sparse component which models gross noise. RCCA thus facilitates the fusion of sequences arising from different modalities while being corrupted with gross noise. Subsequently, we turn to the problem of the accurate temporal alignment of se-

quences under gross noise, a highly challenging problem arising in fields such as computer vision [86, 114, 265, 297, 298, 296], bioinformatics [144] and speech processing [120, 219]. In Chapter 3 we introduced Dynamic Time Warping (DTW) and other related CCA based warping techniques, while in Chapter 9, where we aimed at the fusion of multiple annotations, we pointed out the lack of these models in terms related to the problem, such as learning temporal dynamics and explicitly modelling noise. In this chapter, we focus on the temporal alignment of high dimensional data under the presence of gross noise. In fact, most extensions of Dynamic Time Warping are also based on CCA [298] and while successful, they inherit the same issues as CCA in the presence of gross Gaussian noise. We extend RCCA to handle temporal warpings in RCTW, by learning low-rank projections while simultaneously finding the temporal alignment that maximises the spatial correlation in the error-free space. In other words, the RCTW aligns the corrupted sequences in a error-free common low-rank latent subspace which is robustly estimated, even in the presence of gross errors. The projections are obtained by minimizing the weighted sum of nuclear and ℓ_1 norms, by solving a sequence of convex optimization problems, while the temporal alignment is found by applying the DTW in an alternating fashion. The RCCA and RCTW models are mainly motivated by the success of *robust principal component analysis* (RPCA) [37] and *inductive RPCA* (IRPCA) [15] in gross error correction, and especially from the successful combination of rank minimization principles with spatial alignment [194]. Summarising, the contributions of this chapter are as follows.

- A novel method, i.e. RCCA is proposed as a robust-to-gross-errors variant of CCA. RCCA is further extended to a novel, robust method for the temporal alignment of high-dimensional data sequences despite large occlusions and corruptions.
- An efficient algorithm for RCCA and RCTW is derived by solving a sequence of convex problems. Each of the these convex problems is solved efficiently by employing first-order optimization techniques.
- Different sets of experiments on synthetic and real data validate that the proposed RCCA manages to robustly fuse multiple modalities and features, while RCTW accurately aligns grossly corrupted data sequences compared to state-of-the-art alignment methods.

The chapter is organized as follows. We introduce both the RCCA and RCTW in Section 10.2. Subsequently, in Section 10.3 we perform various experiments utilising both models with experiments both on synthetic and real data. In terms of real data, we utilise RCCA for (i) the fusion of audio-visual data for the detection of interest, and (ii) heterogeneous face recognition. We evaluate RCTW on (i) the temporal alignment of human walking sequences, and (ii) on the problem of temporal action unit alignment. Conclusions are drawn in Section 10.4.

10.2 Methodology

Canonical Correlation Analysis (CCA), as introduced in Chapter 3, is a shared-space component analysis method which is typically utilised for problems such as the fusion of multiple observation sets and modalities, multi-view analysis [231, 45], while often is utilised along with Dynamic Time Warping (DTW) in order to achieve the temporal alignment of multiple sequences with varying dimensionality and varying length [296, 298, 232]. The classical formulation of CCA carries the assumption that all errors follow a Gaussian distribution with a small variance. This is a typical assumption employed in most machine learning systems thus far, as employing Gaussian noise is the simplest assumption one can make without further complicating the model at hand. In this chapter, we propose a robust variant of CCA that is able to handle gross errors in the observation sets and is thus suitable for deployment under real-world conditions, where such errors are in abundance.

More formally, we assume $\mathbf{X}$ and $\mathbf{Y}$ to be high-dimensional observation sets, likely corrupted with gross noise, where $\mathbf{X} \in \mathbb{R}^{dx \times T}$ and $\mathbf{Y} \in \mathbb{R}^{dy \times T}$. Each of these observation sets may represent e.g., features extracted from modalities to be fused (e.g., facial trackings and audio cues). The Robust Canonical Correlation Analysis (RCCA), based on the desired low-rankness of the projections as well as the sparsity of the noise terms, can be formulated as

$$\begin{aligned} &\underset{\mathbf{P}_x,\mathbf{P}_y,\mathbf{E}_x,\mathbf{E}_y}{\operatorname{argmin}} \quad \operatorname{rank}(\mathbf{P}_x) + \operatorname{rank}(\mathbf{P}_y) \\ &+\lambda_1\|\mathbf{E}_x\|_0 + \lambda_2\|\mathbf{E}_y\|_0 + \frac{\mu}{2}\|\mathbf{P}_x\mathbf{X} - \mathbf{P}_y\mathbf{Y}\|_F^2 \\ &\text{s.t. } \mathbf{X} = \mathbf{P}_x\mathbf{X} + \mathbf{E}_y, \mathbf{Y} = \mathbf{P}_y\mathbf{Y} + \mathbf{E}_y. \end{aligned} \tag{10.1}$$

where as can be seen, RCCA uncovers a low-rank subspace $\mathbf{P}_x$, $\mathbf{P}_y$, by estimating the gross distortion terms for each modality, $\mathbf{E}_x$ and $\mathbf{E}_y$, where λ_1, λ_2 and μ are non-negative parameters.

Unfortunately, Problem (10.1) is deemed difficult to solve due to the discrete nature of the rank function [261] and the ℓ_0 norm [170]. Nevertheless, it has been proved that the convex envelope of the ℓ_0 norm is the ℓ_1 norm [63], while the convex envelope of the rank function is the nuclear norm [78]. Therefore, convex relaxations of (10.1) can be obtained by replacing the ℓ_0 norm and the rank function with their convex envelopes. The resulting problem

$$\begin{aligned} &\underset{\mathbf{P}_z,\mathbf{P}_a,\mathbf{E}_z,\mathbf{E}_a}{\operatorname{argmin}} \quad \|\mathbf{P}_z\|_* + \|\mathbf{P}_a\|_* \\ &+\lambda_1\|\mathbf{E}_z\|_1 + \lambda_2\|\mathbf{E}_a\|_1 + \frac{\mu}{2}\|\mathbf{P}_z\mathbf{Z} - \mathbf{P}_a\mathbf{A}\|_F^2 \\ &\text{s.t. } \mathbf{Z} = \mathbf{P}_z\mathbf{Z} + \mathbf{E}_z, \mathbf{A} = \mathbf{P}_a\mathbf{A} + \mathbf{E}_a. \end{aligned} \tag{10.2}$$

can be solved by employing the Linearized Alternating Directions Method (LADM) [143], a variant of the *alternating direction augmented lagrange multiplier method* [25]. The algorithm is detailed in Alg. 5. We note that the singular value thresholding operator can be defined for any matrix $\mathbf{M}$ [35], as: $\mathcal{D}_\tau[\mathbf{M}] = \mathbf{U}\mathcal{S}_\tau\mathbf{V}^T$ where $\mathbf{M} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^T$ is the singular value decomposition (SVD) and $\mathcal{S}_\tau[q] = \text{sign}(q)\max(|q| - \tau, 0)$ the shrinkage operator [37] (extended to matrices via element-wise application). Before moving on to discuss the optimisation, we move on to discuss the extension of this model to include time warpings, as this is in fact a more general case which includes the RCCA.

10.2.1 RCCA with Time Warpings (RCTW)

Problem (10.1) can be easily extended in order to handle sequences of different lengths. Dynamic Time Warping (DTW) has already been described in Chapter 3, but to make this chapter self-sufficient we summarise the definition here. Given two data sequences $\mathbf{X} = [\mathbf{x}_1|\mathbf{x}_2|\dots|\mathbf{x}_{T_x}] \in \mathbb{R}^{d\times T_x}$ and $\mathbf{Y} = [\mathbf{y}_1|\mathbf{y}_2|\dots|\mathbf{y}_{T_y}] \in \mathbb{R}^{d\times T_y}$, the DTW aligns the sequences by solving [219]:

$$\underset{\mathbf{\Delta}_x,\mathbf{\Delta}_y}{\operatorname{argmin}} \ \frac{1}{2}\|\mathbf{X}\mathbf{\Delta}_x - \mathbf{Y}\mathbf{\Delta}_y\|_F^2, \ \text{ s.t. } \ \mathbf{\Delta}_x \in \{0,1\}^{T_x\times T}, \mathbf{\Delta}_y \in \{0,1\}^{T_y\times T}, \tag{10.3}$$

where $\mathbf{\Delta}_x$ and $\mathbf{\Delta}_y$ are binary selection matrices encoding the alignment path. Although the number of possible alignments is exponential in T_xT_y, the DTW is able to recover the optimal alignment path in $\mathcal{O}(T_xT_y)$ by employing dynamic programming.

In order to extend RCCA with time warpings, we formulate an analogous problem to (10.1) as follows.

$$\begin{aligned}\underset{\substack{\mathbf{P}_x,\mathbf{P}_y,\mathbf{E}_x,\\ \mathbf{E}_y,\mathbf{\Delta}_x,\mathbf{\Delta}_y}}{\operatorname{argmin}} \quad & \operatorname{rank}(\mathbf{P}_x)+\operatorname{rank}(\mathbf{P}_y)+\lambda_x\|\mathbf{E}_x\|_0+\lambda_y\|\mathbf{E}_y\|_0 \\ & +\frac{\mu}{2}\|\mathbf{P}_x\mathbf{X}\mathbf{\Delta}_x-\mathbf{P}_y\mathbf{Y}\mathbf{\Delta}_y\|_F^2 \\ \text{s.t.} \quad & \mathbf{X}=\mathbf{P}_x\mathbf{X}+\mathbf{E}_x,\mathbf{Y}=\mathbf{P}_y\mathbf{Y}+\mathbf{E}_y, \\ & \mathbf{\Delta}_x\in\{0,1\}^{T_x\times T},\mathbf{\Delta}_y\in\{0,1\}^{T_y\times T}.\end{aligned} \tag{10.4}$$

where now, the observations $\mathbf{X}$ and $\mathbf{Y}$ are of varying time lengths, i.e. $\mathbf{X}\in\mathbb{R}^{dx\times T_x}$ and $\mathbf{Y}\in\mathbb{R}^{dy\times T_y}$, while $\mathbf{\Delta}_x$ and $\mathbf{\Delta}_y$ are binary selection matrices, same as in the DTW case, which encode the alignment path. Similarly to Problem 10.1, Problem 10.4 is difficult to solve, and by adopting a convex relaxation as above, we arrive at:

$$\begin{aligned}\underset{\substack{\mathbf{P}_x,\mathbf{P}_y,\mathbf{E}_x,\\ \mathbf{E}_y,\mathbf{\Delta}_x,\mathbf{\Delta}_y}}{\operatorname{argmin}} \quad & \|\mathbf{P}_x\|_*+\|\mathbf{P}_y\|_*+\lambda_x\|\mathbf{E}_x\|_1+\lambda_y\|\mathbf{E}_y\|_1 \\ & +\frac{\mu}{2}\|\mathbf{P}_x\mathbf{X}\mathbf{\Delta}_x-\mathbf{P}_y\mathbf{Y}\mathbf{\Delta}_y\|_F^2 \\ \text{s.t.} \quad & \mathbf{X}=\mathbf{P}_x\mathbf{X}+\mathbf{E}_x,\mathbf{Y}=\mathbf{P}_y\mathbf{Y}+\mathbf{E}_y, \\ & \mathbf{\Delta}_x\in\{0,1\}^{T_x\times T},\mathbf{\Delta}_y\in\{0,1\}^{T_y\times T}.\end{aligned} \tag{10.5}$$

where accordingly to Problem (10.2), can be solved via LADM. In what follows, we describe the LADM solution to the RCTW problem defined above. Note that the solution for RCCA problem (in cases our samples are aligned in time and of same length) can be equivalently obtained by setting $\mathbf{\Delta_x}=\mathbf{\Delta_y}=\mathbf{I}$, i.e. by simply omitting the Δ_x and Δ_y matrices and of course, the update step for these parameters.

As aforementioned, problem (10.5) can be solved iteratively by employing the *linearized alternating directions method* (LADM) [143], a variant of the *alternating direction augmented*

Lagrange multiplier method (ADM) [25]. That is, (10.5) is solved by minimizing the (partial) augmented Lagrangian function:

$$\begin{aligned}
&\mathcal{L}(\mathbf{P}_x, \mathbf{P}_y, \mathbf{E}_x, \mathbf{E}_y, \boldsymbol{\Delta}_x, \boldsymbol{\Delta}_y, \boldsymbol{\Lambda}_1, \boldsymbol{\Lambda}_2) \\
&= \|\mathbf{P}_x\|_* + \|\mathbf{P}_y\|_* + \lambda_x \|\mathbf{E}_x\|_y + \lambda_2 \|\mathbf{E}_y\|_1 \\
&+ \frac{\mu}{2} \|\mathbf{P}_x \mathbf{X} \boldsymbol{\Delta}_x - \mathbf{P}_y \mathbf{Y} \boldsymbol{\Delta}_y\|_F^2 \\
&+ \operatorname{tr}\left(\boldsymbol{\Lambda}_1{}^T (\mathbf{X} - \mathbf{P}_x \mathbf{X} - \mathbf{E}_x)\right) \\
&+ \operatorname{tr}\left(\boldsymbol{\Lambda}_2{}^T (\mathbf{Y} - \mathbf{P}_y \mathbf{Y} - \mathbf{E}_y)\right) \\
&+ \frac{\mu_x}{2} \|\mathbf{X} - \mathbf{P}_x \mathbf{X} - \mathbf{E}_x\|_F^2 + \frac{\mu_y}{2} \|\mathbf{Y} - \mathbf{P}_y \mathbf{Y} - \mathbf{E}_y\|_F^2 \\
&\text{s.t. } \boldsymbol{\Delta}_x \in \{0,1\}^{T_x \times T}, \boldsymbol{\Delta}_y \in \{0,1\}^{T_y \times T},
\end{aligned} \tag{10.6}$$

where $\boldsymbol{\Lambda}_1, \boldsymbol{\Lambda}_2$ are the Lagrange multipliers for the equality constraints in (10.5) and μ_x, μ_y are nonnegative penalty parameters. By employing the LADM, (10.6) is minimized with respect to each variable in an alternating fashion and finally the Lagrange multipliers are updated at each iteration as outlined in Algorithm 5. The derivation of Algorithm 5 is provided next.

If only $\mathbf{P}_x$ is varying and all the other variables are kept fixed, we simplify (10.6) writing $\mathcal{L}(\mathbf{P}_x)$ instead of $\mathcal{L}(\mathbf{P}_x, \mathbf{P}_y, \mathbf{E}_x, \mathbf{E}_y, \boldsymbol{\Delta}_x, \boldsymbol{\Delta}_y, \boldsymbol{\Lambda}_1, \boldsymbol{\Lambda}_2)$. Let t denote the iteration index, given $\mathbf{P}_{x[t]}$, $\mathbf{P}_{y[t]}$, $\mathbf{E}_{x[t]}$, $\mathbf{E}_{y[t]}$, $\boldsymbol{\Delta}_{x[t]}$, $\boldsymbol{\Delta}_{y[t]}$,$\boldsymbol{\Lambda}_{1[t]}$, and $\boldsymbol{\Lambda}_{2[t]}$, the iterative scheme of LADM for (10.6) reads as follows:

$$\mathbf{P}_{x[t+1]} = \underset{\mathbf{P}_{x[t]}}{\operatorname{argmin}} \mathcal{L}(\mathbf{P}_{x[t]}) \tag{10.7}$$

$$\mathbf{E}_{x[t+1]} = \underset{\mathbf{E}_{x[t]}}{\operatorname{argmin}} \mathcal{L}(\mathbf{E}_{x[t]}) \tag{10.8}$$

$$\mathbf{P}_{y[t+1]} = \underset{\mathbf{P}_{y[t]}}{\operatorname{argmin}} \mathcal{L}(\mathbf{P}_{y[t]}) \tag{10.9}$$

$$\mathbf{E}_{y[t+1]} = \underset{\mathbf{E}_{y[t]}}{\operatorname{argmin}} \mathcal{L}(\mathbf{E}_{y[t]}) \tag{10.10}$$

$$(\boldsymbol{\Delta}_{x[t+1]}, \boldsymbol{\Delta}_{y[t+1]}) = \underset{\boldsymbol{\Delta}_{x[t]}, \boldsymbol{\Delta}_{y[t]}}{\operatorname{argmin}} \mathcal{L}(\boldsymbol{\Delta}_{x[t]}, \boldsymbol{\Delta}_{y[t]}) \tag{10.11}$$

Solving subproblems (10.7) and (10.9). By fixing the other variables, subproblem (10.7) is

reduced to

$$\begin{aligned}&\underset{\mathbf{P}_{x[t]}}{\operatorname{argmin}} \|\mathbf{P}_x\|_* + \frac{\mu}{2}\|\mathbf{P}_x\mathbf{X}\boldsymbol{\Delta}_x - \mathbf{P}_y\mathbf{Y}\boldsymbol{\Delta}_y\|_F^2 \\ &+ \operatorname{tr}\left({\boldsymbol{\Lambda}_1}^T(\mathbf{X} - \mathbf{P}_x\mathbf{X} - \mathbf{E}_x)\right) + \frac{\mu_x}{2}\|\mathbf{X} - \mathbf{P}_x\mathbf{X} - \mathbf{E}_x\|_F^2.\end{aligned} \tag{10.12}$$

Although the standard procedure for solving nuclear norm regularized least squares problems is the *singular value thresholding* operator [35], it cannot be directly applied in case of (10.12), due to the existence of the second term (i.e., $\frac{\mu}{2}\|\mathbf{P}_x\mathbf{X}\boldsymbol{\Delta}_x - \mathbf{P}_y\mathbf{Y}\boldsymbol{\Delta}_y\|_F^2$). To this end, following [143], the differentiable terms in (10.12) i.e., the function $f(\mathbf{P}_x) = \frac{\mu}{2}\|\mathbf{P}_x\mathbf{X}\boldsymbol{\Delta}_x - \mathbf{P}_y\mathbf{Y}\boldsymbol{\Delta}_y\|_F^2 + \operatorname{tr}\left({\boldsymbol{\Lambda}_1}^T(\mathbf{X} - \mathbf{P}_x\mathbf{X} - \mathbf{E}_x)\right) + \frac{\mu_x}{2}\|\mathbf{X} - \mathbf{P}_x\mathbf{X} - \mathbf{E}_x\|_F^2$ is linearly approximated with respect to $\mathbf{P}_x$ at $\mathbf{P}_{x[t]}$ as follows:

$$\begin{aligned}f(\mathbf{P}_x) \approx f(\mathbf{P}_{x[t]}) &+ \operatorname{tr}\left((\mathbf{P}_x - \mathbf{P}_{x[t]})^T\nabla f(\mathbf{P}_{x[t]})\right) \\ &+ \frac{\mu_x\eta_x}{2}\|\mathbf{P}_x - \mathbf{P}_{x[t]}\|_F^2,\end{aligned} \tag{10.13}$$

where, η_x is a proximal parameter. The gradient of $f(\mathbf{P}_{x[t]})$ with respect to $\mathbf{P}_{x[t]}$ is given by:

$$\begin{aligned}&\nabla f(\mathbf{P}_{x[t]}) = \mu_x(\mathbf{P}_{x[t]}\mathbf{X}\mathbf{X}^T + \mathbf{E}_{x[t]}\mathbf{X}^T - \mathbf{X}\mathbf{X}^T) \\ &+ \mu(\mathbf{P}_{x[t]}\mathbf{X}\boldsymbol{\Delta}_{x[t]}\boldsymbol{\Delta}_{x[t]}^T\mathbf{X}^T - \mathbf{P}_{y[t]}\mathbf{Y}\boldsymbol{\Delta}_{y[t]}\boldsymbol{\Delta}_{x[t]}^T\mathbf{X}^T) \\ &- \boldsymbol{\Lambda}_{1[t]}\mathbf{X}^T.\end{aligned} \tag{10.14}$$

Consequently, an approximate solution of (10.12) can be obtained as follows:

$$\begin{aligned}&\mathbf{P}_{x[t+1]} \approx \underset{\mathbf{P}_x}{\operatorname{argmin}} \|\mathbf{P}_x\|_* + f(\mathbf{P}_{x[t]}) \\ &+ \operatorname{tr}\left((\mathbf{P}_x - \mathbf{P}_{x[t]})^T\nabla f(\mathbf{P}_{x[t]})\right) + \frac{\mu_x\eta_x}{2}\|\mathbf{P}_x - \mathbf{P}_{x[t]}\|_F^2 \\ &= \underset{\mathbf{P}_x}{\operatorname{argmin}} \|\mathbf{P}_x\|_* + \frac{\mu_x\eta_x}{2}\|\mathbf{P}_x - (\mathbf{P}_{[t]} - \frac{1}{\mu_x\eta_x}\nabla f(\mathbf{P}_{x[t]})\|_F^2 \\ &= \mathcal{D}_{\frac{1}{\mu_x\eta_x}}\left[\mathbf{P}_{x[t]} - \frac{1}{\mu_x\eta_x}\nabla f(\mathbf{P}_{x[t]})\right].\end{aligned} \tag{10.15}$$

The singular value thresholding operator defined for any matrix $\mathbf{Q}$ as [35]: $\mathcal{D}_\tau[\mathbf{Q}] = \mathbf{U}\mathcal{S}_\tau\mathbf{V}^T$ with $\mathbf{Q} = \mathbf{U}\boldsymbol{\Sigma}\mathbf{V}^T$ being the singular value decomposition and $\mathcal{S}_\tau[q] = \operatorname{sgn}(q)\max(|q| - \tau, 0)$ is the shrinkage operator [37], which can be extended to matrices by applying it element-wise.

The solution of (10.9) in analogy with (10.7) is given by

$$\mathbf{P}_{y[t+1]} = \mathcal{D}_{\frac{1}{\mu_y \eta_y}} \left[\mathbf{P}_{y[t]} - \frac{1}{\mu_y \eta_y} \nabla f(\mathbf{P}_{y[t]}) \right], \tag{10.16}$$

where $\nabla f(\mathbf{P}_{y[t]}) = \mu_x(\mathbf{P}_{y[t]}\mathbf{Y}\mathbf{Y}^T + \mathbf{E}_{y[t]}\mathbf{Y}^T - \mathbf{Y}\mathbf{Y}^T) + \mu(\mathbf{P}_{y[t]}\mathbf{Y}\boldsymbol{\Delta}_{y[t]}\boldsymbol{\Delta}_{y[t]}^T\mathbf{Y}^T - \mathbf{P}_{x[t]}\mathbf{X}\boldsymbol{\Delta}_{x[t]}\boldsymbol{\Delta}_{y[t]}^T\mathbf{Y}^T) - \boldsymbol{\Lambda}_{2[t]}\mathbf{Y}^T$.

Solving subproblems (10.8) and (10.10). By fixing the other variables, subproblem (10.8) is reduced to

$$\begin{aligned} \underset{\mathbf{E}_{x[t]}}{\operatorname{argmin}} & \lambda_x \|\mathbf{E}_x\|_1 + \operatorname{tr}\left(\boldsymbol{\Lambda}_1{}^T(\mathbf{X} - \mathbf{P}_x\mathbf{X} - \mathbf{E}_x)\right) \\ & + \frac{\mu_x}{2}\|\mathbf{X} - \mathbf{P}_x\mathbf{X} - \mathbf{E}_x\|_F^2. \end{aligned} \tag{10.17}$$

The subgradient of (10.17) provides a closed-form solution for $\mathbf{E}_{x[t+1]}$ by employing the shrinkage operator:

$$\mathbf{E}_{x[t+1]} = \mathcal{S}_{\frac{\lambda_x}{\mu_x}}[\mathbf{X} - \mathbf{P}_{x[t+1]}\mathbf{X} + \frac{1}{\mu_x}\boldsymbol{\Lambda}_{1[t]}]. \tag{10.18}$$

In a similar manner to (10.8), the solution of (10.10) is given by:

$$\mathbf{E}_{y[t+1]} = \mathcal{S}_{\frac{\lambda_y}{\mu_y}}[\mathbf{Y} - \mathbf{P}_{y[t+1]}\mathbf{Y} + \frac{1}{\mu_y}\boldsymbol{\Lambda}_{2[t]}]. \tag{10.19}$$

Solving (10.11). Subproblem (10.11) is solved by applying the DTW on the clean latent spaces defined by $\mathbf{P}_{x[t+1]}\mathbf{X}, \mathbf{P}_{y[t+1]}\mathbf{Y}$. Thus the warping matrices are obtained as follows:

$$[\boldsymbol{\Delta}_{x[t+1]}, \boldsymbol{\Delta}_{y[t+1]}] = \mathrm{DTW}(\mathbf{P}_{x[t+1]}\mathbf{X}, \mathbf{P}_{y[t+1]}\mathbf{Y}). \tag{10.20}$$

The Algorithm 1 terminates when the following criteria are satisfied [143]:

$$\begin{aligned} \max\Big(&\frac{\|\mathbf{X} - \mathbf{P}_{x[t+1]}\mathbf{X} - \mathbf{E}_{x[t+1]}\|_F}{\|\mathbf{X}\|_F}, \\ &\frac{\|\mathbf{Y} - \mathbf{P}_{y[t+1]}\mathbf{Y} - \mathbf{E}_{y[t+1]}\|_F}{\|\mathbf{Y}\|_F}\Big) < \epsilon_1, \end{aligned} \tag{10.21}$$

and

$$\begin{aligned} \max\Big(&\frac{\|\mathbf{P}_{x[t+1]} - \mathbf{P}_{x[t]}\|_F}{\|\mathbf{X}\|_F}, \frac{\|\mathbf{P}_{y[t+1]} - \mathbf{P}_{y[t]}\|_F}{\|\mathbf{Y}\|_F}, \\ &\frac{\|\mathbf{E}_{x[t+1]} - \mathbf{E}_{x[t]}\|_F}{\|\mathbf{X}\|_F}, \frac{\|\mathbf{E}_{y[t+1]} - \mathbf{E}_{y[t]}\|_F}{\|\mathbf{Y}\|_F}\Big) < \epsilon_2. \end{aligned} \tag{10.22}$$

Algorithm 5 Solving (10.6) by the LADM method.

Input: Data sequences: $\mathbf{X} \in \mathbb{R}^{d \times T_x}$ and $\mathbf{Y} =\in \mathbb{R}^{d \times T_y}$, parameters: $\lambda_x = 1/\sqrt{\max(d, T_x)}$, $\lambda_y = 1/\sqrt{\max(d, T_y)}$.
Output: The projection matrices: $\mathbf{P}_x, \mathbf{P}_y$, the warping matrices $\mathbf{\Delta}_x, \mathbf{\Delta}_y$, and the error matrices $\mathbf{E}_x, \mathbf{E}_y$.

Initialize: Set $\mathbf{P}_{x[0]}, \mathbf{P}_{y[0]}, \mathbf{E}_{x[0]}$, and $\mathbf{E}_{y[0]}$ to zero matrices of compatible dimensions. Initialize $\mathbf{\Delta}_{x[0]}$ and $\mathbf{\Delta}_{y[0]}$ by the DTW. $t = 0$ $\mu_{[0]} = \mu_{x[0]} = \mu_{y[0]} = 10^{-6}$, $\rho = 1.9$, $\eta_x = 1.02\sigma_x^2$, $\eta_y = 1.02\sigma_y^2$, where σ_x, σ_y are the largest singular values of $\mathbf{X}$ and $\mathbf{Y}$, respectively. $\epsilon_1 = 10^{-4}, \epsilon_2 = 10^{-5}$.
while not converged **do**
  Fix the other variables, and update $\mathbf{P}_{x[t+1]}$ by:
  $\mathbf{P}_{x[t+1]} \leftarrow \mathcal{D}_{\frac{1}{\mu_{x[t]}\eta_x}}[\mathbf{P}_{x[t]} - 1/(\mu_{x[t]} \cdot \eta_x)\nabla f(\mathbf{P}_{x[t]})]$.
  Fix the other variables, and update $\mathbf{E}_{x[t+1]}$ by:
  $\mathbf{E}_{x[t+1]} \leftarrow \mathcal{S}_{\frac{\lambda_1}{\mu_{x[t]}}}[\mathbf{X} - \mathbf{P}_{x[t+1]}\mathbf{X} + \frac{1}{\mu_{x[t]}}\mathbf{\Lambda}_{1[t]}]$.
  Fix the other variables, and update $\mathbf{P}_{y[t+1]}$ by:
  $\mathbf{P}_{y[t+1]} \leftarrow \mathcal{D}_{\frac{1}{\mu_{y[t]}\eta_y}}[\mathbf{P}_{y[t]} - 1/(\mu_{y[t]} \cdot \eta_y)\nabla f(\mathbf{P}_{y[t]})]$.
  Fix the other variables, and update $\mathbf{E}_{y[t+1]}$ by:
  $\mathbf{E}_{y[t+1]} \leftarrow \mathcal{S}_{\frac{\lambda_2}{\mu_{y[t]}}}[\mathbf{Y} - \mathbf{P}_{y[t+1]}\mathbf{Y} + \frac{1}{\mu_{y[t]}}\mathbf{\Lambda}_{2[t]}]$.
  Fix the other variables, and update the warping paths $\mathbf{\Delta}_{x[t+1]}, \mathbf{\Delta}_{y[t+1]}$ by:
  $[\mathbf{\Delta}_{x[t+1]}, \mathbf{\Delta}_{y[t+1]}] \leftarrow \mathrm{DTW}(\mathbf{P}_{x[t+1]}\mathbf{X}, \mathbf{P}_{y[t+1]}\mathbf{Y})$.
  Update the Lagrange multipliers by:
  $\mathbf{\Lambda}_{1[t+1]} \leftarrow \mathbf{\Lambda}_{1[t]} + \mu_{x[t]}(\mathbf{X} - \mathbf{P}_{x[t+1]}\mathbf{X} - \mathbf{E}_{x[t+1]})$.
  $\mathbf{\Lambda}_{2[t+1]} \leftarrow \mathbf{\Lambda}_{2[t]} + \mu_{y[t]}(\mathbf{Y} - \mathbf{P}_{y[t+1]}\mathbf{Y} - \mathbf{E}_{y[t+1]})$.
  Update $\mu_{x[t+1]}$ by:
  if $\mu_{x[t]}\|\mathbf{P}_{x[t+1]} - \mathbf{P}_{x[t]}\|_F / \|\mathbf{X}\|_F \leq \epsilon_2$ **then**
    $\mu_{x[t+1]} \leftarrow \min(\rho \cdot \mu_{x[t]}, 10^6)$.
  end if
  if $\mu_{y[t]}\|\mathbf{P}_{y[t+1]} - \mathbf{P}_{y[t]}\|_F / \|\mathbf{Y}\|_F \leq \epsilon_2$ **then**
    $\mu_{y[t+1]} \leftarrow \min(\rho \cdot \mu_{y[t]}, 10^6)$.
  end if
  Update $\mu_{[t+1]}$ by: $\mu_{[t+1]} \leftarrow \min(\mu_{x[t+1]}, \mu_{y[t+1]})$
  Check convergence conditions in (10.22) and (10.21).
  $t \leftarrow t + 1$.
end while

The dominant cost of each iteration in Algorithm 5 is the computation the singular value thresholding operator (i.e., Step 3 and Step 5). Thus, the complexity of each iteration is $\mathcal{O}(d^2 \cdot T)$. Regarding the convergence of Algorithm 5, there is no established convergence proof of the ADM for more than two blocks of variables [25, 194].Nevertheless, weak convergence results can be derived if the block of variables is assumed to be bounded. However, the application of ADM in optimization problems with more than two blocks of variables (e.g., [15, 194]) yields algorithms whose convergence is empirically guaranteed. This can be attributed to the convexity of (10.6) with respect to all the blocks of variables.

If the dimensions of the data sequence are different i.e., $\mathbf{X} \in \mathbb{R}^{d_x \times T_x}$ and $\mathbf{Y} \in \mathbb{R}^{d_y \times T_y}$ with $d_y \neq d_x$, then the dimensionality of the largest sequence can be reduced to that of the smallest by a random projection matrix drawn from a normal zero-mean distribution. Such a random projection matrix provides with high probability a *stable embedding* [16] preserving the Euclidean distances between all vectors in the original space in the feature space of reduced dimensions. Furthermore, if both data sequences are high-dimensional such as videos, random projections could be applied to both of the for computational tractability.

10.3 Experimental Evaluation

In this section, we evaluate the performance of RCCA and RCTW with several experiments, both on real and synthetic data. Firstly, we show a set of synthetic experiments in Section 10.3.1. Subsequently, we evaluate RCCA on problems relating to fusion, including audio-visual fusion for the prediction of interest (Section 10.3.2) as well as Heterogeneous Face Recognition and Matching (Section 10.3.3). We compare to state-of-the-art CCA variants, such as the classical CCA, the Common Orthogonal Basis Extraction (COBE) [299], the Joint and Individual Variation Explained (JIVE) [146], as well as least-squares formulations of CCA utilising $l1$ and $l2$ norms [240]. Subsequently, we evaluate RCTW on problems such as the temporal alignment of human behaviour against state-of-the-art temporal alignment methods, namely the CTW [298] and the GTW [296]. We note that the alignment error, similarly to

Chapter 9, is evaluated by employing the following metric [296]:

$$\text{Err} = \frac{\text{dist}(\mathbf{\Pi}^*, \hat{\mathbf{\Pi}}) + \text{dist}(\hat{\mathbf{\Pi}}, \mathbf{\Pi}^*)}{m^* + \hat{m}},$$
$$\text{dist}(\mathbf{\Pi}_1, \mathbf{\Pi}_2) = \sum_{i=1}^{m_1} \min(\{\|\pi_1^{(i)} - \pi_2^{(j)}\|_2\})_{j=1}^{m_2}), \tag{10.23}$$

where m^* is the length of $\mathbf{\Pi}^*$ and $\hat{m}$ is the length $\hat{\mathbf{\Pi}}$.

10.3.1 Synthetic Data

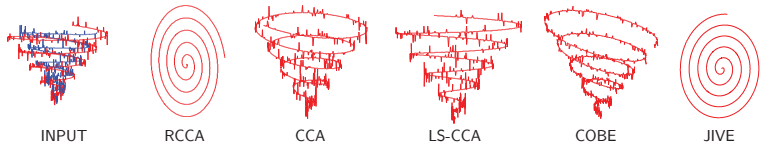


Figure 10.1: Evaluating RCCA and other compared methods on a given input of distorted by noise 3D spirals.

For the synthetic experiments a similar setting to [298] was employed, utilising 3D spirals. We note that this experiment is mostly focused on temporal warping, but as a proof of concept, in Fig. 10.1 we compare the resulting subspace when given a set of 3D spirals distorted by additive non-Gaussian noise. Clearly, RCCA is able to isolate the noise in the error matrices and infer the clean latent space. Regarding RCTW, we generate the 3D spirals data as follows: $\mathbf{X} = \mathbf{S}_x\mathbf{Z}\mathbf{T}_x \in \mathbb{R}^{3\times T_x}$, $\mathbf{Y} = \mathbf{S}_y\mathbf{Z}\mathbf{T}_y \in \mathbb{R}^{3\times T_y}$, where $\mathbf{Z} \in \mathbb{R}^{3\times T}$ is the true latent data sequence. $\mathbf{S}_x, \mathbf{S}_y \in \mathbb{R}^{3\times 3}$ and $\mathbf{T}_x \in \mathbb{R}^{T_x\times T}, \mathbf{T}_y \in \mathbb{R}^{T_y\times T}$ are random spatial and temporal warping matrices, respectively. Next, both $\mathbf{X}$ and $\mathbf{Y}$ are corrupted by adding gross non-gaussian noise to a percentage of samples (i.e., columns of $\mathbf{X}$ and $\mathbf{Y}$) ranging from 5 to 55%. In Fig. 10.2 we present averaged results on 50 data sequences, where the latent data sequence $\mathbf{Z}$ is perturbed by 50 different random spatial and temporal transformations. The mean alignment error of the compared techniques is presented in Fig. 10.5a. It is clear from the results that RCTW outperforms the compared approaches, exhibiting a stable and low path alignment error.

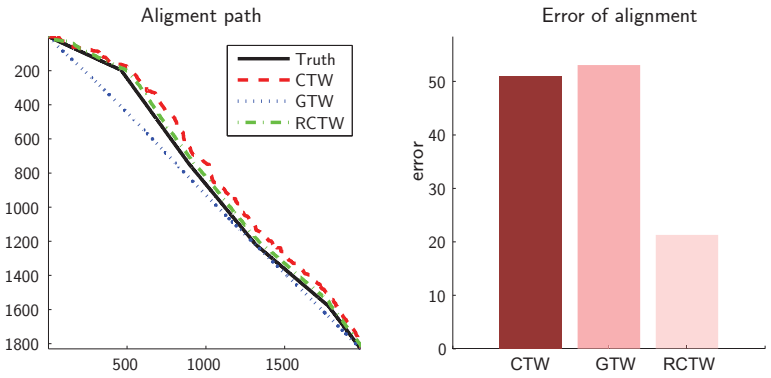


Figure 10.2: Comparison of the performance of the CTW, the GTW, and the RCTW on synthetic data alignment. The mean alignment path (left) and the mean alignment error (right) obtained by the CTW, the GTW, and the RCTW (left) by applying 50 different random spatial and temporal transformations on the latent data sequence $\mathbf{Z}$.

10.3.2 Audio-Visual Fusion via RCCA

The automatic detection of the level of interest in audiovisual sequences is a problem which has been gaining rising attention in the field of machine learning and pattern recognition [195, 227, 228], as it has crucial value for a vast span of applications such as affect-sensitive interfaces, interactive learning systems etc. In this section, we evaluate RCCA on the problem of fusion multi-modal signals for the automatic estimation of the level of interest. The experimental setting we follow is precisely the same as the one used in Chapter 7, where we utilise interest annotations we obtained for the SEMAINE database [157], which contains recordings of naturalistic dyadic interactions (Chapter 2). Similarly to Chapter 7, we utilise an Active Appearance Model (AAM) based tracker [182], designed for simultaneous tracking of 3D head pose, lips, eyebrows, eyelids and irises in videos and thus obtain 113 2D-points, resulting in an 226 dimensional feature vector per frame, while utilising 13 MFCC cepstrum coefficients for each audio frameCross-validation is performed given the features and annotations. Regression was performed via a Relevance Vector Machine (RVM) [246] (c.f., Chapter 3). Given the input-output pair $(\mathbf{x}_i, \mathbf{y}_i)$, RVM models the function $\mathbf{y}_i = \mathbf{w}^T \phi(\mathbf{x_i}) + \epsilon_i, \epsilon_i \sim \mathcal{N}(0, \sigma^2)$. For the design matrix, we use an RBF Kernel, $\phi(\mathbf{x}_i, \mathbf{x}_j) = \exp\left\{-\frac{||\mathbf{x}_i - \mathbf{x}_j||}{l}\right\}$. Results are evaluated

based on the Mean Squared Error (MSE) and the Correlation Coefficient (COR).

Results are presented in Table 10.1. We focus our discussion mostly on the COR, since the MSE is typically very small. There are several interesting observations. Firstly, audio cues appear better for predicting interest in contrast to facial features. This is expected, since according to theory [130], interest is more correlated with arousal, which is the primary dimension for which audio cues are known to perform better [174, 99], while this has also been confirmed by other works on interest recognition (c.f., [227]). Furthermore, it is clear that feature level fusion and classical CCA fusion are not able to out-perform single-cue prediction. In fact, CCA fusion merely manages to achieve equal accuracy to using simply audio cues. COBE, JIVE and LS-CCA$_{\ell 2}$ achieve similar results, while they are outperformed by LS-CCA$_{\ell 1}$. It is clear that RCCA outperforms all compared techniques, by correctly estimating a low-rank subspace where the input modalities are maximally correlated, free of gross noise contaminations, capturing both intra and inter-cue correlations.

Table 10.1: Results for predicting interest from emotion dimensions in the SEMAINE database, using facial trackings (Face), audio cues (Audio), feature-level fusion (F_l), CCA-based fusion (CCA$_f$), Robust CCA fusion (RCCA$_f$) and other compared techniques. Evaluation is based on the Mean Squared Error (MSE) and the correlation coefficient (COR).

	Face	**Audio**	**F$_l$**	**CCA$_f$**	**RCCA$_f$**	**COBE$_f$**	**JIVE$_f$**	**LS-CCA$_{\ell 1,f}$**	**LS-CCA$_{\ell 2,f}$**
MSE	0.033	0.031	0.031	0.031	0.029	0.030	0.03	0.031	0.032
COR	0.432	0.460	0.443	0.458	**0.490**	0.463	0.46	0.48	0.464

10.3.3 Heterogeneous Face Recognition via RCCA

For our experiments, we utilise the CASIA Heterogeneous Face Biometrics database [140], which consists of static face images captured in different (heterogeneous) spectral bands, e.g. visual (VIS) spectrum, the near infrared (NIR) spectrum or measurements of the 3D facial shape (3D). The database contains a total of 100 subject, with 4 VIS and 4 NIR face images per subject, while for 3D faces, 2 images are included per subject for 92 subjects, and 1 per subject for the other 8 subjects. An example of the data is shown in Fig. 10.3. In our experiment, we use a subset of the data for which all VIS, NIR and 3D spectrum images are

available, consisting of 100 subjects and a total of *approx* 600 images. We perform two sets of distinct experiments, where in each we consider two modalities: Firstly, $\mathbf{X}$ =VIS and $\mathbf{Y}$ =3D, and secondly $\mathbf{X}$ =VIS and $\mathbf{Y}$ =NIR, where $\mathbf{X}$ and $\mathbf{Y}$ represent the relevant data matrices. In each experiment, we train using both modalities, inferring projections to the shared space. During testing, only one modality is present; therefore, the shared space is recovered by projecting the queried modality onto the shared space, via the projections inferred during training. Secondly, we utilise the CUHK [268] database. We utilise a portion of the database

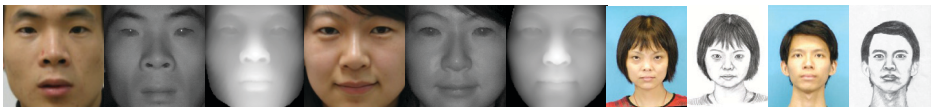


Figure 10.3: Example data included in the CASIA HFB [140] (male and female subject, visual, infra-red and 3d) and CUHK [268] (female and male subject, visual and sketch) databases.

containing 188 subjects, where for each subject a visual image along with a sketch is provided (See Fig. 10.3). We use 100 subjects for training and 88 for testing. Since the sets of training and testing identities are disjoint, we perform correlation-based matching on the testing set in order to match the sketches to the visual images and vice-versa in the projected space learnt during training. Finally, in order to evaluate the compared methods under noisy scenarios, we adopt six noise levels for our experiments, with each level corresponding to the percentage of corrupted images and the percentage of the image which is corrupted. We uniformly select a number of images from the dataset, which are subsequently corrupted by superimposing black patches on a certain percentage of the image area. Results utilising all compared methods are presented in Fig. 10.4. Clearly, the results indicate that RCCA overperforms compared methods in the presence of noise, which is the typical case under real-world scenarios.

10.3.4 Temporal Alignment of Human Walking

In this set of experiments, the performance of the RCTW in alignment of human actions is assessed by conducting experiments on the KTH database [224]. To this end, 25 pairs of sequences consisting of videos performing the same action (walking) were randomly selected. Variations within the pairs appear in clothing, background or view angle. To make the experiment more challenging, we occlude 30% of each frame. In Fig. 10.5b the mean alignment error

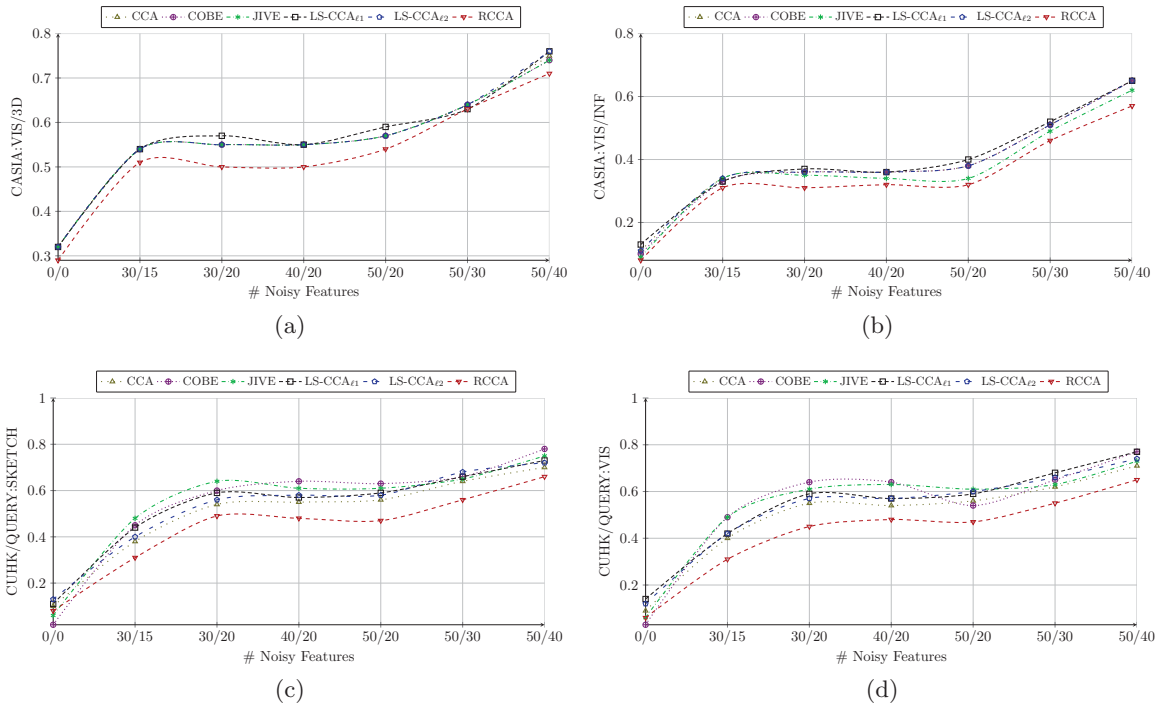
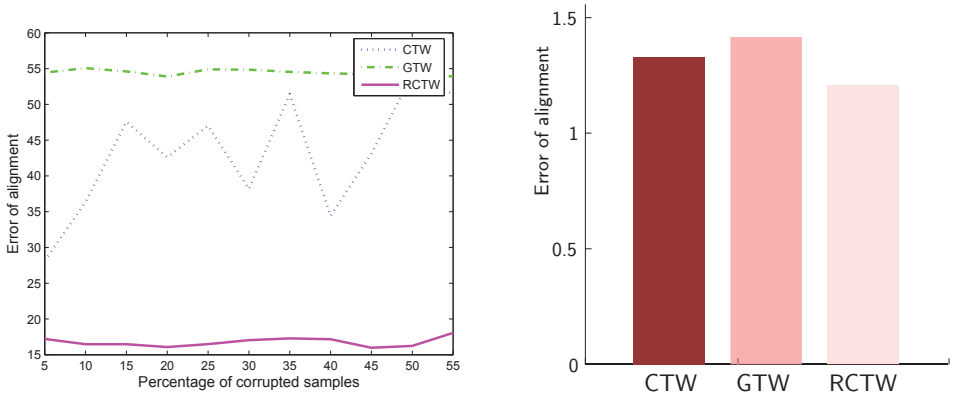


Figure 10.4: Error resulting from compared methods on the CASIA HFB and CUHK databases.

obtained by the CTW, the GTW and the RCTW on corrupted human walking sequences is depicted. Clearly, the RCTW outperforms the CTW and the GTW with respect to alignment error. An illustrative example of aligning occluded human walking sequences with the RCTW is depicted in Fig. 10.6. It can be observed that the occlusions have been removed.

10.3.5 Temporal Action Unit Alignment

The MMI dataset [189] has been employed in order to assess the performance of the RCTW on the temporal alignment of facial expressions. The MMI database [189] consists of more than 300 videos which have been annotated in terms of *action units* (AUs). In particular, each video contains frame-by-frame annotations of each action unit activated covering all temporal phases (i.e., neutral, onset, apex, offset) of each AU. We use a subset of the database with



(a) (b)

Figure 10.5: (a) Mean alignment error obtained by the CTW, the GTW, and the RCTW, as a function of the percentage of corrupted samples on synthetic data sequences. (b) Mean alignment error obtained by the CTW, the GTW and the RCTW on human walking sequences by the KTH.

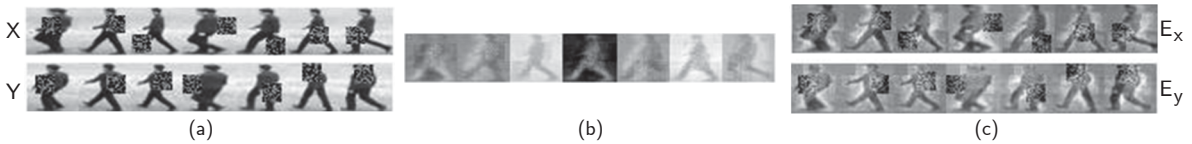


Figure 10.6: Alignment of occluded human walking sequences obtained by the RCTW. (a) The initial occluded walking sequences i.e., $\mathbf{X}$, $\mathbf{Y}$. (b) Aligned sequences onto the error-free latent common space which has been robustly estimated by the RCTW. (c) Magnitude of the recovered gross errors.

approximately 50 pairs of videos of 8 different subjects where action unit 12 is activated.

The experiment proceeds as follows. Firstly, we extract a set of 20 facial points using a person independent tracker presented in [190]. We use 8 2D points (16 dimensional feature vector) which refer to the lower face. Subsequently, we corrupt the facial features with sparse spike noise in order to evaluate the robustness of the compared algorithms. In particular, we draw values from a random normal distribution and add uniformly to 5% of the frames of

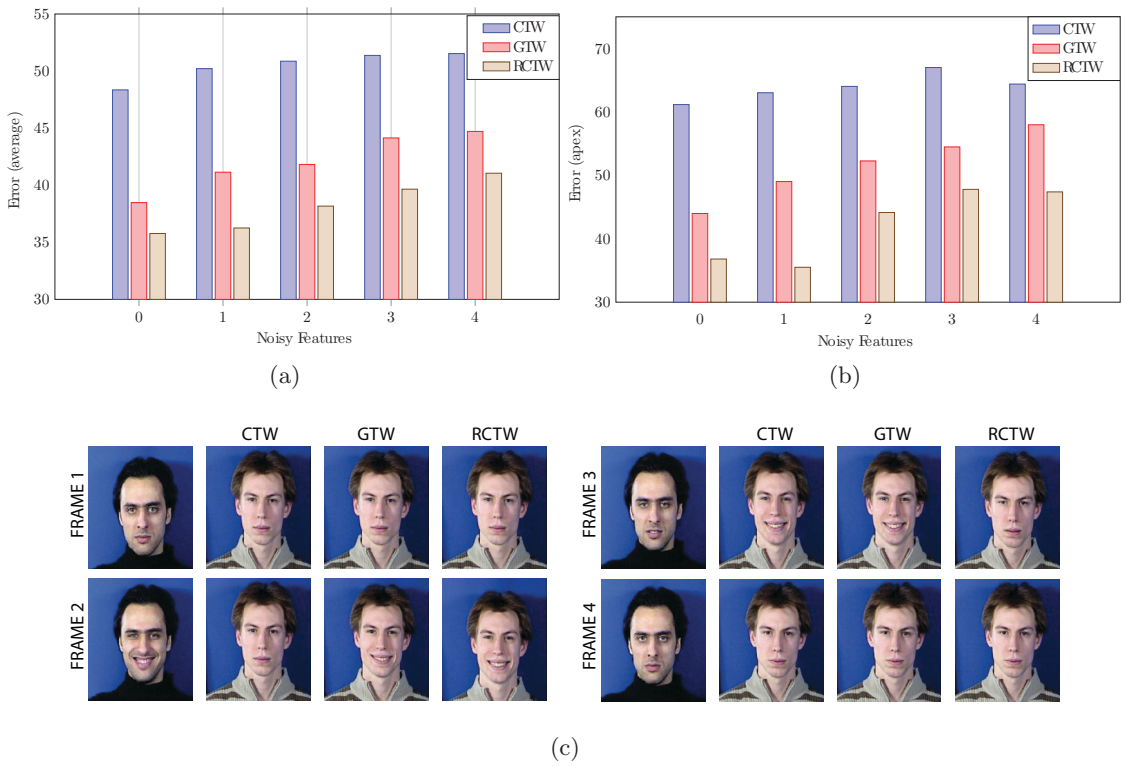


Figure 10.7: Action Unit alignment comparing the RCTW, the CTW, and the GTW. (a) Average error, (b) error for apex phase, (c) example video, where four frames grabbed from the entire duration of both videos are shown. For each frame the first image shows the first video, while the rest of the three show the corresponding (aligned) frame of the second video for each of the methods employed.

each video. This type of noise is common when using detection-based trackers, in which case a point can be misdirected for several frames.

Results are presented in Fig. 10.7. The error we used is the percentage of misaligned frames for each pair of videos, normalised per frame (i.e. divided by the aligned video length). We present results on average (for the entire video, Fig. 10.7a) and results regarding the apex (which is the 'peak' of the expression, Fig. 10.7b). In the presented results, the number of features corrupted by noise increases to 4 out of 8 (which essentially means that 50% of our features are corrupted by noise). It is clear from the results that the RCTW can outperform

both the CTW and the GTW in this scenario, maintaining relatively low error even when heavily increasing the presence of noise.

10.4 Conclusions

In this chapter, by exploiting recent advances on matrix rank minimization we proposed one of the first robust variants of CCA, and the first method which simultaneously discovers a subspace in which two sequences maximally correlate, and at the same time removes possibly gross errors from the data. The proposed method outperforms state-of-the-art techniques in many problems related to fusion and alignment, such as (i) the temporal alignment of action units and human walking sequences in the presence of gross errors, (ii) the robust audio-visual fusion under various noise levels for the detection of interest, as well as (iii) the problem of heterogeneous face recognition.

Chapter 11

A Unified Framework for Probabilistic Component Analysis

Contents

11.1 Introduction

Unification frameworks in machine learning provide valuable material towards the deeper understanding of various methodologies, while also they form a flexible basis upon which further extensions can be easily built. One of the first attempts to unify methodologies was made in [212]. In this seminal work, models such as Factor analysis (FA), Principal Component Analysis (PCA), mixtures of Gaussian clusters (MGC), vector quantization (VQ), Linear Dynamic Systems (LDS), Hidden Markov Models (HMM) and Independent Component Analysis (ICA) were unified as variations of unsupervised learning under a single basic generative model.

Component Analysis (CA)[1] unification frameworks proposed in previous works, such as [52], [2], [123], [30], [58] and [241], provide significant insights on how CA methods such as Principal Component Analysis (PCA), Linear Discriminant Analysis (LDA), Laplacian Eigenmaps and others can be formulated as (i) least squares problems under mild conditions, (ii) graph embedding schemes solved as generalised eigenvalue problems, (iii) as trace optimisation problems with generalised orthogonalities, and (iv) as optimisation problems over manifold spaces. Nevertheless, while some probabilistic equivalents of, e.g. PCA have been developed (c.f., [248] [211]), to this date no unification framework has been proposed for *probabilistic* component analysis.

Motivated by the latter, in this chapter we propose the *first* probabilistic unified framework for component analysis. Based on Markov Random Fields (MRFs), our framework unifies *all* component analysis techniques whose corresponding deterministic problem is solved as a trace optimisation problem without domain constraints for the parameters, such as Principal Component Analysis (PCA), Linear Discriminant Analysis (LDA), Locality Preserving Projections (LPP) and Slow Feature Analysis (SFA). Our framework provides further insight on component analysis methods from a probabilistic perspective. This entails providing probabilistic explanations for the data at hand with explicit variance modelling, as well as reduced complexity compared to the deterministic equivalents. These qualities are even more valuable

[1] Component Analysis has been introduced in Chapter 3.

in case of methods for which no probabilistic equivalent exists in literature so far (such as LPP, a probabilistic equivalent of which is presented in this chapter). Furthermore, our generalised framework provides a straight-forward methodology for producing novel component analysis techniques by imposing specific parametrisations on products of MRFs.

The rest of the chapter is organised as follows. We initially introduce previous work on component analysis, while highlighting the novelties/advantages of methods generated following the proposed framework (Section 11.2). Subsequently, we formulate the joint (complete-data) Probability Density Function (PDF) of a set of observations and latent variables. We show that the Maximum Likelihood (ML) solution of this joint PDF is co-directional to the solutions obtained when solving the deterministic PCA, LDA, LPP and SFA, by changing only the prior distribution of the latent variable (Section 11.3), thus theoretically proving the equivalence of our probabilistic models to the corresponding deterministic. As we show, the prior distribution models the latent dependencies and thus determines the resulting component analysis technique. E.g, when using a fully connected Markov Random Field (MRF) for the latent prior distribution, we derive PCA. When choosing the product of a fully connected MRF and an MRF connected only to within-class data, we derive LDA. LPP is derived by choosing a locally connected MRF, while finally, SFA is produced when the joint prior is a linear Markov-chain. Based on the aforementioned PDF we subsequently propose Expectation Maximization (EM) algorithms for learning the parameters of the model (Section 11.4). Furthermore, we generalize the algorithm to products of arbitrarily-many MRFs with arbitrary parametrization, thus providing an elegant framework for producing novel component analysis techniques. An example is shown in Section 11.6, where we propose a novel, part-based component analysis technique. In Section 11.7, with a set of both synthetic and real data, we demonstrate the usefulness and advantages of this family of probabilistic component analysis methods, which are shown to outperform their deterministic (and probabilistic, given they exist) equivalents, while finally, we conclude the chapter in Section 11.8.

11.2 Prior Art and Novelties

An important contribution of this chapter lies in the proposed unification of probabilistic component techniques, giving rise to the first framework that reduces the construction of probabilistic component analysis models to the design of an appropriate prior, thus defining the latent connectivity.

In Chapter 3, we already introduced component analysis and detailed some common variants such as PCA and CCA. In this section, we describe several component analysis techniques which are utilised in this chapter, such as LDA, LPP and SFA, while also describing the state-of-the-art in probabilistic alternatives. While doing so, we highlight the other novelties and advantages that our proposed framework entails wrt. each alternative formulation. To make the chapter self-sufficient, we also include a reminder of methods related to PCA. Throughout this chapter we consider, without any loss of generality, a zero mean set of F-dimensional observations of length T, $\{\mathbf{x}_1, \ldots, \mathbf{x}_T\}$, represented by a matrix $\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_T]$. All CA methods discover an N-dimensional latent space $\mathbf{Y} = [\mathbf{y}_1, \ldots, \mathbf{y}_T]$ which preserves certain properties of $\mathbf{X}$.

11.2.1 Principal Component Analysis (PCA)

As described in Chapter 3, PCA[2] recovers a set of loadings $\mathbf{W}$, satisfying $\mathbf{Y} = \mathbf{W}^T\mathbf{X}$ where $\mathbf{Y}$ denotes the recovered latent space. By considering $\mathbf{S} = \frac{1}{T}\sum_{i=1}^{T}\mathbf{x}_i\mathbf{x}_i^T$ to be the total scatter matrix and, the optimisation problem is as follows

$$\mathbf{W}_o = \arg\max_{\mathbf{W}} \operatorname{tr}\left[\mathbf{W}^T\mathbf{S}\mathbf{W}\right], \text{ s.t. } \mathbf{W}^T\mathbf{W} = \mathbf{I} \tag{11.1}$$

where $\mathbf{S} = \frac{1}{T}\sum_{i=1}^{T}\mathbf{x}_i\mathbf{x}_i^T$ The optimal N projection basis $\mathbf{W}_o$ are recovered correspond to the N eigenvectors of $\mathbf{S}$ which in turn correspond to the N largest eigenvalues. Probabilistic variants of PCA have been proposed independently in [211] and [248], where the following

[2] We denote deterministic component analysis methods by their initials, e.g. PCA for Principal Component Analysis. Other, existing probabilistic techniques are prefixed with P, e.g. PPCA for Probabilistic PCA. The methods we propose in this chapter are prefixed with ML and EM for Maximum Likelihood and Expectation Maximisation respectively, e.g. EM-PCA.

linear generative model was adopted,

$$\mathbf{x}_i = \mathbf{W}\mathbf{y}_i + \boldsymbol{\epsilon}_i,\ \mathbf{y}_i \sim \mathcal{N}(\mathbf{0}, \mathbf{I}),\ \boldsymbol{\epsilon}_i \sim \mathcal{N}(\mathbf{0}, \sigma^2\mathbf{I}) \tag{11.2}$$

where $\mathbf{W} \in \Re^{F \times N}$ is the loading matrix and $\boldsymbol{\epsilon}_i$ represents noise. When $N < F$, the latent variables are expected to offer a parsimonious explanation of the dependencies between observations.

11.2.2 Linear Discriminant Analysis (LDA)

Let us now further assume that our data $\mathbf{X}$ is further separated into K disjoint classes $\mathcal{C}_1, \ldots, \mathcal{C}_K$ having T_i samples and $T = \sum_{c=1}^{K} |\mathcal{C}_c|$. The Fisher's Linear Discriminant Analysis (LDA) finds a set of projection bases $\mathbf{W}$ s.t. [280]

$$\mathbf{W}_o \ = \arg\min_{\mathbf{W}} \operatorname{tr}\left[\mathbf{W}^T\mathbf{S}_w\mathbf{W}\right],\ \text{s.t. } \mathbf{W}^T\mathbf{S}\mathbf{W} = \mathbf{I} \tag{11.3}$$

where $\mathbf{S}_w = \sum_{c=1}^{K} \sum_{\mathbf{x}_i \in \mathcal{C}_c} (\mathbf{x}_i - \boldsymbol{\mu}_{\mathcal{C}_i})(\mathbf{x}_i - \boldsymbol{\mu}_{\mathcal{C}_i})^T$ and $\boldsymbol{\mu}_{\mathcal{C}_i}$ the mean of class i. The idea is to find a latent space $\mathbf{Y} = \mathbf{W}^T\mathbf{X}$ such that the within-class variance is minimized in a whitened space. The solution is given by the eigenvectors of $\mathbf{S}_w$ that correspond to the $N - K$ eigenvectors (corresponding to the $N - K$ smallest eigenvalues) of the whitened data (i.e. by removing the variance after applying PCA). [3]

Several probabilistic latent variable models which exploit class information have been recently proposed (c.f., [202, 290, 111]). In [202, 290] another two related attempts were made to formulate a PLDA. Considering $\mathbf{x}_i$ to be the i-th sample of the c-th class, the generative model of [202] can be described as:

$$\mathbf{x}_i = \mathbf{F}\mathbf{h}_c + \mathbf{G}\mathbf{w}_{ic} + \boldsymbol{\epsilon}_{ic},\ \mathbf{h}_c, \mathbf{w}_{ic} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}),\ \boldsymbol{\epsilon}_{ic} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}) \tag{11.4}$$

where $\mathbf{h}_c$ represents the class-specific weights and $\mathbf{w}_{ic}$ the weights of each individual sample, with $\mathbf{G}$ and $\mathbf{F}$ denoting the corresponding loadings. Regarding [290], the probabilistic model is as follows:

$$\mathbf{x}_i = \mathbf{F}_c\mathbf{h}_c + \boldsymbol{\epsilon}_{ic},\ \mathbf{h}_c, \mathbf{F}_{ic} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}),\ \boldsymbol{\epsilon}_{ic} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}) \tag{11.5}$$

[3] We adopt this formulation of LDA instead of the equivalent of maximizing the trace of the between-class scatter matrix [22], since this facilitates our following discussion on Probabilistic LDA alternatives.

We note that the two models become equivalent when choosing a common $\mathbf{F}$ (Eq. 11.5) for all classes while also disregarding the matrix $\mathbf{G}$. In this case, the ML solution is given by obtaining the eigenvectors corresponding to the largest eigenvalues of $\mathbf{S}_w$. Hence, the solution is vastly different than the one obtained by deterministic LDA (which keeps the smallest ones, Eq. 11.3), resembling more to the solution of problems which retain the maximum variance. In fact, when learning a different $\mathbf{F}_c$ per class, the model of [290] reduces to applying PPCA per class.

To the best of our knowledge the only probabilistic model where the ML solution is closely related to that of deterministic LDA is [111]. The probabilistic model is defined as follows: $\mathbf{x} \in \mathcal{C}_i,\ \mathbf{x}|\mathbf{y} \sim \mathcal{N}(\mathbf{y}, \mathbf{\Phi}_w),\ \mathbf{y} \sim \mathcal{N}(\mathbf{m}, \mathbf{\Phi}_b),\ \mathbf{V}^T\mathbf{\Phi}_b\mathbf{V} = \mathbf{\Psi}$ and $\mathbf{V}^T\mathbf{\Phi}_w\mathbf{V} = \mathbf{I},\ \mathbf{A} = \mathbf{V}^{-T}$, $\mathbf{\Phi}_w = \mathbf{A}\mathbf{A}^T\ \mathbf{\Phi} = \mathbf{A}\mathbf{\Psi}\mathbf{A}^T$, where the observations are generated as:

$$\mathbf{x}_i = \mathbf{A}\mathbf{u},\ \mathbf{u} \sim \mathcal{N}(\mathbf{V}, \mathbf{I}),\ \mathbf{v} \sim \mathcal{N}(\mathbf{0}, \mathbf{\Psi}). \tag{11.6}$$

The drawback of this model is that it requires all classes to contain the same number of samples [111]. As we will show, we overcome this limitation in our formulation.

11.2.3 Locality Preserving Projections (LPP)

Locality Preserving Projections (LPP) is the linear alternative of Laplacian Eigenmaps [179]. The aim is to obtain a set of projection bases $\mathbf{W}$ and a latent space $\mathbf{Y} = \mathbf{W}^T\mathbf{X}$ which preserves the local neighbourhoods of the original samples. First, let us define a set of weights that represent locality. Common choices for the weights are the heat kernel $u_{ij} = e^{-\frac{||\mathbf{x}_i - \mathbf{x}_j||^2}{\gamma}}$ or a set of constant weights ($u_{ij} = 1$ if the i-th and the j-th vectors are adjacent and $u_{ij} = 0$ otherwise, while $u_{ij} = u_{ji}$). LPP finds a set of projection basis matrix $\mathbf{W}$ by solving the following problem:

$$\begin{aligned} \mathbf{W}_o &= \arg\min_{\mathbf{W}} \textstyle\sum_{i,j=1}^{T} \sum_{n=1}^{N} u_{ij} ||\mathbf{w}_n^T\mathbf{x}_i - \mathbf{w}_n^T\mathbf{x}_j||^2 \\ &= \arg\min_{\mathbf{W}} \operatorname{tr}\left[\mathbf{W}^T\mathbf{X}\mathbf{L}\mathbf{X}^T\mathbf{W}\right] \\ \text{s.t.} \quad & \mathbf{W}^T\mathbf{X}\mathbf{D}\mathbf{X}^T\mathbf{W} = \mathbf{I} \end{aligned} \tag{11.7}$$

where $\mathbf{U} = [u_{ij}]$, $\mathbf{L} = \mathbf{D} - \mathbf{U}$ and $\mathbf{D} = \operatorname{diag}(\mathbf{U}\mathbf{1})$ (where $\operatorname{diag}(\mathbf{a})$ is the diagonal matrix having as main diagonal vector $\mathbf{a}$ and $\mathbf{1}$ is a vector of ones). The objective function with the chosen weights w_{ij} results in a heavy penalty if the neighbouring points $\mathbf{x}_i$ and $\mathbf{x}_j$ are

mapped far apart. Therefore, its minimization ensures that if $\mathbf{x}_i$ and $\mathbf{x}_j$ are near, then the projected features $\mathbf{y}_i = \mathbf{W}^T\mathbf{x}_i$ and $\mathbf{y}_j = \mathbf{W}^T\mathbf{x}_i$ are near as well. To the best of our knowledge no probabilistic models exist for LPPs. In the following (Section 11.3, 11.4), we show how a probabilistic version of LPPs arises by choosing an appropriate prior over the latent space $\mathbf{y}_i$.

11.2.4 Slow Feature Analysis

Now let us consider the case that the columns of $\mathbf{x}_i$ are samples of a time series of length T. The aim of Slow Feature Analysis (SFA) is given T sequential observation vectors $\mathbf{X} = [\mathbf{x}_1 \dots \mathbf{x}_T]$, to find an output signal representation $\mathbf{Y} = [\mathbf{y}_1 \dots \mathbf{y}_T]$ for which the features change slowest over time [273]. By assuming again a linear mapping $\mathbf{Y} = \mathbf{W}^T\mathbf{X}$ for the output representation, SFA minimizes the slowness for these values, defined as the variance of the first derivative of $\mathbf{Y}$. Formally, $\mathbf{W}$ of SFA is computed as

$$\mathbf{W}_o = \arg\min_{\mathbf{W}} \operatorname{tr}\left[\mathbf{W}^T\dot{\mathbf{X}}\dot{\mathbf{X}}\mathbf{W}\right], \text{ s.t. } \mathbf{W}^T\mathbf{S}\mathbf{W} = \mathbf{I}, \tag{11.8}$$

where $\dot{\mathbf{X}}$ is the first derivative matrix (usually computed as the first order difference i.e., $\dot{\mathbf{x}}_j = \mathbf{x}_j - \mathbf{x}_{j-1}$). An ML solution of the SFA was recently proposed in [254]. The idea was to incorporate a Gaussian linear dynamical system prior over the latent space $\mathbf{Y}$. The proposed generative model is

$$\begin{aligned} P(\mathbf{x}_t|\mathbf{W},\mathbf{y}_t,\sigma_x) &= \mathcal{N}(\mathbf{W}^{-1}\mathbf{y}_t,\sigma_x^2\mathbf{I}) \\ P(\mathbf{y}_t|\mathbf{y}_{t-1},\lambda_{1:N},\sigma_{1:N}) &= \textstyle\prod_{n=1}^{N} P(y_{n,t}|y_{n,t-1},\lambda_n,\sigma_n^2) \\ P(y_{n,t}|y_{n,t-1},\lambda_n,\sigma_n^2) &= \mathcal{N}(\lambda_n y_{n,t-1},\sigma_n^2) \\ P(y_{n,1}|\sigma_{n,1}^2) &= \mathcal{N}\left(0,\sigma_{n,1}^2\right). \end{aligned} \tag{11.9}$$

As we will show, SFA is indeed a special case of our general model.

Summarizing, in the following sections we formulate a unified, probabilistic framework for component analysis which: (1) incorporates PCA as a special case, (2) produces a probabilistic LDA which (i) has an ML solution for the loading matrix $\mathbf{W}$ which is co-directional to the deterministic LDA (Eq. 11.3) and (ii) does not make any assumptions regarding the number of samples per class (as in [111]), (3) provides the first, to the best of our knowledge, probabilistic

model which explains LPP, (4) naturally incorporates the recently proposed ML framework of SFA [254] as a special case, (5) provides variance estimates for observations as well as latent dimensions (differentiating our approach from existing probabilistic component analysis techniques (e.g., PPCA, PLDA) by providing more robust estimates, and (6) provides an elegant framework for producing novel component analysis techniques (as we show in Section 11.6).

11.3 A Unified ML Framework for Component Analysis

In this section, we present the proposed Maximum Likelihood (ML) framework for probabilistic component analysis and show how PCA, LDA, LPP and SFA can be generated within this framework, also proving equivalence with known deterministic models. Furthermore, to demonstrate how our framework can be easily used in order to generate novel component analysis techniques, in Section 11.6 we introduce a novel component analysis method, whose ML solution is found merely by following the this section.

Firstly, to ease computations, we assume the generative model for the i-th observation, $\mathbf{x}_i$ to be defined as

$$\mathbf{x}_i = \mathbf{W}^{-1}\mathbf{y}_i + \boldsymbol{\epsilon}_i,\ \boldsymbol{\epsilon}_i \sim N(\mathbf{0}, \sigma_x^2\mathbf{I}). \tag{11.10}$$

In order to fully define the likelihood we need to define a prior distribution on the latent variables $\mathbf{y}$. We will prove that by choosing one of the priors defined below and subsequently taking the ML solution wrt. parameters, we end up generating the aforementioned family of probabilistic component models. The priors, parametrised by $\beta = \{\sigma_{1:N}, \lambda_{1:N}\}$ (illustrated in Fig. 11.1) are:

• An MRF with full connectivity - each latent node $\mathbf{y}_i$ is connected to all other latent nodes $\mathbf{y}_j, j \neq i$.

$$\begin{aligned} P(\mathbf{Y}|\beta) &= \tfrac{1}{Z}\exp\left\{-\tfrac{1}{2}\sum_{n=1}^{N}\sum_{i\in\mathcal{T}}\tfrac{1}{|\mathcal{T}_i|}\sum_{j\in\mathcal{T}_i}\tfrac{1}{\sigma_n^2}(y_{n,i}-\lambda_n y_{n,j})^2\right\} \\ &\approx \tfrac{1}{Z}\exp\left\{-\tfrac{1}{2}\sum_{n=1}^{N}\sum_{i\in\mathcal{T}}\tfrac{1}{|\mathcal{T}|}\sum_{j\in\mathcal{T}}\tfrac{1}{\sigma_n^2}(y_{n,i}-\lambda_n y_{n,j})^2\right\} \\ &= \tfrac{1}{Z}\exp\left\{-\tfrac{1}{2}\left(\mathrm{tr}\left[\boldsymbol{\Lambda}^{(1)}\mathbf{Y}\mathbf{Y}^T\right]+\mathrm{tr}\left[\boldsymbol{\Lambda}^{(2)}\mathbf{Y}\mathbf{M}\mathbf{Y}^T\right]\right)\right\}, \end{aligned} \tag{11.11}$$

where $\mathcal{T} = \{1, \ldots, T\}$, $\mathcal{T}_i = \mathcal{T} \setminus i$, $\mathbf{M} \triangleq -\frac{1}{|\mathcal{T}|}\mathbf{1}\mathbf{1}^T$, $\mathbf{\Lambda}^{(1)} \triangleq \left[\delta_{mn}\frac{\lambda_n^2+1}{\sigma_n^2}\right]$ and $\mathbf{\Lambda}^{(2)} \triangleq \left[\delta_{mn}\frac{2\lambda_n}{\sigma_n^2}\right]$.

• A product of two MRFs. In the first, each latent node $\mathbf{y}_i$ is connected only to other latent nodes in the same class ($\mathbf{y}_j, j \in \tilde{\mathcal{C}}_i$). In the second, each latent node ($\mathbf{y}_i$) is connected to all other latent nodes ($\mathbf{y}_j, j \neq i$).

$$\begin{aligned} P(\mathbf{Y}|\beta) = & \frac{1}{Z}\exp\left\{-\frac{1}{2}\sum_{n=1}^{N}\sum_{i=1}^{T}\frac{1}{|\tilde{\mathcal{C}}_i|}\sum_{j\in\tilde{\mathcal{C}}_i}\frac{\lambda_n}{\sigma_n^2}(y_{n,i}-y_{n,j})^2\right\} \\ & \exp\left\{-\frac{1}{2}\sum_{n=1}^{N}\sum_{i=1}^{T}\frac{1}{T-1}\sum_{j=1}^{T}\frac{(1-\lambda_n)^2}{\sigma_n^2}(y_{n,i}-y_{n,j})^2\right\} \\ = & \frac{1}{Z}\exp\left\{-\frac{1}{2}\left(\mathrm{tr}\left[\mathbf{\Lambda}^{(1)}\mathbf{Y}\mathbf{M}_c\mathbf{Y}^T\right]+\mathrm{tr}\left[\mathbf{\Lambda}^{(2)}\mathbf{Y}\mathbf{M}_t\mathbf{Y}^T\right]\right)\right\}, \end{aligned} \tag{11.12}$$

where $\mathbf{M}_c \triangleq \mathbf{I} - \mathrm{diag}[\mathbf{C}_1, \ldots, \mathbf{C}_C]$, $\mathbf{C}_c \triangleq \frac{1}{N_c}\mathbf{1}_c\mathbf{1}_c^T$, $\mathbf{M}_t \triangleq \mathbf{I} + \mathbf{M}$, $\mathbf{\Lambda}^{(1)} \triangleq \left[\delta_{mn}(\frac{\lambda_n}{\sigma_n^2})\right]$ and $\mathbf{\Lambda}^{(2)} \triangleq \left[\delta_{mn}\frac{(1-\lambda_n)^2}{\sigma_n^2}\right]$, while $\tilde{\mathcal{C}}_i = \{j : \exists\, \mathcal{C}_l \text{ s.t. } \{\mathbf{x}_j, \mathbf{x}_i\} \in \mathcal{C}_l, i \neq j\}$.

• A product of two MRFs. In the first, each latent node $\mathbf{y}_i$ is connected to all other latent nodes that belong in $\mathbf{y}_i$'s neighbourhood. This neighbourhood is symmetrically defined as $\mathcal{N}_i^s = \mathcal{N}_j^s = \{i \in \mathcal{N}_j \cup j \in \mathcal{N}_i\}$. In the second, we only have individual potentials per node.

$$\begin{aligned} P(\mathbf{Y}|\beta) = & \frac{1}{Z}\exp\left(-\frac{1}{2}\sum_{n=1}^{N}\sum_{i=1}^{T}\frac{1}{|\mathcal{N}_i^s|}\sum_{j\in\mathcal{N}_i^s}\frac{\lambda_n}{\sigma_n^2}(y_{n,i}-y_{n,j})^2\right) \\ & \exp\left(-\frac{1}{2}\sum_{n=1}^{N}\sum_{i=1}^{T}\frac{(1-\lambda_n)^2}{\sigma_n^2}y_{n,i}^2\right) \\ = & \frac{1}{Z}\exp\left\{-\frac{1}{2}\left(\mathrm{tr}\left[\mathbf{\Lambda}^{(1)}\mathbf{Y}\tilde{\mathbf{L}}\mathbf{Y}^T\right]+\mathrm{tr}\left[\mathbf{\Lambda}^{(2)}\mathbf{Y}\tilde{\mathbf{D}}\mathbf{Y}^T\right]\right)\right\} \end{aligned} \tag{11.13}$$

where $\tilde{\mathbf{L}}$ and $\tilde{\mathbf{D}}$ are normalised versions of $\mathbf{L}$ and $\mathbf{D}$ as defined in the relevant section for LPPs (Section 11.2.3) i.e. $\tilde{\mathbf{L}} = \mathbf{D}^{-1}\mathbf{L}$ and $\tilde{\mathbf{D}} = \mathbf{I}$, while $\mathbf{\Lambda}^{(1)}$ and $\mathbf{\Lambda}^{(2)}$ are defined as above.

• A linear dynamical system prior over the latent space.

$$\begin{aligned} P(\mathbf{Y}|\beta) & = \frac{1}{Z}\exp\left\{-\sum_{n=1}^{N}\left(\frac{1}{2\sigma_{n,1}^2}y_{n,1}^2+\frac{1}{2\sigma_n^2}\sum_{t=2}^{T}[y_{n,t}-\lambda_n y_{n,t-1}]^2\right)\right\} \\ & \approx \frac{1}{Z}\exp\left\{-\frac{1}{2}\left(\mathrm{tr}\left[\mathbf{\Lambda}^{(1)}\mathbf{Y}\mathbf{K_1}\mathbf{Y}^T\right]+\mathrm{tr}\left[\mathbf{\Lambda}^{(2)}\mathbf{Y}\mathbf{Y}^T\right]\right)\right\} \end{aligned} \tag{11.14}$$

where $\mathbf{K}_1 = \mathbf{P}_1\mathbf{P}_1^T$ and $\mathbf{P}_1$ is a $T \times (T-1)$ matrix with elements $p_{ii} = 1$ and $p_{(i+1)i} = -1$ (the rest are zero). The approximation holds when $T \to \infty$. Again, $\mathbf{\Lambda}^{(1)}$ and $\mathbf{\Lambda}^{(2)}$ are defined as above.

In all cases the partition function Z is defined as $Z = \int P(\mathbf{Y})d\mathbf{Y}$. The motivation behind choosing the above priors over the latent space was given by the influential analysis made in

[103] where the connection between (the deterministic) LPPs, PCA and LDA was explored. A further piece of the puzzle was added by the recent work [254] where the linear dynamical system prior (Eq. 11.14) was used in order to provide a derivation of SFA in a ML framework. By formulating the proper priors for these models we unify these subspace methods in a single probabilistic framework of a linear generative model along with a prior of the form

$$P(\mathbf{Y}) \propto \exp\left\{-\tfrac{1}{2}\left(\mathrm{tr}\left[\mathbf{\Lambda}^{(1)}\mathbf{Y}\mathbf{B}^{(1)}\mathbf{Y}^T\right] + \mathrm{tr}\left[\mathbf{\Lambda}^{(2)}\mathbf{Y}\mathbf{B}^{(2)}\mathbf{Y}^T\right]\right)\right\}. \tag{11.15}$$

The differentiation amongst these models lies in the neighbourhood over which the potentials are defined. In fact, the varying neighbouring system is translated into the matrices $\mathbf{B}^{(1)}$ and $\mathbf{B}^{(2)}$ in the functional form of the potentials, essentially encapsulating the latent covariance connectivity. E.g., for Eq. 11.11, $\mathbf{B}^{(1)} = \mathbf{I}$ and $\mathbf{B}^{(2)} = \mathbf{M}$, for Eq. 11.12, $\mathbf{B}^{(1)} = \mathbf{M}_c$ and $\mathbf{B}^{(2)} = \mathbf{M}_t$, for Eq. 11.13, $\mathbf{B}^{(1)} = \mathbf{L}$ and $\mathbf{B}^{(2)} = \mathbf{D}$ and finally for Eq. 11.14, $\mathbf{B}^{(1)} = \mathbf{K}$ and $\mathbf{B}^{(2)} = \mathbf{I}$ (also see Table 11.2).

In the following we will show that ML estimation using these potentials is equivalent to the deterministic formulations of PCA, LDA and LPP. SFA is a special case for which it was already shown in [254] that a potential of the form of Eq. 11.14 within an ML framework produces a projection with the same direction as Eq. 11.8.

Adopting the linear generative model in Eq. 11.10, the corresponding conditional data (observation) probability is a Gaussian,

$$P(\mathbf{x}_t|\mathbf{y}_t, \mathbf{W}, \sigma_x^2) = \mathcal{N}(\mathbf{W}^{-1}\mathbf{y}_t, \sigma_x^2). \tag{11.16}$$

Having chosen a prior of the form described in Eq. 11.15 we can now derive the likelihood of our model as follows:

$$P(\mathbf{X}|\Psi) = \int \prod_{t=1}^{T} P(\mathbf{x}_t|\mathbf{y}_t, \mathbf{W}, \sigma^2) P(\mathbf{Y}|\sigma_{1:N}^2, \lambda_{1:N}) d\mathbf{Y} \tag{11.17}$$

where the model parameters are defined as $\Psi = \{\sigma_x^2, \mathbf{W}, \sigma_{1:N}^2, \lambda_{1:N}\}$. In the following we will show that by substituting the above priors in Eq. 11.17 and maximising the likelihood we obtain loadings $\mathbf{W}$ which are the same (up to a scale ambiguity) to the deterministic PCA, LDA and LPPs and SFA.

Firstly, by substituting the general prior (Eq. 11.15) in the likelihood (Eq.

$$\begin{aligned} P(\mathbf{X}|\Psi) &= \int \prod_{t=1}^{T} P(\mathbf{x}_t|\mathbf{y}_t, \mathbf{W}, \sigma^2) \frac{1}{Z} \exp \\ &\left\{-\frac{1}{2}\left(\text{tr}\left[\mathbf{\Lambda}^{(1)}\mathbf{Y}\mathbf{B}^{(1)}\mathbf{Y}^T\right] + \text{tr}\left[\mathbf{\Lambda}^{(2)}\mathbf{Y}\mathbf{B}^{(2)}\mathbf{Y}^T\right]\right)\right\} d\mathbf{Y}. \end{aligned} \tag{11.18}$$

In order to obtain a zero-variance limit ML solution, we map $\sigma_x \to 0$

$$\begin{aligned} P(\mathbf{X}|\Psi) &= \int \prod_{t=1}^{T} \delta(\mathbf{x}_t - \mathbf{W}^{-1}\mathbf{y}_t) \frac{1}{Z} \exp \\ &\left\{-\frac{1}{2}\left(\text{tr}\left[\mathbf{\Lambda}^{(1)}\mathbf{Y}\mathbf{B}^{(1)}\mathbf{Y}^T\right] + \text{tr}\left[\mathbf{\Lambda}^{(2)}\mathbf{Y}\mathbf{B}^{(2)}Y^T\right]\right)\right\} d\mathbf{Y} \end{aligned} \tag{11.19}$$

By completing the integrals and taking the logarithms, we obtain the conditional log-likelihood:

$$\begin{aligned} L(\Psi) = \log P(\mathbf{X}|\theta) = &- \log Z + T \log |\mathbf{W}| - \frac{1}{2} \\ &\text{tr}\left[\mathbf{\Lambda}^{(1)}\mathbf{W}\mathbf{X}\mathbf{B}^{(1)}\mathbf{X}^T\mathbf{W}^T + \mathbf{\Lambda}^{(2)}\mathbf{W}\mathbf{X}\mathbf{B}^{(2)}\mathbf{X}^T\mathbf{W}^T\right] \end{aligned} \tag{11.20}$$

where $\log Z$ is a constant term independent of $\mathbf{W}$. By maximising for $\mathbf{W}$ ($\frac{\partial L}{\partial \mathbf{W}} = 0$) we obtain

$$\begin{aligned} T\mathbf{W}^{-T} - \left(\mathbf{\Lambda}^{(1)}\mathbf{W}\mathbf{X}\mathbf{B}^{(1)}\mathbf{X}^T + \mathbf{\Lambda}^{(2)}\mathbf{W}\mathbf{X}\mathbf{B}^{(2)}\mathbf{X}^T\right) &= \mathbf{0} \\ \mathbf{I} = \mathbf{\Lambda}^{(1)}\mathbf{W}\mathbf{X}\mathbf{B}^{(1)}\mathbf{X}^T\mathbf{W}^T + \mathbf{\Lambda}^{(2)}\mathbf{W}\mathbf{X}\mathbf{B}^{(2)}\mathbf{X}^T\mathbf{W}^T. \end{aligned} \tag{11.21}$$

It is easy to prove that since $\mathbf{\Lambda}^{(1)}, \mathbf{\Lambda}^{(2)}$ are diagonal matrices, the $\mathbf{W}$ which satisfies Eq. 11.21 simultaneously diagonalises (up to a scale ambiguity) $\mathbf{X}\mathbf{B}^{(1)}\mathbf{X}^T$ and $\mathbf{X}\mathbf{B}^{(2)}\mathbf{X}^T$. By substituting the $\mathbf{B}$ matrices (as defined in Table 11.2) in Eq. 11.21, we now consider all cases separately:

- **PCA.** By utilising Eq. 11.11, Eq. 11.21 is reformulated as $\mathbf{W}\mathbf{X}\mathbf{X}^T\mathbf{W}^T = \left[\mathbf{\Lambda}^{(1)}\right]^{-1}$ hence $\mathbf{W}$ is given by (up to a scale ambiguity) the eigenvectors of the total scatter matrix $\mathbf{S}$.
- **LDA.** By substituting Eq. 11.12 in Eq. 11.21, we arrive at $\mathbf{\Lambda}^{(1)}\mathbf{W}\mathbf{X}\mathbf{M}\mathbf{X}^T\mathbf{W}^T + \mathbf{\Lambda}^{(2)}\mathbf{W}\mathbf{X}\mathbf{X}^T\mathbf{W}^T = \mathbf{I}$. Thus, $\mathbf{W}$ is given by the directions that simultaneously diagonalise $\mathbf{S}$ and $\mathbf{S}_w$.
- **LPP.** By using Eq. 11.13 then Eq. 11.21 yields $\mathbf{\Lambda}^{(1)}\mathbf{W}\mathbf{X}\tilde{\mathbf{L}}\mathbf{X}^T\mathbf{W}^T + \mathbf{\Lambda}^{(2)}\mathbf{W}\mathbf{X}\tilde{\mathbf{D}}^T\mathbf{X}^T\mathbf{W}^T = \mathbf{I}$, therefore $\mathbf{W}$ is given by the directions that simultaneously diagonalise $\mathbf{X}\mathbf{L}\mathbf{X}^T$ and $\mathbf{X}\mathbf{D}\mathbf{X}^T$.

- **SFA.** Finally, for SFA, by utilising Eq. 11.14, Eq. 11.21 becomes $\mathbf{\Lambda}^{(1)}\mathbf{W}\mathbf{X}\mathbf{K}\mathbf{X}^T\mathbf{W}^T + \mathbf{\Lambda}^{(2)}\mathbf{W}\mathbf{X}\mathbf{X}^T\mathbf{W}^T = \mathbf{I}$, and $\mathbf{W}$ is given by the directions that simultaneously diagonalise $\mathbf{X}\mathbf{K}\mathbf{X}^T$ and $\mathbf{X}\mathbf{X}^T$.

The above shows that the ML solution following our framework is equivalent to the deterministic models of PCA, LDA, LPP and SFA. The direction of $\mathbf{W}$ does not depend of σ_n^2 and λ_n, which can be estimated by optimizing Eq. 11.20 with regards to these parameters. In this work we will provide update rules for σ_n and λ_n using an EM framework (Section 11.4). As can be observed, the ML loading $\mathbf{W}$ does not depend on the exact setting of λ_n, so long as they are all different. If $0 < \lambda_n < 1$, $\forall\ n$, then larger values of λ_n correspond to more expressive (in case of PCA), more discriminant (for LDA), more local (regarding LPP) and slower latents (in case of SFA). This corresponds directly to the ordering of the solutions from PCA, LDA, LPP and SFA. To recover exact equivalence to LDA, LPP, SFA another limit is required that corrects the scales. There are several choices, but a natural one is to let $\sigma_n^2 = 1 - \lambda_n^2$. This choice in case of LDA and SFA fixes the prior covariance of the latent variables to be one ($\mathbf{W}^T\mathbf{X}\mathbf{X}\mathbf{W} = \mathbf{I}$) and it forces $\mathbf{W}^T\mathbf{X}\mathbf{D}\mathbf{X}\mathbf{W} = \mathbf{I}$ in case of LPP. This choice of σ_n has been also discussed in [254] for slow feature analysis. We note that in case of PCA, we should set σ_n to be analogous to the corresponding eigenvalue of the covariance matrix, since otherwise the method will result to a *minor* component analysis.

11.4 A Unified Expectation Maximization for Component Analysis

In the following we propose a unified EM framework for component analysis. This framework can treat all priors with undirected links (such as Eq. 11.11, Eq. 11.12 and Eq. 11.13). The EM of the prior in Eq. 11.14 contains only directed links with no loops, and thus can be solved (without any approximations) similarly to the EM of a linear dynamical system [26]. If we treat the SFA links as undirected, we end up with an autoregressive component analysis (see Section 11.4.1).

In order to perform EM with an MRF prior we adopt the simple and elegant mean field

approximation theory [204, 40, 287], which essentially allows computationally favourable factorizations within an EM framework. Let us consider a generalisation of the priors we defined in Section 11.3 to $\mathcal{M}$ MRFs:

$$P(\mathbf{Y}|\beta) = \prod_{\mu \in \mathcal{M}} \frac{1}{Z^\mu} \exp\{Q^\mu\} \tag{11.22}$$

$$Q^\mu = -\sum_{n=1}^{N} \frac{f_\mu(\lambda_n)}{2\sigma_n^2} \frac{1}{c} \sum_{i \in \omega_i} \frac{1}{c_j^\mu} \sum_{j \in \omega_j^\mu} (y_{n,i} - \phi_\mu(\lambda_n) y_{n,j})^2$$

where c and c_j^μ are normalisation constants, while f_μ and ϕ_μ are functions of λ_n. Without loss of generality and in order to preserve clarity of notation, we assume that $c = 1$, $c_j^\mu = |\omega_j^\mu|$ and $\omega_i^\mu = [1, \ldots, T]$. Furthermore, we now assume the linear model

$$\mathbf{x}_i = \mathbf{W}\mathbf{y}_i + \epsilon_i, \epsilon_i \sim \mathcal{N}(0, \sigma_x^2). \tag{11.23}$$

For clarity, the set of parameters associated with the prior (i.e. energy function) are denoted as $\beta = \{\sigma_{1:N}, \lambda_{1:N}\}$, the parameters related to the observation model $\theta = \{\mathbf{W}, \sigma_x\}$, while the total parameter set is denoted as $\Psi = \{\theta, \beta\}$.

In agreement with [40], we replace the marginal distribution $P(\mathbf{Y}|\beta)$ by the mean-field

$$P(\mathbf{Y}|\beta) \approx \prod_{i=1}^{T} P(\mathbf{y}_i|\mathbf{m}_i^{\mathcal{M}}, \beta^{\mathcal{M}}). \tag{11.24}$$

Since different CA models have different latent connectivities (and thus different MRF configurations), the mean-field influence on each latent point $\mathbf{y}_i$ now depends on the model-specific connectivity via $\mathbf{m}_i^{\mathcal{M}}$, a function of $\mathbb{E}[\mathbf{y}_j]$. After calculating the normalising integral for the priors Eq. 11.11-11.13 and given the mean-field, it can be easily shown that Eq. 11.22 follows a Gaussian distribution,

$$P(\mathbf{y}_i|\mathbf{m}_i^{\mathcal{M}}, \beta) = \mathcal{N}(\mathbf{m}_i^{\mathcal{M}}, \mathbf{\Sigma}^{\mathcal{M}}), \tag{11.25}$$

$$\mathbf{m}_i^{\mathcal{M}} = \sum_{\mu \in \mathcal{M}} \left(\frac{f_\mu(\lambda_n)\phi_\mu(\lambda_n)}{F^M(\lambda_n)} \boldsymbol{\mu}_{\omega_j^\mu} \right) = \sum_{\mu \in \mathcal{M}} \mathbf{\Lambda}^\mu \boldsymbol{\mu}_{\omega_j^\mu} \tag{11.26}$$

$$\mathbf{\Sigma}^{\mathcal{M}} = \left[\delta_{mn} \frac{\sigma_n^2}{F^M(\lambda_n)} \right] \tag{11.27}$$

Table 11.1: MRF configuration for PCA, LDA and LPP, where $\mathcal{T}_i = \{1 \dots T\} \setminus \{i\}$

$\mathcal{M} = \{\alpha, \beta\}$	$F^{\mathcal{M}} = \sum_\mu f_\mu$	f_a	ϕ_α	ω_j^α	f_β	ϕ_β	ω_j^β
PCA (11.11)	1	1	λ_n	$\mathcal{T}_i$			
LDA (11.12)	$\lambda_n + (1-\lambda_n)^2$	λ_n	1	$\tilde{\mathcal{C}}_i$	$(1-\lambda_n)^2$	1	$\mathcal{T}_i$
LPP (11.13)	$\lambda_n + (1-\lambda_n)^2$	λ_n	1	$\mathcal{N}_i^s$	$(1-\lambda_n)^2$	0	$\{1\}$

with $\boldsymbol{\mu}_{\omega_j^\mu} = \frac{1}{|\omega_j^\mu|} \sum_{j \in \omega_j^\mu} \mathbb{E}[\mathbf{y}_{n,j}]$ and $F^M(\lambda_n) = \sum_{\mu \in \mathcal{M}} f_\mu(\lambda_n)$.

Therefore, by simply replacing the parametrisation of the priors we defined in Eq. 11.11 (PCA), 11.12 (LDA) and 11.13 (LPP) (see also Table 11.1) for the mean and variance (Eq. 11.26 and Eq. 11.27), we obtain the posterior distribution for each CA method we propose. The means $\mathbf{m}_i^{\mathcal{M}}$ for PCA, LDA and LPP are obtained as

$$\begin{aligned} \mathbf{m}_i^{(\text{PCA})} &= \boldsymbol{\Lambda} \boldsymbol{\mu}_{-i} \\ \mathbf{m}_i^{(\text{LDA})} &= \boldsymbol{\Lambda}^{(\alpha)} \boldsymbol{\mu}_{-i} + \boldsymbol{\Lambda}^{(\beta)} \boldsymbol{\mu}_{\tilde{\mathcal{C}}_i} \\ \mathbf{m}_i^{(\text{LPP})} &= \boldsymbol{\Lambda}^{(\alpha)} \boldsymbol{\mu}_{\mathcal{N}_i^s} \end{aligned} \tag{11.28}$$

and the variances $\boldsymbol{\Sigma}^{\mathcal{M}}$ as

$$\begin{aligned} \boldsymbol{\Sigma}^{(\text{PCA})} &= \left[\delta_{mn} \sigma_n^2\right] \\ \boldsymbol{\Sigma}^{(\text{LDA})} = \boldsymbol{\Sigma}^{(\text{LPP})} &= \left[\delta_{mn} \left(\frac{\sigma_n^2}{\lambda_n + (1-\lambda_n)^2}\right)\right] \end{aligned} \tag{11.29}$$

where $\boldsymbol{\mu}_{-i} = \frac{1}{T-1} \sum_{j \neq i}^{T} \mathbb{E}^{\mathcal{M}}[\mathbf{y}_j]$ is the mean, $\boldsymbol{\mu}_{\tilde{\mathcal{C}}_i} = \frac{1}{|\tilde{\mathcal{C}}_i|} \sum_{j \in \tilde{\mathcal{C}}_i} \mathbb{E}^{\mathcal{M}}[\mathbf{y}_j]$ the class mean, and $\boldsymbol{\mu}_{\mathcal{N}_i^s} = \frac{1}{|\mathcal{N}_i^s|} \sum_{j \in \mathcal{N}_i^s}^{T} \mathbb{E}^{\mathcal{M}}[\mathbf{y}_j]$ the neighbourhood mean. Furthermore, $\boldsymbol{\Lambda} = [\delta_{mn} \lambda_n]$, $\boldsymbol{\Lambda}^{(\alpha)} = \left[\delta_{mn} \left(\frac{\lambda_n}{\lambda_n + (1-\lambda_n)^2}\right)\right]$ and $\boldsymbol{\Lambda}^{(\beta)} = \left[\delta_{mn} \left(\frac{(1-\lambda_n)^2}{\lambda_n + (1-\lambda_n)^2}\right)\right]$.

In order to complete the expectation step, we infer the first order moments of the latent posterior, defined as

$$P(\mathbf{y}_i | \mathbf{x}_i, \mathbf{m}_i^{\mathcal{M}}, \Psi^{\mathcal{M}}) = \frac{P(\mathbf{x}_i | \mathbf{y}_i, \theta^{\mathcal{M}}) P(\mathbf{y}_i | \mathbf{m}_i^{\mathcal{M}}, \beta^{\mathcal{M}})}{\int_{\mathbf{y}_i} P(\mathbf{x}_i | \mathbf{y}_i, \theta^{\mathcal{M}}) P(\mathbf{y}_i | \mathbf{m}_i^{\mathcal{M}}, \beta^{\mathcal{M}}) d\mathbf{y}_i}. \tag{11.30}$$

Since the posterior is a product of Gaussians[4] , we have

$$P(\mathbf{y}_i | \mathbf{x}_i, \mathbf{m}_i^{\mathcal{M}}, \Psi^{\mathcal{M}}) = \mathcal{N}(\mathbf{y}_i | (\mathbf{W}^T \mathbf{x}_i + \boldsymbol{\Sigma}^{\mathcal{M}^{-1}} \mathbf{m}_i^{\mathcal{M}}) \mathbf{A}, \sigma_x^{\mathcal{M}^2} \mathbf{A}) \tag{11.31}$$

[4]The result can be easily obtained by completing the square for $\mathbf{y}_i$.

with $\mathbf{A} = (\mathbf{W}^T\mathbf{W} + (\hat{\mathbf{\Sigma}}^{\mathcal{M}})^{-1})^{-1}$ and $\hat{\mathbf{\Sigma}}^{\mathcal{M}} = \left[\delta_{mn}(\Sigma^{\mathcal{M}}_{mn}/\sigma_x^{\mathcal{M}^2})\right]$. Therefore $\mathbb{E}^{\mathcal{M}}[\mathbf{y}_i]$ is equal to the mean, and $\mathbb{E}^{\mathcal{M}}[\mathbf{y}_i\mathbf{y}_i^T] = \sigma_x^{\mathcal{M}^2}\mathbf{A} + \mathbb{E}[\mathbf{y}_i]\mathbb{E}[\mathbf{y}_i]^T$.

Having recovered the first order moments, we move on to the maximisation step. In order to maximize the marginal log-likelihood, $\log P(\mathbf{X}|\Psi^{\mathcal{M}})$ we adopt the usual EM bound [212], $\int_{\mathbf{Y}} P(\mathbf{Y}|\mathbf{X}, \Psi^{\mathcal{M}}) \log P(\mathbf{X}, \mathbf{Y}) d\mathbf{Y}$. By adopting the approximation proposed in [40], the complete-data likelihood is factorised as

$$P(\mathbf{Y}, \mathbf{X}|\Psi^{\mathcal{M}}) \approx \prod_{i=1}^{T} P(\mathbf{x}_i|\mathbf{y}_i, \theta^{\mathcal{M}}) P(\mathbf{y}_i|\mathbf{m}_i^{\mathcal{M}}, \beta^{\mathcal{M}}). \tag{11.32}$$

Therefore, the maximisation term (EM bound) becomes

$$\sum_{i=1}^{T} \int_{\mathbf{y}_i} P(\mathbf{y}_i|\mathbf{x}_i, \mathbf{m}_i^{\mathcal{M}}, \Psi^{\mathcal{M}}) \log P(\mathbf{x}_i, \mathbf{y}_i|\Psi^{\mathcal{M}}) d\mathbf{y}_i. \tag{11.33}$$

As can be seen the likelihood can be separated due to the logarithm for estimating $\theta^{\mathcal{M}} = \{\mathbf{W}^{\mathcal{M}}, \sigma_x^{\mathcal{M}}\}$ and $\beta = \{\sigma_{1:N}^{\mathcal{M}}, \lambda_{1:N}^{\mathcal{M}}\}$ as follows:

$$\theta^{\mathcal{M}} = \arg\max \left\{ \sum_{i=1}^{T} \int_{\mathbf{y}_i} P(\mathbf{y}_i|\mathbf{x}_i, \mathbf{m}_i^{\mathcal{M}}, \Psi^{\mathcal{M}}) \log P(\mathbf{x}_i|\mathbf{y}_i, \theta^{\mathcal{M}}) d\mathbf{y}_i \right\}. \tag{11.34}$$

$$\beta^{\mathcal{M}} = \arg\max \left\{ \sum_{i=1}^{T} \int_{\mathbf{y}_i} P(\mathbf{y}_i|\mathbf{x}_i, \mathbf{m}_i^{\mathcal{M}}, \Psi^{\mathcal{M}}) \log P(\mathbf{y}_i|\mathbf{m}_i^{\mathcal{M}}, \beta^{\mathcal{M}}) d\mathbf{y}_i \right\}. \tag{11.35}$$

Subsequently, we maximise the log-likelihoods wrt. the parameters, recovering the update equations (as detailed in the appendix). For θ, by maximising Eq. 11.34, we obtain

$$\mathbf{W}^{\mathcal{M}} = \left(\sum_{i=1}^{T} \mathbf{x}_i \mathbb{E}^{\mathcal{M}}[\mathbf{y}_i]^T \right) \left(\sum_{i=1}^{T} \mathbb{E}^{\mathcal{M}}[\mathbf{y}_i\mathbf{y}_i^T] \right)^{-1} \tag{11.36}$$

$$\begin{aligned} \sigma_x^{\mathcal{M}^2} &= \tfrac{1}{FT} \sum_{i=1}^{T} \{||\mathbf{x}_i||^2 - 2\mathbb{E}^{\mathcal{M}}[\mathbf{y}_i]^T (\mathbf{W}^{\mathcal{M}})^T \mathbf{x}_i \\ &\quad + \mathrm{Tr}[\mathbb{E}^{\mathcal{M}}[\mathbf{y}_i\mathbf{y}_i^T](\mathbf{W}^{\mathcal{M}})^T \mathbf{W}^{\mathcal{M}}]\}. \end{aligned} \tag{11.37}$$

Similarly, by maximising Eq. 11.35 for β, we obtain:

$$\sigma_n^{\mathcal{M}^2} = \frac{F^{\mathcal{M}}(\lambda_n)}{T} \sum_{i=1}^{T} (\mathbb{E}^{\mathcal{M}}[y_{n,i}^2] - 2\mathbb{E}^{\mathcal{M}}[y_{n,i}] m_{n,i}^{\mathcal{M}} + m_{n,i}^{\mathcal{M}^2}) \tag{11.38}$$

where, as defined in Eq. 11.27, for PCA $F^{\mathcal{M}}(\lambda_n) = 1$, and for LDA and LPP $F^{\mathcal{M}}(\lambda_n) = \lambda_n + (1-\lambda_n)^2$. For λ_n we choose the updates as described in Section 11.3. In what follows, we discuss some further points wrt. the proposed EM framework.

11.4.1 Discussion

Comparison to other probabilistic variants of PCA. It is clear that regarding the proposed EM-PCA, the updates for $\theta = \{\mathbf{W}, \sigma_x^2\}$ as well as the distribution of the latent variable $\mathbf{y}_i$ are the same with previously proposed probabilistic approaches [211, 248]. The only variation is the mean of $\mathbf{y}_i$, which in our case is shifted by the mean field, $\hat{\mathbf{\Sigma}}^{(\text{PCA})^{-1}}\mathbf{m}_i^{(\text{PCA})}$, while in addition, our method models per-dimension variance (σ_n), deeming the framework suitable for scenarios where the noise varies amongst dimensions. Note that in order to fully identify with the PPCA proposed in [248], we can set $\lambda_n = 0$, and $\sigma_n = 1$. In case we set just $\lambda_n = 0$ and $\sigma_n \neq 1$ we attain a Factor Analysis variant.

EM for SFA. The SFA prior in Eq. 11.14 allows for two interpretations of the SFA graphical model: both as an undirected MRF and directed graphical model (Dynamic Bayesian Network, DBN). Based on the undirected MRF interpretation, SFA would trivially fit into the EM framework described in this section, where $\mathbf{m}_i^{(\text{SFA})} = \mathbf{\Lambda}\mathbb{E}[\mathbf{y}_{i-1}]$ and $\mathbf{\Sigma}^{(\text{SFA})} = \mathbf{\Sigma}^{(\text{PCA})}$ (Eq. 11.28 and 11.29). In fact, this undirected reading of SFA can lead to an autoregressive [215] SFA model, able to model bi-directional dependencies over the latent variables, which can be easily extended to higher orders. When considering the SFA prior as a directed Markov chain, one can resort to exact inference techniques applied on DBNs. In fact, the EM for SFA can be reduced to solving a standard Linear Dynamic System [26] (LDS). The observed distribution follows Eq. 11.23, while the latent space is generated as $P(\mathbf{y}_t|\mathbf{y}_{t-1}) \sim \mathcal{N}(\mathbf{\Lambda}\mathbf{y}_{t-1}, \Gamma)$, where $\Gamma_{mn} = \delta_{mn}\sigma_n^2$ with the constraint that $\mathbf{\Lambda}$ is diagonal and that $\sigma_n^2 = 1 - \lambda_n^2$. By applying smoothing (e.g., Rauch-Tung-Striebel) we obtain $\mathbb{E}[\mathbf{y}_t], \text{Var}[\mathbf{y}_t\mathbf{y}_t^T]$ and $\text{Var}[\mathbf{y}_t\mathbf{y}_{t-1}^T]$. The updates for $\mathbf{W}$ and σ_x^2 are the same as Eq. 11.36 and Eq. 11.37. The updates for $\mathbf{\Lambda} = [\delta_{mn}\lambda_n]$ are derived similarly to LDS [26], while enforcing $\sigma_n^2 = 1 - \lambda_n^2$ as discussed in Section 11.3

Complexity. The EM algorithm for our models is an iterative procedure for recovering the latent space which preserves the characteristics enforced by the selected latent neighbourhood. Our analysis is similar to PCCA [211, 248]. For $N << T, F$ the complexity at each iteration is bounded by $O(TNF)$, unlike deterministic models which is $O(T^3)$. This is due to the covariance appearing only in trace operations, and is of high value for our proposed EM based models, especially in case of EM-LPP where no probabilistic equivalent exists.

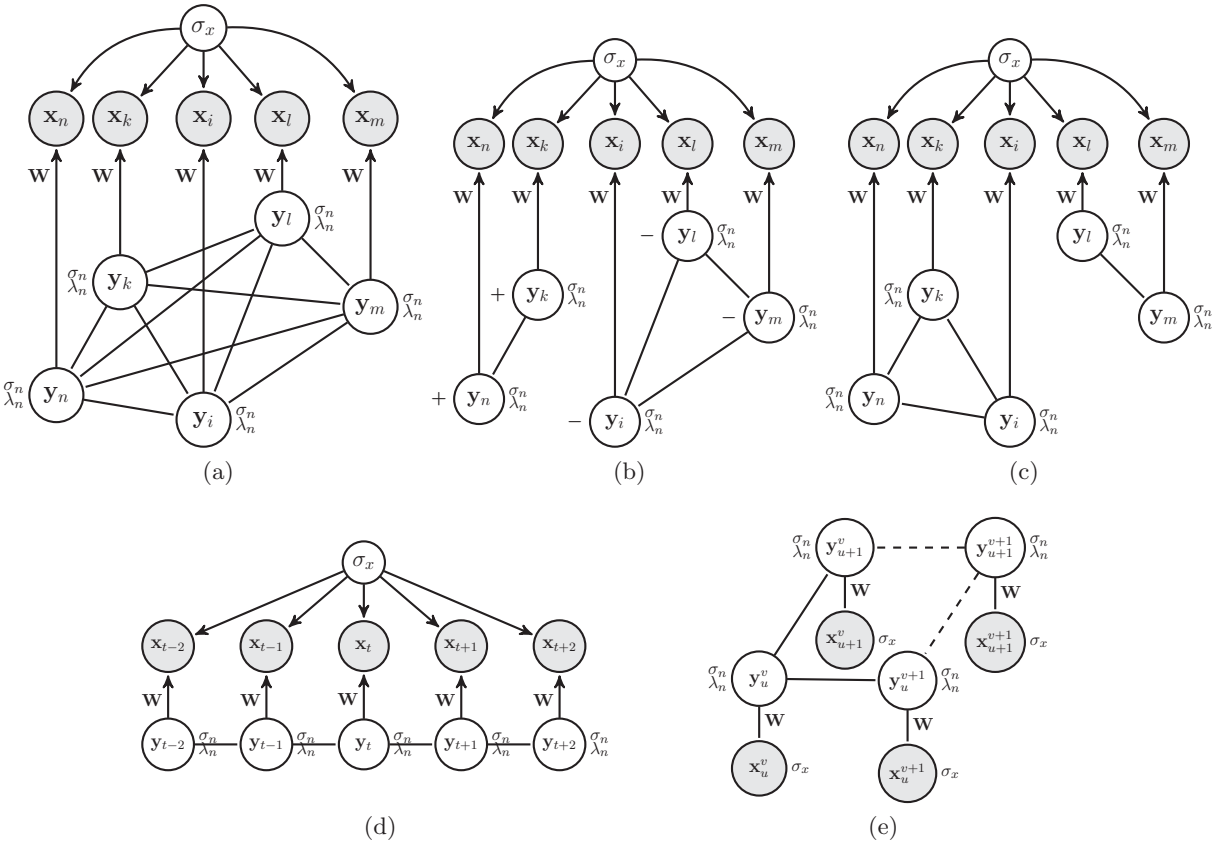


Figure 11.1: MRF connectivies utilised for deriving PCA, LDA, LPP, SFA and Spatial Structure-aware Component Analysis under our unifying framework. (a) Fully connected MRF (for PCA), (b) within-class connected MRF (LDA, along with (a)), (c) locally connected MRF with individual potentials for LPP, (d) A linear chain leading to Autoregressive SFA (when directed leading to SFA), (e) Spatial Structure-Aware method (Section 11.6).

Table 11.2: Matrices $\mathbf{B}^{(1)}$ and $\mathbf{B}^{(2)}$ which determine latent connectivity and in conclusion, the derived component analysis model.

	PCA	LDA	LPP	SFA	SLDA	MFA	NPDA
$\mathbf{B}^{(1)} =$	$\mathbf{I}$	$\mathbf{I} - \text{diag}[\mathbf{C}_1, \ldots, \mathbf{C}_C]$ $\mathbf{C}_c \triangleq \frac{1}{N_c}\mathbf{1}_c\mathbf{1}_c^T$	$\mathbf{L} = \mathbf{D} - \mathbf{W}$	$\mathbf{K} = \mathbf{P}\mathbf{P}^T$	$\mathbf{I} - \text{diag}(\mathbf{C}_1, \ldots, \mathbf{C}_r)$ $\mathbf{C}_r = \frac{1}{\lvert\mathcal{C}^r\rvert}$	$\mathbf{L}_{\tilde{\mathcal{C}}_i^{\mathcal{N}_i^s}}$	$\mathbf{L}_{\tilde{\mathcal{C}}_i^{\mathcal{N}_i^s}}$
$\mathbf{B}^{(2)} =$	$-\frac{1}{T}\mathbf{1}\mathbf{1}^T$	$\mathbf{I} - \mathbf{1}\mathbf{1}^T$	$\mathbf{D} = diag(\mathbf{W}\mathbf{1})$	$\mathbf{I}$	$\mathbf{I} - \frac{1}{T}\mathbf{1}_T\mathbf{1}_T^T$	$\mathbf{L}_{\tilde{\mathcal{C}}_{j\neq i}^{\mathcal{N}_i^s}}$	$\mathbf{L}_{\tilde{\mathcal{C}}_{j\neq i}^{\mathcal{N}_i^s}, w_{i,j}^{\text{NLDA}}}$

Table 11.3: The mean $\mathbf{m}_i^{\mathcal{M}}$ of the posterior latent distribution of each component analysis technique, which along with the covariance (which is the same for all methods except PCA) defines the model at hand.

PCA	LDA	LPP	SLDA	MFA	NPDA
$\mathbf{m}_i^{\mathcal{M}} \quad \mathbf{\Lambda}\boldsymbol{\mu}_{-i}$	$\mathbf{\Lambda}^{(\alpha)}\boldsymbol{\mu}_{-i} + \mathbf{\Lambda}^{(\beta)}\boldsymbol{\mu}_{\tilde{\mathcal{C}}_i}$	$\mathbf{\Lambda}^{(\alpha)}\boldsymbol{\mu}_{\mathcal{N}_i^s}$	$\mathbf{\Lambda}^{(\alpha)}\boldsymbol{\mu}_{-i} + \mathbf{\Lambda}^{(\beta)}\boldsymbol{\mu}_{\mathcal{C}_i^r}$	$\mathbf{\Lambda}^{(\alpha)}\boldsymbol{\mu}_{\tilde{\mathcal{C}}_i^{\mathcal{N}_i}} + \mathbf{\Lambda}^{(\beta)}\boldsymbol{\mu}_{\tilde{\mathcal{C}}_{j\neq i}^{\mathcal{N}_i}}$	$\mathbf{\Lambda}^{(\alpha)}\boldsymbol{\mu}_{\tilde{\mathcal{C}}_i^{\mathcal{N}_i}} + \mathbf{\Lambda}^{(\beta)}\boldsymbol{\mu}_{\tilde{\mathcal{C}}_{j\neq i}^{w,\mathcal{N}_i}}$

Mixtures. The family of models presented in this chapter can be easily extended to handle mixtures of component analysers. This is extremely important in many cases, where more than one Gaussians are required to fit the data. Particularly in our case, we can have mixtures of different component analysis methods. The derivation follows [26] and is detailed in the appendix of this thesis.

Probabilistic LDA Classification. We can exploit the probabilistic nature of the proposed EM-LDA in order to probabilistically infer the most probable class assignment for unseen data. Instead of using the inferred projection, we can essentially utilise the log-likelihood of the model. In more detail, we can estimate the marginal log-likelihood for each test point $\mathbf{x}^*$ being assigned to each class c:

$$\arg_c \max \left\{ \log P(\mathbf{x}^* | \mathbf{m}^{\mathcal{M}_c}, \Psi^{\mathcal{M}}) \right\} \tag{11.39}$$

where by adopting the usual EM bound (as shown in Eq. 11.33), this boils down to

$$\arg_c \max \int_{\mathbf{y}_i^*} P(\mathbf{y}_i^* | \mathbf{x}_i^*, \mathbf{m}^{\mathcal{M}_c}, \Psi^{\mathcal{M}}) \log P(\mathbf{x}_i^*, \mathbf{y}_i^* | \Psi^{\mathcal{M}}) d\mathbf{y}_i^* \tag{11.40}$$

where $P(\mathbf{y}_i^* | \mathbf{x}_i^*, \mathbf{m}^{\mathcal{M}}, \Psi^{\mathcal{M}})$ is estimated as in Eq. 11.31, by utilising the inferred model parameters ($\Psi^{\mathcal{M}}$) along with the class model. Note that since the posterior mean given $\mathbf{x}_i$ depends

on all *other* observations excluding i (Eq. 11.28), we only need to store the class mean estimated as a weighted average of all training data and all training data in class c, as

$$\mathbf{m}^{\mathcal{M}_c} = \mathbf{\Lambda}^{(\alpha)} \frac{1}{T} \sum_{j=1}^{T} \mathbb{E}^{\mathcal{M}}[\mathbf{y}_j] + \mathbf{\Lambda}^{(\beta)} \frac{1}{|\mathcal{C}_c|} \sum_{j \in \mathcal{C}_c} \mathbb{E}^{\mathcal{M}}[\mathbf{y}_j] \tag{11.41}$$

This is in contrast to traditional methods where all the (projected) training data have to be kept. Furthermore, during evaluation, we only need to estimate the likelihood of each test datum's assignment to each class ($\mathcal{O}(|C|)$), rather than compare each test datum to the entire training set ($\mathcal{O}(T)$).

11.5 Variants of LDA / Supervised LPP

In order to demonstrate the flexibility of the proposed framework, in this section we discuss several variants of LDA such as Subclass Discriminant Analysis (SDA) [300], Marginal Fisher Analysis (MFA) [279] and Nonparametric Discriminant Analysis (NPDA) [141], and show how they can be easily incorporated into the proposed framework. We note that MFA and NPDA can also be considered as variants of LPP since (i) the locality is preserved (by accounting for nearest neighbours of each point) and (ii) the class information is used to impose further constraints on locality (i.e. supervision).

Firstly, Subclass Discriminant Analysis determines the number of subclasses in each class via clustering, essentially estimating the optimal number of Gaussians per class. This results in an LDA model where the underlying distribution of each class is a mixture of Gaussians. In our framework, assuming we discover r subclasses for each class, our LDA prior (Eq. 11.12) is reformulated as follows

$$\begin{aligned} P(\mathbf{Y}|\beta) = \; & \tfrac{1}{Z} \exp\left\{ -\tfrac{1}{2} \textstyle\sum_{n=1}^{N} \sum_{i=1}^{T} \frac{1}{|\tilde{\mathcal{C}}_i^r|} \sum_{j \in \tilde{\mathcal{C}}_i^r} \frac{\lambda_n}{\sigma_n^2} (y_{n,i} - y_{n,j})^2 \right\} \\ & \exp\left\{ -\tfrac{1}{2} \textstyle\sum_{n=1}^{N} \frac{1}{T} \sum_{i=1}^{T} \frac{1}{T-1} \sum_{j=1}^{T} \frac{(1-\lambda_n)^2}{\sigma_n^2} (y_{n,i} - y_{n,j})^2 \right\} \end{aligned} \tag{11.42}$$

where $\mathcal{C}_l^r$ denotes the set of data points $\mathbf{x}_j$ belonging to subclass l, while we define $\tilde{\mathcal{C}}_i^r = \{j : \exists\, \mathcal{C}_l^r \text{ s.t. } \{\mathbf{x}_j, \mathbf{x}_i\} \in \mathcal{C}_l^r, i \neq j\}$. For the ML solution, we replace the matrices $\mathbf{B}^{(1)}$ and $\mathbf{B}^{(2)}$ (see Eq. 11.15) to $\mathbf{B}^{(1)} = \mathbf{I} - diag(\mathbf{C}_1, \ldots, \mathbf{C}_r)$ where $\mathbf{C}_r = \frac{1}{|\mathcal{C}^r|} \mathbf{1}_r \mathbf{1}_r^T$ and $\mathbf{B}^{(2)} = \mathbf{I} - \frac{1}{T} \mathbf{1}_T \mathbf{1}_T^T$. For EM, $\mathbf{m}_i^{(\text{SLDA})} = \mathbf{\Lambda}^{(\alpha)} \boldsymbol{\mu}_{-i} + \mathbf{\Lambda}^{(\beta)} \boldsymbol{\mu}_{\mathcal{C}_i^r}$ where now, $\boldsymbol{\mu}_{\tilde{\mathcal{C}}_i} = \frac{1}{|\tilde{\mathcal{C}}_i^r|} \sum_{j \in \tilde{\mathcal{C}}_i^r} \mathbb{E}[\mathbf{y}_j]$ and $\Sigma^{\text{SLDA}} = \Sigma^{\text{LDA}}$.

Marginal Fischer Analysis (MFA) imposes a local structure on discriminant analysis techniques by using local neighbourhoods in the definition of the scatter within-class and between-class scatter matrices. Following MFA, the within-class matrix is measured as the sum of distances between each sample $\mathbf{x}_i$ and its nearest neighbours within the same class. We denote the latter set as $\tilde{\mathcal{C}}_i^{\mathcal{N}_i^s}$, where $j \in \tilde{\mathcal{C}}_i^{\mathcal{N}_i^s}$ iff $\{j \in \mathcal{N}_i^s \cap j \in \tilde{\mathcal{C}}_i\}$. The between-class matrix penalises the nearest neighbours $\mathbf{x}_j$ of each sample $\mathbf{x}_i$ belonging though to different classes. We denote the set of (indices) neighbours of $\mathbf{x}_i$ belonging to a different class than i as $\tilde{\mathcal{C}}_{j\neq i}^{\mathcal{N}_i^s}$, where $j \in \tilde{\mathcal{C}}_{j\neq i}^{\mathcal{N}_i^s}$ iff $\{j \in \mathcal{N}_i^s \cap j \notin \tilde{\mathcal{C}}_i\}$. In this case, we can straight-forwardly reformulate our LDA prior (Eq. 11.12) as follows

$$\begin{aligned} P(\mathbf{Y}|\beta) \quad &= \tfrac{1}{Z} \exp\left\{ -\tfrac{1}{2} \sum_{n=1}^{N} \sum_{i=1}^{T} \tfrac{1}{|\tilde{\mathcal{C}}_i^{\mathcal{N}_i^s}|} \sum_{j \in \tilde{\mathcal{C}}_i^{\mathcal{N}_i^s}} \tfrac{\lambda_n}{\sigma_n^2} (y_{n,i} - y_{n,j})^2 \right\} \\ &\exp\left\{ -\tfrac{1}{2} \sum_{n=1}^{N} \sum_{i=1}^{T} \tfrac{1}{T-1} \sum_{j \in \tilde{\mathcal{C}}_{j\neq i}^{\mathcal{N}_i^s}}^{T} \tfrac{(1-\lambda_n)^2}{\sigma_n^2} (y_{n,i} - y_{n,j})^2 \right\} \end{aligned} \tag{11.43}$$

Regarding the ML solution, given our general model (Eq. 11.15), for MFA we have $\mathbf{B}^{(1)} = \mathbf{L}_{\tilde{\mathcal{C}}_i^{\mathcal{N}_i^s}}$ and $\mathbf{B}^{(2)} = \mathbf{L}_{\tilde{\mathcal{C}}_{j\neq i}^{\mathcal{N}_i^s}}$, where $\mathbf{L}_{\tilde{\mathcal{C}}_i^{\mathcal{N}_i^s}}$ denotes the Laplacian built on the neighbourhood defined by $\tilde{\mathcal{C}}_i^{\mathcal{N}_i^s}$ and $\mathbf{L}_{\tilde{\mathcal{C}}_{j\neq i}^{\mathcal{N}_i^s}}$ the Laplacian built on the neighbourhood defined by $\tilde{\mathcal{C}}_{j\neq i}^{\mathcal{N}_i^s}$. The ML solution diagonalises both $\mathbf{X}\mathbf{L}_{\tilde{\mathcal{C}}_i^{\mathcal{N}_i^s}}\mathbf{X}$ and $\mathbf{X}\mathbf{L}_{\tilde{\mathcal{C}}_{j\neq i}^{\mathcal{N}_i^s}}\mathbf{X}$ (see Eq. 11.21). Regarding our EM framework, $\mathbf{m}_i^{(\text{MFA})} = \mathbf{\Lambda}^{(\alpha)}\boldsymbol{\mu}_{\tilde{\mathcal{C}}_i^{\mathcal{N}_i}} + \mathbf{\Lambda}^{(\beta)}\boldsymbol{\mu}_{\tilde{\mathcal{C}}_{j\neq i}^{\mathcal{N}_i}}$ where now, $\boldsymbol{\mu}_{\tilde{\mathcal{C}}_i^{\mathcal{N}_i}} = \frac{1}{|\tilde{\mathcal{C}}_i^{\mathcal{N}_i}|} \sum_{j \in \tilde{\mathcal{C}}_i^{\mathcal{N}_i}} \mathbb{E}[\mathbf{y}_j]$, $\boldsymbol{\mu}_{\tilde{\mathcal{C}}_{j\neq i}^{\mathcal{N}_i}} = \frac{1}{|\tilde{\mathcal{C}}_{j\neq i}^{\mathcal{N}_i}|} \sum_{j \in \tilde{\mathcal{C}}_{j\neq i}^{\mathcal{N}_i}} \mathbb{E}[\mathbf{y}_j]$, and $\Sigma^{\text{MFA}} = \Sigma^{\text{LDA}}$.

The formulation for Nonparametric LDA (NLDA) [141] is quite similar to the above discussion on MFA, with an introduction of an extra weighting for the scatter-between matrix $w_{i,j}^{\text{NLDA}}$, which depends on the classes of $\mathbf{y}_i$ and $\mathbf{y}_j$ and the nearest neighbours of $\mathbf{y}_i$ [141]. The weight is essentially the ratio of the minimum distance of $\mathbf{y}_i$ to the nearest neighbours (of the same or different class) to the sum of those distances. Within our framework, this method is very similar to the MFA formulation. For the ML solution, we have $\mathbf{B}^{(1)} = \mathbf{L}_{\tilde{\mathcal{C}}_i^{\mathcal{N}_i^s}}$ and $\mathbf{B}^{(2)} = \mathbf{L}_{\tilde{\mathcal{C}}_{j\neq i}^{\mathcal{N}_i^s}, w_{i,j}^{\text{NLDA}}}$, where $\mathbf{L}_{\tilde{\mathcal{C}}_{j\neq i}^{\mathcal{N}_i^s}, w_{i,j}^{\text{NLDA}}}$ is the Laplacian built on $\tilde{\mathcal{C}}_{j\neq i}^{\mathcal{N}_i^s}, w_{i,j}^{\text{NLDA}}$, where each entry has been weighted by $w_{i,j}^{\text{NLDA}}$. Regarding EM, this only changes the corresponding mean

of $\mathbf{m}_i^{(\text{MFA})}$, $\boldsymbol{\mu}_{\tilde{\mathcal{C}}_{j\neq i}^{\mathcal{N}_i}}$ to $\boldsymbol{\mu}_{\tilde{\mathcal{C}}_{j\neq i}^{w,\mathcal{N}_i}} = \frac{1}{|\tilde{\mathcal{C}}_{j\neq i}^{\mathcal{N}_i}|} \sum_{j\in\tilde{\mathcal{C}}_{j\neq i}^{\mathcal{N}_i}} w_{i,j}^{\text{NLDA}} \mathbb{E}[\mathbf{y}_j]$.

11.6 Spatial Structure-Aware Dimensionality Reduction

In order to demonstrate the utilisation of the proposed framework in terms of generating new component analysis techniques, we propose a novel 2D component analysis method, which we coin as *Spatial Structure-aware Component Analysis* (SACA). The model consists of imposing an MRF over a grid (i.e., an image, or segments/parts of a structure), thus preserving spatial relationships in the discovered latent space. We now assume our observations (which can be image descriptors) $\mathbf{X}$ are taken over a discrete grid $uv \in \mathcal{I} = [1 \dots T_1] \times [1 \dots T_2]$, with $T = T_1 \times T_2$. We follow the linear generative model described in Eq. 11.10 and define an MRF prior over the latent space as

$$\begin{aligned} P(\mathbf{Y}|\beta) =& \frac{1}{Z}\exp\left\{-\frac{1}{2}\sum_{n=1}^{N}\sum_{uv}\left\{\left\{\frac{\lambda_n}{2{\sigma_n}^2}(y_{n,uv} - y_{n,(u+1)v})^2\right\}\right.\right. \\ &\left.\left. + \frac{\lambda_n}{2{\sigma_n}^2}(y_{n,uv} - y_{n,u(v+1)})^2 + \frac{(1-\lambda_n)^2}{{\sigma_n}^2}y_{n,uv}^2\right\}\right\} \\ =& \frac{1}{Z}\exp\left\{-\frac{1}{2}\left(\text{tr}[(\dot{\mathbf{Y}}_u\dot{\mathbf{Y}}_u^T + \dot{\mathbf{Y}}_v\dot{\mathbf{Y}}_v^T)\boldsymbol{\Lambda}^{(1)}] \ + \text{tr}[\mathbf{Y}\mathbf{Y}^T\boldsymbol{\Lambda}^{(2)}]\right)\right\} \\ =& \frac{1}{Z}\exp\left\{\text{tr}\left[\boldsymbol{\Lambda}^{(1)}\mathbf{Y}\hat{\mathbf{K}}\mathbf{Y}^T\right] + \text{tr}[\boldsymbol{\Lambda}^{(2)}\mathbf{Y}\mathbf{Y}^T]\right\} \end{aligned} \tag{11.44}$$

where uv now runs through the 2D grid, $\Lambda^{(1)}$ and $\Lambda^{(2)}$ are defined as above and $\dot{\mathbf{Y}}_u = [\mathbf{y}_{uv} - \mathbf{y}_{(u+1)v}]$, $\dot{\mathbf{Y}}_v = [\mathbf{y}_{uv} - \mathbf{y}_{u(v+1)}]$ and $\hat{\mathbf{K}} = [\mathbf{K}_1\mathbf{K}_{T_1}]$ where $\mathbf{K}_j = \mathbf{P}_j\mathbf{P}_j^T$ with each $\mathbf{P}_j$ being a $T \times (T-1)$ matrix with elements $p_{jj} = 1$ and $p_{(i+j)i} = -1$ (the rest of the elements are zero). One can easily observe that the SACA prior falls into the general prior category we defined in Eq. 11.15, with $\mathbf{B}^{(1)} = \hat{\mathbf{K}}$ and $\mathbf{B}^{(2)} = \mathbf{I}$.

Therefore, following Eq. 11.21, the optimal weight matrix is found by setting

$$\mathbf{I} = \boldsymbol{\Lambda}^{(1)}\mathbf{W}\mathbf{X}\hat{\mathbf{K}}\mathbf{X}^T\mathbf{W}^T + \boldsymbol{\Lambda}^{(2)}\mathbf{W}\mathbf{X}\mathbf{X}^T\mathbf{W}^T. \tag{11.45}$$

It should be clear that since $\boldsymbol{\Lambda}^{(1)}$, $\boldsymbol{\Lambda}^{(2)}$ are diagonal matrices the above is satisfied if and only if $\mathbf{W}$ performs joint diagonalization of $\dot{\mathbf{X}}_u\dot{\mathbf{X}}_u^T + \dot{\mathbf{X}}_v\dot{\mathbf{X}}_v^T$ (or, equivalently, $\mathbf{X}\hat{\mathbf{K}}\mathbf{X}$) and $\mathbf{X}\mathbf{X}^T$.

We note that this is also the solution of the deterministic optimization problem

$$\begin{aligned}\mathbf{W} &= \arg\min_{\mathbf{W}} \sum_{j=1}^{p}\sum_{u=1}^{T_1}\sum_{v=1}^{T_2}\left(||\mathbf{w}_j^T(\mathbf{x}_{(u+1)v}-\mathbf{x}_{uv}||^2\right.\\ &\quad \left.+||\mathbf{w}_j^T(\mathbf{x}_{u(v+1)}-\mathbf{x}_{uv})||^2\right)\\ &= \arg\min_{\mathbf{W}} \mathrm{tr}[\mathbf{W}^T(\dot{\mathbf{X}}_u\dot{\mathbf{X}}_u^T+\dot{\mathbf{X}}_v\dot{\mathbf{X}}_v^T)\mathbf{W}]\\ &\quad \text{s.t.}\mathbf{W}^T\mathbf{X}\mathbf{X}^T\mathbf{W}=\mathbf{I}. \end{aligned} \tag{11.46}$$

We highlight the fact that we formulated a novel component analysis technique just by defining the the latent MRF (Eq. 11.44). The weight updates were then just retrieved by replacing in Eq. 11.21. We note that the SACA MRF connectivity falls trivially into the generalised framework we defined in Section 11.4, with $\mathbf{m}_{uv}^{\text{SACA}} = \frac{1}{2}\frac{\lambda_n}{\lambda_n+(1-\lambda_n)^2}\left(\mathbb{E}[\mathbf{y}_{u+1,v}]+\mathbb{E}[\mathbf{y}_{u,v+1}]\right)$ and $\mathbf{\Sigma}^{(\text{SACA})} = \mathbf{\Sigma}^{(\text{LDA})}$.

11.7 Experimental Evaluation

As proof of concept, we provide experiments both on synthetic (Section 11.7.1) and real-world data (Section 11.7.2, 11.7.3, 11.7.4). By the presented experiments, we aim to (i) experimentally validate the equivalence of the proposed probabilistic models to other models belonging in the same class (be it deterministic or probabilistic), and (ii) experimentally evaluate the performance of our models against others in the same class.

11.7.1 Synthetic Data

We demonstrate the application of our proposed probabilistic component analysis techniques on a set of synthetic data (see Fig. 11.2), generated utilising the Dimensionality Reduction Toolbox. In more detail, we compare the corresponding deterministic formulations of PCA, LDA and LLE to our proposed probabilistic models. The aim of this experiment is mainly to qualitatively illustrate the equivalence of the proposed methods (by observing how the probabilistic projections match the deterministic equivalents). Furthermore, the variance modelling per latent dimension in our EM-LDA is clear in $\mathbb{E}[\mathbf{y}]$ of the proposed EM-LDA (Fig. 11.2, Col. 3). This will prove beneficial prediction-wise, as we show in the following section.

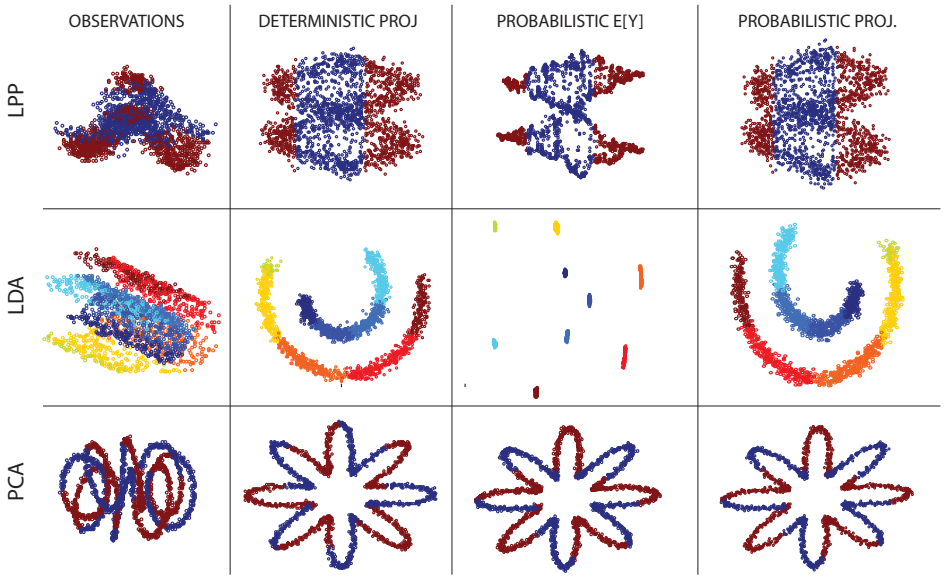


Figure 11.2: Synthetic experiments with deterministic LLE, LDA and PCA compared to our proposed probabilistic methods. For the deterministic models, the projections are shown in the 2nd column. For our probabilistic equivalents, we show the $\mathbb{E}[y]$ (3rd column) along with the projections (4th column). A neighbourhood of 12 was used in the case of LLE.

11.7.2 Real Data: Face Recognition via EM-LDA

One of the most common applications of LDA is face recognition. Therefore, we utilise various databases in order to verify the performance of our proposed EM-LDA. In more detail, we utilise the popular Extended Yale B database [81], as well as the PIE [237] and AR databases [153]. The experiments span a wide range of variability, such as various facial expressions (PIE, AR), illumination changes (Yale B, PIE) as well as pose changes (PIE).

Database Description: PIE, Yale B and AR

The CMU PIE database [237] contains faces under varying pose, illumination, and expression,consisting of more than 41000 images for a total of 68 subjects. We used a total of 170 images near frontal images for each subject. For training, we randomly selected a subset consisting of 5 images per subject, while for testing, the remaining images were used.

The extended Yale B database [81] contains a total of 16128 images of 38 subjects under 9 poses and 64 illumination conditions. We utilised a subset of 64 near frontal images per subject. For training, a random selection of a subset with 5 images per subject was used, while the rest of the images where used for testing.

Finally, the AR database [153] consists of more than 4000 frontal view face images of 126 subjects, while each subject is portrayed in upto 26 images, taken in two sessions, where the second was captured two weeks later from the first. Each session contains images under different facial expressions, illumination changes and occlusions. In our experiment, we focus on facial expressions. We firstly randomly select 100 subjects. Subsequently, use the images which portray varying facial expressions from session 1, while using the corresponding images from session 2 for testing.

Experimental Setting and Results

In related experiments, we compared our EM-LDA against deterministic LDA, the Fukunaga-Koontz variant (FK-LDA) [288] and PLDA [202] (which has been shown to outperform other probabilistic methods such as [111] in [139]) under the presence of Gaussian noise. We used the gradients of each image pixel as features, since as we experimentally verified, this improved the results for all compared methods. The errors of each compared method applied each database, accompanied by increasing Gaussian noise in the input, is shown in Fig. 11.3. Although PLDA offers a substantial improvement wrt. deterministic LDA and performs better than FK-LDA, it is clear that the proposed EM-LDA outperforms other compared LDA variants. This can be attributed to the explicit variance modelling (both for observations and per dimension) in our models, which appears to enable more robust classification.

11.7.3 Real Data: Level of Interest Detection

Automatically estimating the level of interest is a problem which has been gaining much attention by researchers lately, mostly to the vast applicability of such models, ranging from virtual guides to interactive learning systems as well as other applications pertaining to human-machine interaction. In this section, we aim to evaluate the performance of the proposed

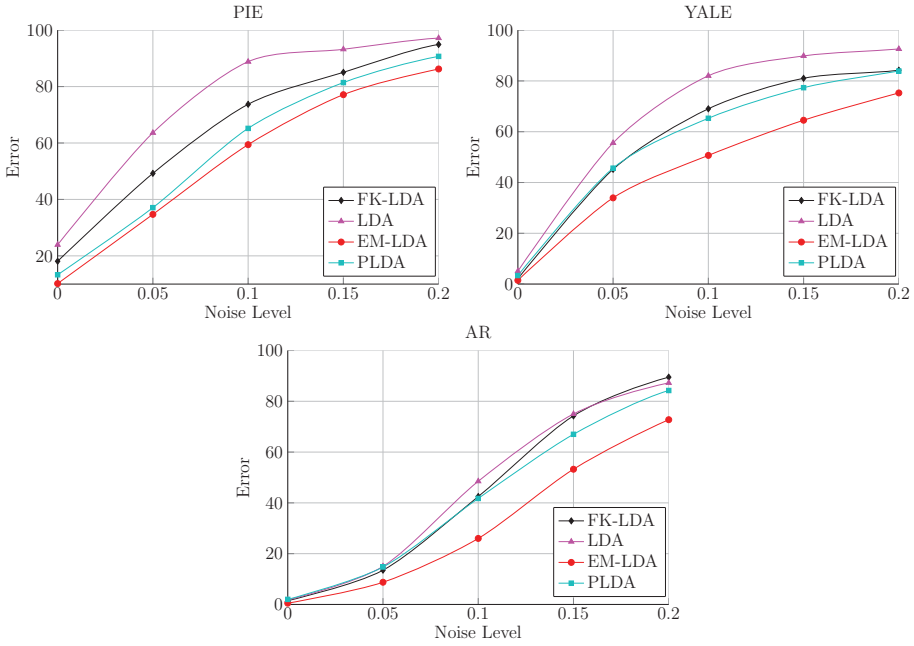


Figure 11.3: Recognition error for the databases PIE, YALE and AR under increasing Gaussian noise, comparing LDA, FK-LDA [288] the proposed EM-LDA and PLDA [202].

EM-LDA on the problem of level of interest detection. To this end, we focus on data consisting of video recordings of visitors to the Lisbon Zoo in 2013, interacting with a robot acting as a virtual guide. Sample images of the dataset can be seen in Fig. 11.4. The aforementioned data has been labelled in terms of three classes: *no interest*, When the subject is not interested in the interaction, is unmotivated and possibly wants to terminate it, *interest*, when the subject appears interested and eager to participate in the interaction, and *high interest*, when the subject appears pleased to participate in the interaction, and may show signs of enthusiasm or positive emotion expressions (e.g. laughter). We utilise the tracker described in [9], which is based on a discriminative regression based approach for Constrained Local Models (Discriminative Response Map Fitting). We utilise both face pose estimation (pitch, yaw and roll angles) along with 66 estimated facial landmarks (capturing eyebrows, eyes, nose and mouth/lips). For these experiments, we perform frame-based evaluation on both a binary interest detection problem (no-interest vs. interest) as well as the more complex 3 class prob-

Figure 11.4: Sample frames from the data used for the level of interest experiment.

lem of no-interest vs. interest vs. high interest. We maintain a balanced set for both training and testing, by selecting a random 1000 frames for each class, later separated into balanced training and testing sets. In order to evaluate in terms of noise-resilience, as in Section 11.7.2, we add increasing Gaussian noise to the features. As in previous experiments, we evaluate the proposed EM-LDA to deterministic LDA, Fukunaga-Koontz LDA (FK-LDA) [288] and PLDA [202]. Results of the described experiments are presented in Table 11.4. While FK-LDA seems to perform better than LDA under noisy scenarios, PLDA appears to overperform both, while clearly, EM-LDA achieves the best error rates against all compared methods.

11.7.4 Real Data: Face Visualisation via EM-LPP

One of the typical applications of Neighbour Embedding methods is the visualisation of , usually high-dimensional, data at hand. In particular, LPPs have often been used in visualising faces, providing an intuitive understanding of the variance and structural properties of the data [211], [103]. In order to evaluate the proposed EM-LPP, which is to the best of our knowledge

Table 11.4: Error obtained by applying the proposed EM-LDA, PLDA [202], classical LDA and FK-LDA [288] to the problem of interest detection (2-CLASS Interest vs. No Interest, 3-CLASS: No Interest vs. Low Interest vs. High Interest)

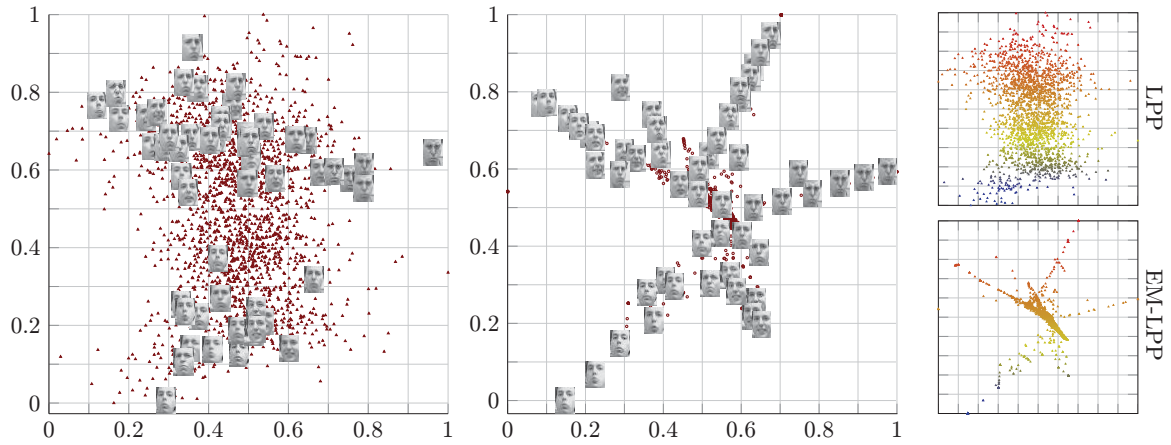
	2-CLASS Interest Detection				3-CLASS Interest Detection			
NOISE	**FK-LDA**	**LDA**	**EM-LDA**	**PLDA**	**FK-LDA**	**LDA**	**EM-LDA**	**PLDA**
0.00	0.21	0.26	**0.15**	0.23	0.33	0.31	**0.26**	0.32
0.05	0.33	0.29	**0.26**	0.29	0.40	0.39	**0.25**	0.38
0.10	0.38	0.37	**0.32**	0.36	0.45	0.43	**0.37**	0.39
0.15	0.43	0.41	**0.37**	0.41	0.50	0.50	**0.41**	0.44
0.20	0.46	0.46	**0.40**	0.44	0.53	0.56	**0.47**	0.47
0.25	0.45	0.46	**0.42**	0.45	0.56	0.57	**0.50**	0.51

the first probabilistic equivalent to LPP [179], we experiment on the Frey Faces database[5] [213], which contains 1965 images, captured as sequential frames of a video sequence. We apply a similar experiment to [103]. We firstly perturbed the images with random Gaussian noise, while subsequently we apply EM-LPP and LPP. The resulting space is illustrated in Fig. 11.5. It is clear that the deterministic LPP was unable to cope with the added Gaussian noise, failing to capture a meaningful data clustering. Note that the proposed EM-LPP was able to well capture the structure of the input data, modelling both pose and expression within the inferred latent space.

11.8 Conclusions

In this chapter we introduced a novel, unifying probabilistic component analysis framework, which reduces the construction of probabilistic component analysis models to essentially selecting the proper latent neighbourhood via the design of the latent connectivity. Our framework can thus be used to introduce novel probabilistic component analysis techniques by formulating new latent priors as products of MRFs. In this work, we have shown specific priors which when used, generate probabilistic models corresponding to PCA, LPP, LDA and SFA, while by doing so we introduced the first, favourable complexity-wise, probabilistic equivalent to LPP. Furthermore, we introduced a novel component analysis technique via our framework,

[5] http://www.cs.nyu.edu/~roweis/data.html



(a) LPP (b) EM-LPP (c)

Figure 11.5: Applying the proposed EM-LPP to the Frey Faces database, where each image is perturbed with random Gaussian noise. The latent projections obtained via LPP [179] and EM-LPP are illustrated in the figure. The inferred space in (a,b) is also annotated with face images from the database.

suitable for part-based dimensionality reduction. Finally, by means of theoretical analysis and experiments, we have demonstrated various advantages that our proposed methods pose against existing probabilistic and deterministic techniques.

Chapter 12

Discussion and Conclusions

12.1 Thesis Summary

In this thesis, a set of various novel methodologies were presented, aimed at solving a set of emerging challenges related to the fields of affective computing, machine learning and computer vision. During the last decade, the vast amounts of data that have been made available (initiating the so-called "Big-Data" era), along with the ongoing demand for applications that are able to cope under real-world conditions led to a series of shifts in the research direction employed in these fields. As a consequence, researchers transitioned from analysing posed expressions, usually from static images, to utilising high-quality video sequences with subjects portraying spontaneous behaviour. A further shift relates to moving away from adopting basic emotion categories in order to describe the affective state of the subject (e.g., anger, surprise), towards utilising more flexible and versatile emotion descriptions, such as the *continuous emotion dimensions*, which in effect model a much wider range of affective variability and can better capture the majority of emotions experienced in our everyday lives (e.g., excitement, boredom, interest). The work described in this thesis follows these shifts, by tackling highly challenging problems emerging from the adoption of these new directions in the affective sciences.

The first part of this thesis focused on learning continuous emotion dimensions. With the problem of analysing affect based on latent, continuous emotion dimensions still at its infancy

at the time this research was initiated, the problem was firstly approached by providing some of the first studies on predicting continuous emotion dimensions from multiple modalities (i.e. exploiting facial expressions, shoulder movements as well as acoustic features). An important contribution of our work lies in proposing and implementing the novel idea of modelling correlations and temporal patterns amongst output-dimensions in order to improve the accuracy of learning algorithms. To this end, we present several methods fitted to the task, such as utilising stacked Bidirectional Long-Short Term Memory Neural Networks for fusion (Chapter 5) in order to model both temporal and spatial relationships amongst the outputs. In Chapter 6 we derive a novel probabilistic regression technique based on the Relevance Vector Machine (RVM), which is able to account for such correlations. Finally, in Chapter 7 we describe Correlated Spaces Regression (CSR), a framework based on Canonical Correlation Analysis (CCA) which is suitable for both capturing the structure of output vectors, correlating the inputs with the outputs, as well as removing output redundancy. Since the outputs are projected into an uncorrelated space, this method facilitates the utilisation of single-output regression. In the same chapter, we also presented a set of experimental results regarding questions such as the correlations arising amongst emotion dimensions, as well as measuring the correlation of other emotions (such as the *level of interest* and other basic emotions) to the typically employed set of emotion dimensions. These empirical results end up motivating the utilisation of learning models which take into account the correlation of output dimensions, while also motivating the application of component analysis in order to extract more meaningful (as in more correlated with output emotion dimensions) features.

Conclusions drawn in the first part of this thesis highlight significant features which aid the development of systems aimed towards continuous and dimensional emotion analysis, such as modelling the temporal dynamics of both the input modalities (e.g., facial expressions, acoustic features) as well as the output emotion dimensions, the fusion of cues and modalities which may convey complementary information as well as the exploitation of the correlation of emotion dimensions. Nevertheless, an important limitation of the models presented in the first part, as well as most of the state-of-the-art methodologies, lies in utilising a simple average operation for fusing the multiple expert annotations for each emotion dimension. This approach is deemed suboptimal, since as repeatedly highlighted in this thesis, the annotation of emotion di-

mensions, usually performed by multiple experts, exhibits a significant spatio-temporal person-specific bias, while also being exposed to other forms of noise and irrelevant to-the-task information. By simply taking the average of all annotations as the ground truth, one makes implicit assumptions, namely (i) that all annotators are equally capable, (ii) that each annotation sample corresponds to the analogous sample of the sequence at hand, and (iii) that spatial noise can be cancelled out. Essentially, these assumptions are not valid in realistic scenarios, e.g., since the annotators usually have varying response times, the annotation sample will always have a positive temporal shift with respect to the sequence being annotated. In fact, the temporal discrepancies arising in the annotations can partially justify why in Chapter 7, we found that emotion dimensions seem to be better correlated to each other rather than to features such as facial expressions. As understandable, the fusion of multiple, continuous annotations is one of the most significant challenges in terms of modelling continuous emotion dimensions, as it is crucial to obtain a clean ground truth in order to properly train machine learning techniques.

In the second part of the thesis, we firstly dealt particularly with the problem of fusing multiple continuous annotations, having in mind that in order to solve this problem we need to eliminate both person-specific bias and noise, as well as heal any temporal discrepancies amongst the annotations. In Chapter 9, we turn to methods related to component analysis, initially focusing on a particular subset that we refer to as shared-space component analysis. In general, shared-space methods aim to capture a commonality manifesting amongst all observations, while some methods also model the individual (private) portions of the signal, which are specific to one set of observations. Based on the intuitive similarity of the shared-space of multiple observations to the ground truth derived from multiple annotations, in Chapter 9, we proposed a novel, shared-space method (Dynamic Probabilistic Canonical Correlation Analysis, DPCCA) which aims to provide a probabilistic representation of the ground truth as the inferred shared space, clean from spatio-temporal bias. DPCCA is able to decontaminate the annotations from any bias and person-specific characteristics by isolating them in the *private* space of the model, while learning the *shared* information conveyed by all annotators. Moreover, by utilising a time warping process, DPCCA temporally aligns the clean annotations, thus resolving any temporal discrepancies amongst the nnotations. DPCCA, as shown in Chapter 9, is also able to incorporate feature sets (such as e.g., tracked points encapsulating

facial expressions) during inference, in order to improve the derived annotation.

Subsequently, in Chapter 10, we follow on with our work on shared-space component analysis by proposing a novel, robust framework for multi-modal fusion and temporal alignment (Robust Canonical Correlation Analysis, RCCA), suitable for grossly corrupted high-dimensional observations. The robust property entails that the method can handle instances of non-Gaussian noise, commonly occurring in data acquired under real-world conditions. The method is evaluated on problems such as the robust temporal alignment of human behaviour, the robust audio-visual fusion for predicting the level of interest, as well as heterogeneous face recognition. It is worth noting that for some of the fusion experiments, we adopt the challenging scenario where one of the fused modalities is missing during testing.

Finally, Chapter 11 is focused on the problem of dimensionality reduction (and in particular, feature extraction) via probabilistic component analysis. In more detail, the chapter is focused on providing a theoretical unification of component analysis methods. In general, unifying frameworks are of crucial importance to sciences in general, as they facilitate the deeper understanding of a collection of methodologies. In this light, in Chapter 11 we presented the first, unifying framework for probabilistic component analysis, which unifies most well known component analysis techniques which can be formulated as a trace optimisation problem without domain constraints for the parameters. In particular, by formulating a probabilistic framework utilising Markov Random Fields (MRFs), we show how methods such as Principal Component Analysis (PCA), Linear Discriminant Analysis (LDA), Locality Preserving Projections (LPP) and others can be easily generated by manipulating the latent connectivity imposed by the utilised MRFs. By means of various experiments, we demonstrated the efficacy of the generated component analysis methods, on problems such as the detection of the level of interest and the visualisation and analysis of facial images. As we can conclude by the various experiments presented as well as the theoretical justification, methods derived via our framework pose several advantages against other related formulations, while our unifying framework facilitates the straightforward generation of novel component analysis techniques.

12.2 Future Work

There are many future research directions that arise from the work presented in this thesis. Application-wise, the proposed methods may be trained and evaluated on larger datasets, with the utilisation of cross-database evaluation. In fact, we expect cross-database evaluations to further demonstrate the advantage of applying fusion of multiple annotations as presented in Chapter 9, since the individual bias of disjoint sets of annotators (typically the case when utilising more than one databases) is expected to have a much higher variance. Furthermore, the methods presented in Chapters 5 and 6 for exploiting output correlations may be tested utilising more emotion dimensions than valence and arousal, in order to gain insight into how the models perform when more outputs are provided and if any performance gain is to be achieved. Specifically for OA-RVM (Chapter 6), it would be interesting to increment the model in order to learn temporal dependencies arising in the input features, while a further extension is to incorporate the temporal window parameter formally into the optimisation function in order to eliminate the need for cross-validation. Moreover, it would be interesting to explore the utilisation of other temporal kernels for constructing the design matrix of OA-RVM and perform related comparisons. Regarding Correlated-Spaces Regression (CSR, Chapter 7), a possible extension is to utilise a robust variant of Canonical Correlation Analysis (CCA) within CSR, in order to accommodate for other types of noise than Gaussian. Furthermore, CSR is an inherently static method due to the dependence on CCA. The modelling of temporal dependencies within such methods may be approached either via feature transformations or by reformulating the method and incorporating temporal constraints.

In Chapter 9, we presented DPCCA, a novel probabilistic method aimed at the temporal alignment and fusion of multiple sequences, such as dimensional emotion annotations. As aforementioned, the model can also take into consideration any input features (such as facial and acoustic features) when deriving the ground truth. An extension that may be explored lies in incrementing DPCCA in order to discover non-linear relationships. This is likely to be helpful e.g., when considering the relationship of high-dimensional feature spaces to the low-dimensional ground truth, which is likely to be non-linear. Furthermore, it is interesting to experiment with the warping process itself, e.g., by utilising warping constraints adapted

to the problem and data at hand in order to favour correcting lags which are more typical in annotations or to avoid warping more than the expected delay time for a given annotation task. Such experimentation is deemed highly appropriate in order to deal with specific idiosyncrasies which may arise in the data-at-hand. In Chapter 10, we presented a robust-to-gross-noise variant of Canonical Correlation Analysis (CCA). Future work on this model lies directly in the limitations posed, namely (i) being confined to two observation sets, and (ii) requiring the observation sets to be equal in terms of dimensionality (or pre-processed to be so). Extensions would require reformulating the problem to learn projections onto a common space regardless of dimensionality (thus rendering any relevant pre-processing unnecessary), as well as being able to handle multiple observation sets (i.e. a multi-set variant of RCCA). Finally, Chapter 11 is essentially a technical work which introduces a novel viewpoint in terms of providing a unifying framework for probabilistic component analysis methods. There is a multitude of future work spurring from the unifying framework. For example, exploiting the advantages of discriminative models by utilising e.g., Gaussian Processes instead of Markov Random Fields in order to learn non-linear mappings while still preserving the interesting characteristics of the observations. Furthermore, by introducing hierarchical hyperpriors, the unifying framework can be incremented in order to provide robust and sparse properties, thus inducing more easily generalisable models via the unifying framework. Finally, the framework can be extended to a supervised setting where some form of label information is available, in order to constraint the latent spaces. The framework may also be extended to multiset settings, where more than one observation sets are considered.

12.3 Conclusions

This work, in its entirety consists of a novel set of solutions for multiple highly challenging problems, revolving around the fields of affect sensing, machine learning and computer vision. The thesis is diverse, both in terms of focus and contribution, adopting both application oriented challenges and devising novel models aimed at solving them, such as the problem of fusing multiple expert annotations, as well as reaching out for theoretical answers to problems of a unifying nature, such as the unification of probabilistic component analysis. This is a

characteristic not only of the field itself, but also of a modern scientific direction, based on the increasing amount of information and knowledge exchanged between scientific disciplines as well the ever-easy access to scientific knowledge itself. This is in fact, a discourse of scientific languages, and the base of such communication lies in a *common* language amongst disciplines. Stay for too long confined within the symbolic barriers of your scientific language, you become constraint by it, limit your creativity and gradually loose the ability to understand the language of others while becoming less understandable to them, rendering any attempts on the communication and exchange of knowledge impotent. Nevertheless, spend too little time within it, and you can not understand it yourself. As it is often said, *only one who truly understands something can explain it simply*, that is, in the non-scientific, *common* language. There is clearly a balance to be struck. It is the authors hope that this thesis provides a small contribution towards this direction.

Bibliography

[1] N. Ahmed, T. Natarajan, and K. R. Rao. Discrete cosine transform. *Computers, IEEE Transactions on*, 100(1):90–93, 1974. 52

[2] K. Akisato, S. Masashi, S. Hitoshi, and K. Hirokazu. Designing various multivariate analysis at will via generalized pairwise expression. *Journal of Information Processing*, 6(1):136–145, 2013. 192

[3] G. Allport. *Social Psychology*. Houghton Mifflin, Boston, 1924. 44

[4] T. R. Almaev and M. F. Valstar. Local gabor binary patterns from three orthogonal planes for automatic facial expression recognition. In *Affective Computing and Intelligent Interaction (ACII)*, pages 356–361. IEEE, 2013. 51

[5] N. Alvarado. Arousal and valence in the direct scaling of emotional response to film clips. *Motivation and Emotion*, 21:323–348, 1997. 17, 86, 95, 104

[6] N. Ambady and R. Rosenthal. Half a minute: predicting teacher evaluations from thin slices of nonverbal behavior and physical attractiveness. *Journal of Personality and Social Psychology*, 1993. 48

[7] F. Arnold. *Attention and interest: A study in psychology and education*. Macmillan, 1910. 41

[8] A. Ashraf, S. Lucey, J. Cohn, T. Chen, Z. Ambadar, K. Prkachin, P.Solomon, and B. J. Theobald. The painful face: Pain expression recognition using active appearance models. *Proceedings of the 9th international conference on Multimodal interfaces*, 2007. 42

[9] A. Asthana, S. Zafeiriou, S. Cheng, and M. Pantic. Robust discriminative response map fitting with constrained local models. In *Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on*, pages 3444–3451. IEEE, 2013. 50, 215

[10] R. J. Aumann. Agreeing to disagree. *The Annals of Statistics*, 4(6):pp. 1236–1239. 160

[11] F. R. Bach and M. I. Jordan. A Probabilistic Interpretation of Canonical Correlation Analysis. Technical report, 2006. 75, 132, 147, 157

[12] P. Baldi, S. Brunak, P. Frasconi, G. Pollastri, and G. Soda. Exploiting the past and the future in protein secondary structure prediction. *Bioinformatics*, 15:937–946, 1999. 69

[13] T. Baltrušaitis, N. Banda, and P. Robinson. Dimensional affect recognition using continuous conditional random fields. In *IEEE FG*, 2013. 25, 41, 125, 129

[14] T. Bänziger and K. Scherer. Using actor portrayals to systematically study multimodal emotion expression: The GEMEP corpus. In *ACII '07: Proceedings of the Second International Conference on Affective Computing and Intelligent Interaction*, pages 476–487. Springer, 2007. 55

[15] B. Bao, G. Liu, C. Xu, and Y. S. Inductive robust principal component analysis. *IEEE Trans. Image Processing*, 21(8):3794–3800, 2012. 174, 182

[16] R. Baraniuk, V. Cevher, and M. Wakin. Low-dimensional models for dimensionality reduction and signal recovery: A geometric perspective. *Proceedings of the IEEE*, 98(6):959–971, 2010. 182

[17] M. Bartlett, J. Hager, P. Ekman, and T. Sejnowski. Measuring facial expressions by computer image analysis. *Psychophysiology*, 1999. 52

[18] M. Bartlett, G. Littlewort, M. Frank, C. Lainscsek, I. Fasel, and J. Movellan. Recognizing facial expression: machine learning and application to spontaneous behavior. *IEEE Conf. on Computer Vision and Pattern Recognition*, 2:568–573 vol. 2, June 2005. 42

[19] M. Bartlett, G. Littlewort, M. Frank, C. Lainscsek, I. Fasel, and J. Movellan. Fully automatic facial action recognition in spontaneous behavior. In *Proc. of the IEEE Int. Conf. on Automatic Face and Gesture Recognition*, 2006. 51

[20] A. Batliner, K. Fischer, R. Huber, J. Spilker, and E. Noth. How to find trouble in communication. *Speech Communication*, 2003. 42

[21] M. Belge, M. E. Kilmer, and E. L. Miller. Efficient determination of multiple regularization parameters in a generalized l-curve framework. *Inverse Problems*, 18(4):1161, 2002. 51

[22] P. Belhumeur, J. Hespanha, and D. Kriegman. Eigenfaces vs. fisherfaces: Recognition using class specific linear projection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19(7):711–720, 1997. 195

[23] Y. Bengio, P. Simard, and P. Frasconi. Learning long-term dependencies with gradient descent is difficult. *IEEE Trans. on Neural Networks*, 5(2):157–166, 1994. 68

[24] S. Bermejo and J. Cabestany. Oriented principal component analysis for large margin classifiers. *Neural Netw.*, 14(10):1447–1461, 2001. 91

[25] D. P. Bertsekas. *Constrained Optimization and Lagrange Multiplier Methods*. Athena Scientific, Belmont, MA, 2nd edition, 1996. 176, 178, 182

[26] C. M. Bishop. *Pattern Recognition and Machine Learning*. Springer-Verlag New York, Inc., Secaucus, NJ, USA, 2006. 72, 74, 150, 202, 206, 208

[27] M. J. Black and A. D. Jepson. Eigentracking: Robust matching and tracking of articulated objects using a view-based representation. *International Journal of Computer Vision*, 26(1):63–84, 1998. 50

[28] L. Bo and C. Sminchisescu. Structured output-associative regression. In *Proc. of IEEE Conf. on Computer Vision and Pattern Recognition*, pages 2403–2410, 2009. 105

[29] L. Bo and C. Sminchisescu. Twin Gaussian Processes for Structured Prediction. *International Journal of Computer Vision*, 87:28–52, 2010. 105

[30] M. Borga, T. Landelius, and H. Knutsson. A unified approach to PCA, PLS, MLR and CCA. 1997. 192

[31] L. Breiman. Stacked regressions. *Machine Learning*, 24:49–64, 1996. 111

[32] M. W. Browne. The maximum-likelihood solution in inter-battery factor analysis. *British Journal of Mathematical and Statistical Psychology*, 32(1):75–86, 1979. 75, 147

[33] D. Buller and J. Burgoon. Interpersonal deception theory. *Communication Theory*, 6(5):203–242, 1996. 43

[34] D. Buller, J. Burgoon, C. White, and A. Ebesu. Interpersonal deception: Vii. behavioral profiles of falsification, equivocation and concealment. *Journal of Language and Social Psychology*, 13(5):366–395, 1994. 43

[35] J. F. Cai, E. J. Candes, and Z. Shen. A singular value thresholding algorithm for matrix completion. *SIAM Journal Optimization*, 2(2):569–592, 2009. 176, 179

[36] N. Campbell and P. Mokhtari. Voice quality: the 4th prosodic dimension. *15 th International Congress of Phonetic Sciences*, 2003. 54

[37] E. J. Candes, X. Li, Y. Ma, and J. Wright. Robust principal component analysis? *Journal of ACM*, 58(3):1–37, 2011. 174, 176, 179

[38] G. Caridakis, K. Karpouzis, and S. Kollias. User and context adaptive neural networks for emotion recognition. *Neurocomput.*, 71(13-15):2553–2562, 2008. 88

[39] G. Caridakis, L. Malatesta, L. Kessous, N. Amir, A. Raouzaiou, and K. Karpouzis. Modeling naturalistic affective states via facial and vocal expressions recognition. In *Proc. of ACM Int. Conf. on Multimodal Interfaces*, pages 146–154, 2006. 40

[40] G. Celeux, F. Forbes, and N. Peyrard. Em procedures using mean field-like approximations for markov model-based image segmentation. *Pattern recognition*, 36(1):131–144, 2003. 203, 205

[41] G. Chanel, K. Ansari-Asl, and T. Pun. Valence-arousal evaluation using physiological signals in an emotion recall paradigm. In *Proc. of IEEE Int. Conf. on Systems, Man and Cybernetics*, pages 2662–2667, Oct. 2007. 40

[42] G. Chanel, J. Kronegg, D. Grandjean, and T. Pun. Emotion assessment: Arousal evaluation using eeg's and peripheral physiological signals. In *LNCS Vol. 4105*, pages 530–537, 2006. 88

[43] I. Cohen and et al. Facial expression recognition from video sequences: temporal and static modeling. *Computer Vision and Image Understanding*, 91:160– 187, 2003. 51

[44] J. Cohn and K. Schmidt. The timing of facial motion in posed and spontaneous smiles. *Active Media Technology*, 2003. 42, 48

[45] N. M. Correa, Y.-O. Li, T. Adali, and V. D. Calhoun. Fusion of fMRI, sMRI, and EEG data using canonical correlation analysis. In *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2009.*, pages 385–388. IEEE, 2009. 175

[46] C. Cortes, M. Mohri, and J. Weston. A general regression technique for learning transductions. In *Proc. of Int. Conf. on Machine Learning*, pages 153–160, New York, NY, USA, 2005. ACM. 105

[47] M. Coulson. Attributing emotion to static body postures: Recognition accuracy, confusions, and viewpoint dependence. *Nonverbal Behavior*, 28(2):117139, 2004. 44

[48] R. Cowie, E. Douglas-Cowie, S. Savvidou, E. McMahon, M. Sawey, and M. Schroder. Feeltrace: An instrument for recording perceived emotion in real time. In *Proc. of ISCA Workshop on Speech and Emotion*, pages 19–24, 2000. 56, 57

[49] R. Cowie, E. Douglas-Cowie, N. Tsapatsoulis, G. Votsis, S. Kollias, W. Fellenz, and J. Taylor. Emotion recognition in human-computer interaction. *Signal Processing Magazine, IEEE*, 18(1):32–80, Jan 2001. 46, 54

[50] N. Cristianini and J. Shawe-Taylor. *An Introduction to Support Vector Machines and Other Kernel-based Learning Methods*. Cambridge University Press, Cambridge, UK, March 2000. 72

[51] A. Cruttenden. *Intonation*. Cambridge University Press, Cambridge; New York, 1986. 46

[52] J. P. Cunningham and Z. Ghahramani. Unifying linear dimensionality reduction. *arXiv preprint arXiv:1406.0873*, 2014. 24, 192

[53] M. Dahmane and J. Meunier. Emotion recognition using dynamic grid-based hog features. In *Automatic Face & Gesture Recognition and Workshops (FG 2011), 2011 IEEE International Conference on*, pages 884–888. IEEE, 2011. 51

[54] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*, volume 1, pages 886–893. IEEE, 2005. 51, 52

[55] C. Darwin. *The expression of the emotions in man and animals*, volume 526. University of Chicago press, 1965. 11

[56] J. Davitz. Auditory correlates of vocal expressions of emotional meanings. *The Communication of Emotional Meaning*, pages 101–112, 1964. 38

[57] B. de Gelder, K. B. E. Bfcker, J. Tuomainen, M. Hensen, and J. Vroomen. The combined perception of emotion from voice and face: early interaction revealed by human electric brain responses. *Neuroscience Letters*, 260(2):133 – 136, 1999. 47

[58] F. De la Torre. A least-squares framework for component analysis. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 34(6):1041–1055, 2012. 24, 74, 76, 192

[59] F. De la Torre and J. F. Cohn. Facial expression analysis. In *Visual Analysis of Humans*, pages 377–409. Springer, 2011. 51

[60] B. DePaulo. Cues to deception. *Psychological Bulletin*, 129(1):74–118, 2003. 43

[61] R. Descartes. The passions of the soul, 1649. *The Philosophical Writings of Descartes*, 1, 1989. 11

[62] L. Devillers, I. Vasilescu, and L. Vidrascu. Anger versus fear detection in recorded conversations. *Proceedings of speech prosody*, 2004. 54

[63] D. Donoho. For most large underdetermined systems of equations, the minimal l1-norm near-solution approximates the sparsest near-solution. *Communications on Pure and Applied Mathematics*, 59(7):907–934, 2006. 176

[64] E. Douglas-Cowie, R. Cowie, I. Sneddon, C. Cox, L. Lowry, M. McRorie, L. Jean-Claude Martin, J.-C. Devillers, A. Abrilian, S. Batliner, A. Noam, and K. Karpouzis. The humaine database: addressing the needs of the affective computing community. In *Proc. of the Second Int. Conf. on Affective Computing and Intelligent Interaction*, pages 488–500, 2007. 14, 39, 56, 87, 112

[65] P. Dreuw, T. Deselaers, D. Rybach, D. Keysers, and H. Ney. Tracking using dynamic programming for appearance-based sign language recognition. *IEEE Automatic Face and Gesture Recognition*, 2006. 53

[66] H. Drucker, C. J. C. Burges, L. Kaufman, A. J. Smola, and V. Vapnik. Support vector regression machines. In *NIPS*, pages 155–161, 1996. 63, 72

[67] R. Edgeworth, B. Keen, E. Crane, M. Gross, and A. Arbor. Effect of speed on emotion-related kinematics during walking. *North American Congress on Biomechanics*, 2008. 45

[68] P. Ekman. About brows: Emotional and conversational signals. In M. Cranach, K. Foppa, W. Lepenies, and D. Ploog, editors, *Human Ethology: Claims and Limits of a New Discipline: Contributions to the Colloquium*, pages 169–248. Cambridge University Press, New York, 1979. 48

[69] P. Ekman. *Emotions in the Human Faces*. Studies in Emotion and Social Interaction. Cambridge University Press, 2 edition, 1982. 12, 14

[70] P. Ekman. Facial expression and emotion. *American psychologist*, 48(4):384, 1993. 12

[71] P. Ekman. Darwin, deception, and facial expression. *Ann. NY Acad. Sci*, 2003. 43

[72] P. Ekman and W. Friesen. *Facial Action Coding System: A Technique for the Measurement of Facial Movement*. Consulting Psychologists Press, Palo Alto, 1978. 44, 45

[73] P. Ekman, W. V. Friesen, and J. C. Hager. *Facial Action Coding System*. A Human Face, 2002. Salt Lake City. 16

[74] P. Ekman and E. L. Rosenberg. *What the face reveals: Basic and applied studies of spontaneous expression using the Facial Action Coding System*. Oxford University Press, Oxford, 2005. Oxford, UK. 12, 49

[75] F. Eyben, M. Wllmer, M. F. Valstar, H. Gunes, B. Schuller, and M. Pantic. String-based audiovisual fusion of behavioural events for the assessment of dimensional affect. In *Proc. of IEEE Int. Conf. on Automatic Face and Gesture Recognition*, pages 322–329, 2011. 41

[76] F. Eyben, M. Wöllmer, and B. Schuller. Opensmile: the munich versatile and fast open-source audio feature extractor. In *Proceedings of the international conference on Multimedia*, pages 1459–1462. ACM, 2010. 54

[77] I. Fasel, B. Fortenberry, and J. Movellan. A generative framework for real time object detection and classification. *Computer Vision and Image Understanding*, 2005. 50

[78] M. Fazel. *Matrix Rank Minimization with Applications*. PhD thesis, Dept. Electrical Engineering, Stanford University, CA, USA, 2002. 176

[79] J. R. Fontaine, K. R. Scherer, E. B. Roesch, and P. C. Ellsworth. The world of emotions is not two-dimensional. *Psychological science*, 18(12):1050–1057, 2007. 38

[80] N. Fragopanagos and J. G. Taylor. Emotion recognition in human-computer interaction. *Neural Networks*, 18(4):389–405, 2005. 40

[81] A. Georghiades, P. Belhumeur, and D. Kriegman. From few to many: Illumination cone models for face recognition under variable lighting and pose. *IEEE Trans. PAMI*, 23(6):643–660, 2001. 213, 214

[82] Z. Ghahramani, M. I. Jordan, and P. Smyth. Factorial hidden markov models. In *Machine Learning*. MIT Press, 1997. 150, 151

[83] Z. Ghahramani and S. T. Roweis. Learning nonlinear dynamical systems using an em algorithm. In *Advances in Neural Information Processing Systems 11*, pages 599–605. MIT Press, 1999. 149, 157

[84] S. Gilroy, M. Cavazza, M. Niiranen, E. Andre, T. Vogt, J. Urbain, M. Benayoun, H. Seichter, and M. Billinghurst. Pad-based multimodal affective fusion. In *Proc. of Int. Conf. on Affective Computing and Intelligent Interaction Workshops*, pages 1–8, 2009. 41

[85] D. Glowinski, A. Camurri, G. Volpe, N. Dael, and K. Scherer. Technique for automatic emotion recognition by body gesture analysis. In *Proc. of Computer Vision and Pattern Recognition Workshops*, pages 1–6, 2008. 40

[86] D. Gong and G. Medioni. Dynamic manifold warping for view invariant action recognition. In *Proc. 13th IEEE Int. Conf. Computer Vision*, pages 571–578, 2011. 75, 77, 174

[87] D. Grandjean, D. Sander, and K. R. Scherer. Conscious emotional experience emerges as a function of multilevel, appraisal-driven response synchronization. *Consciousness and Cognition*, 17:484–495, 2008. 39

[88] M. Graver. *Cicero on the emotions: Tusculan Disputations 3 and 4*. University of Chicago Press, 2002. 11

[89] A. Graves and J. Schmidhuber. Framewise phoneme classification with bidirectional LSTM and other neural network architectures. *Neural Networks*, 18:602–610, 2005. 68, 70

[90] S. Graves, G. Hooker, and J. Ramsay. Functional data analysis with r and matlab, 2009. 77

[91] M. Grimm and K. Kroschel. Emotion estimation in speech using a 3d emotion space concept. In *In Proc. IEEE Automatic Speech Recognition and Understanding Workshop*, pages 381–385, 2005. 40, 72, 91, 97, 101, 118

[92] M. Gross, E. Crane, and B. Fredrickson. Effect of felt and recognized emotions on gait kinematics. *American Society of Biomechanics Conference, Palo Alto, CA*, 2007. 45

[93] M. Gross, E. Crane, and B. Fredrickson. Expression of emotion changes gait kinematics. *International Society for Posture and Gait Research, Burlington, VT*, 2007. 45

[94] M. Gross, G. Gerstner, D. Koditschek, B. Fredrickson, and E. Crane. Emotion recognition from body movement kinematics. *American Society of Biomechanics, Portland, OR*, 2004. 45

[95] H. Gunes and M. Pantic. Automatic, dimensional and continuous emotion recognition. *Int. Journal of Synthetic Emotions*, 1(1):68–99, 2010. 12, 15, 88, 89, 91, 104, 120

[96] H. Gunes and M. Pantic. Dimensional emotion prediction from spontaneous head gestures for interaction with sensitive artificial listeners. In *Proc. of International Conference on Intelligent Virtual Agents*, pages 371–377, 2010. 40

[97] H. Gunes, M. Piccardi, and M. Pantic. From the lab to the real world: affect recognition using multiple cues and modalities. In *Affective computing: focus on emotion expression, synthesis, and recognition*, pages 185–218. InTech Education and Publishing, Vienna, Austria, 2008. 12, 46, 47, 48, 50, 53

[98] H. Gunes and B. Schuller. Categorical and dimensional affect analysis in continuous input: Current trends and future directions. *Image and Vision Computing*, 2012. 38

[99] H. Gunes, B. Schuller, M. Pantic, and R. Cowie. Emotion representation, analysis and synthesis in continuous space: A survey. In *Automatic Face & Gesture Recognition and Workshops (FG 2011), 2011 IEEE International Conference on*, pages 827–834. IEEE, 2011. 145, 155, 185

[100] N. Hadjikhani and B. de Gelder. Seeing fearful body expressions activates the fusiform cortex and amygdala. *Current Biology*, 13(24):2201–2205, December 2003. 44

[101] M. A. Hasan. On multi-set canonical correlation analysis. In *Proc. of the Int. Joint Conf. on Neural Networks*, IJCNN'09, pages 2640–2645, Piscataway, NJ, USA, 2009. IEEE Press. 152

[102] T. S. Hava and D. S. Eduardo. Turing computability with neural nets. *Applied Mathematics Letters*, 4:77–80, 1991. 67

[103] X. He, S. Yan, Y. Hu, P. Niyogi, and H. Zhang. Face recognition using laplacianfaces. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27(3):328–340, 2005. 200, 216, 217

[104] C. Hjortsjo. *"Man's face and mimic language"*. "Malmo, Studentlitteratur", "1970". 44

[105] S. Hochreiter. Untersuchungen zu dynamischen neuronalen Netzen. Diploma thesis, Institut für Informatik, Lehrstuhl Prof. Brauer, Technische Universität München, 1991. 68

[106] S. Hochreiter. The vanishing gradient problem during learning recurrent neural nets and problem solutions. *Int. J. Uncertain. Fuzziness Knowl.-Based Syst.*, 6(2):107–116, 1998. 68

[107] S. Hochreiter and J. Schmidhuber. Long short-term memory. *Neural Computation*, 9:1735–1780, 1997. 21

[108] H. Hotelling. Analysis of a complex of statistical variables into principal components. *J. Educational Psychology*, 24:417–441, 1933. 72, 73

[109] H. Hotelling. Relations between two sets of variates. *Biometrika*, 8:321–377, 1936. 72

[110] S. Ioannou, A. Raouzaiou, V. Tzouvaras, T. Mailis, K. Karpouzis, and S. Kollias. Emotion recognition through facial expression analysis based on a neurofuzzy method. *Journal of Neural Networks*, 18:423–435, 2005. 40

[111] S. Ioffe. Probabilistic Linear Discriminant Analysis. In *Computer Vision ECCV 2006*, pages 531–542, 2006. 195, 196, 197, 214

[112] A. K. Jain and S. Z. Li. *Encyclopedia of Biometrics: I-Z.*, volume 1. Springer, 2009. 51

[113] L. C. Jain and L. R. Medsker. *Recurrent Neural Networks: Design and Applications*. CRC Press, Inc., Boca Raton, FL, USA, 1999. 68

[114] I. N. Junejo, E. Dexter, I. Laptev, and P. Perez. View-independent action recognition from temporal self-similarities. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 33(1):172–185, 2011. 75, 174

[115] D. Jurafsky and J. H. Martin. *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics and Speech Recognition*. Prentice Hall, second edition, 2008. 87, 90, 113

[116] I. Kanluan, M. Grimm, and K. Kroschel. Audio-visual emotion recognition using an emotion recognition space concept. *Proc of the 16th European Signal Processing Conference*, 2008. 40, 72, 91

[117] A. Kapoor, Y. Qi, and R. Picard. Fully automatic upper facial action recognition. In *Proc. of the IEEE Int. Workshop on Analysis and Modeling of Faces and Gestures*, pages 195–202, 2003. 51

[118] M. Kim and V. Pavlovic. Discriminative Learning for Dynamic State Prediction. *IEEE Trans. Pattern Anal. Mach. Intell.*, 31(10):1847–1861, 2009. 149

[119] M. Kim and V. Pavlovic. Central subspace dimensionality reduction using covariance operators. *IEEE TPAMI*, 33(4):657–670, 2011. 125

[120] B. King, P. Smaragdis, and J. Mysore. Noise-robust dynamic time warping using plca features. pages 1973–1976, 2012. 75, 174

[121] A. Klami and S. Kaski. Probabilistic approach to detecting dependencies between data sets. *Neurocomput.*, 72(1-3):3946, Dec. 2008. 146, 147, 148

[122] A. Kleinsmith and N. Bianchi-Berthouze. Recognizing affective dimensions from body posture. In *Proc. of the Int. Conf. on Affective Computing and Intelligent Interaction*, pages 48–58, 2007. 40

[123] E. Kokiopoulou, J. Chen, and Y. Saad. Trace optimization and eigenproblems in dimension reduction methods. *Numerical Linear Algebra with Applications*, 18(3):565–602, 2011. 24, 192

[124] A. M. Kring and A. H. Gordon. Sex differences in emotion: Expression, experience, and physiology. *Journal of Personality & Social Psychology*, 74(3):686 – 703, 1998. 120

[125] M. Kubat. Neural networks: a comprehensive foundation by Simon Haykin, Macmillan, 1994, ISBN 0-02-352781-7. *Knowl. Eng. Rev.*, 13(4):409–412, 1999. 68

[126] D. Kulic and E. A. Croft. Affective state estimation for human-robot interaction. *IEEE Trans. on Robotics*, 23(5):991–1000, 2007. 40

[127] R. Laban and L. Ullmann. *The Mastery of Movement.* Princeton Book Company Publishers; 4 Revised edition, 1988. 45

[128] L. Lam and S. Suen. Application of majority voting to pattern recognition: an analysis of its behavior and performance. *Systems, Man and Cybernetics, Part A: Systems and Humans, IEEE Transactions on*, 27(5):553–568, 1997. 59

[129] R. Lane et al. *Cognitive Neuroscience of Emotion.* Oxford University Press, 2000. 14, 17, 86, 104, 124

[130] P. J. Lang, M. K. Greenwald, M. M. Bradley, and A. O. Hamm. Looking at pictures: Affective, facial, visceral, and behavioral reactions. *Psychophysiology*, 30(3):261–273, 1993. 41, 125, 130, 185

[131] M. A. Larkin, G. Blackshields, N. Brown, R. Chenna, P. McGettigan, et al. Clustal w and clustal x version 2.0. *Bioinformatics*, 23(21):2947–2948, 2007. 154

[132] R. Larsen and E. Diener. Affect intensity as an individual difference characteristic: A review. *Journal of research in personality*, 1987. 38, 39

[133] J. Laver. *Principles of Phonetics (Cambridge Textbooks in Linguistics).* Cambridge University Press, June 1994. 45

[134] C. Lee and S. Narayanan. Toward detecting emotions in spoken dialogs. *IEEE Trans. on Speech and Audio Processing*, 2005. 42

[135] T. S. Lee. Image representation using 2d gabor wavelets. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 18(10):959–971, 1996. 52

[136] R. Levenson. Emotion and the autonomic nervous system: A prospectus for research on autonomic specificity. *Social Psychophysiology and Emotion: Theory and Clinical Applications*, pages 17–42, 1988. 88, 89

[137] M. Lewis. *Handbook of emotions*. Guilford Press, 2008. 46

[138] P. A. Lewis, H. D. Critchley, P. Rotshtein, and R. J. Dolan. Neural correlates of processing valence and arousal in affective words. *Cerebral Cortex*, 17(3):742–748, MAR 2007. 14, 17, 86, 104

[139] P. Li, Y. Fu, U. Mohammed, J. H. Elder, and S. J. Prince. Probabilistic models for inference about identity. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 34(1):144–157, 2012. 214

[140] S. Z. Li, Z. Lei, and M. Ao. The HFB face database for heterogeneous face biometrics research. In *Computer Vision and Pattern Recognition Workshops, 2009. CVPR Workshops 2009. IEEE Computer Society Conference on*, pages 1–8. IEEE, 2009. 185, 186

[141] Z. Li, D. Lin, and X. Tang. Nonparametric discriminant analysis for face recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 31(4):755–761, 2009. 209, 210

[142] R. Lienhart and J. Maydt. an extended set of haar-like features for rapid object detection. *IEEE ICIP*, 2002. 50

[143] Z. Lin, R. Liu, and Z. Su. Linearized alternating direction method with adaptive penalty for low-rank representation. In *Proc. 2011 Neural Information Processing Systems Conf.*, pages 612–620, Granada, Spain, 2011. 176, 177, 179, 180

[144] J. Listgarten, R. Neal, S. Roweis, and A. Emili. Multiple alignment of continuous time series. volume 17, 2005. 75, 174

[145] G. Littlewort, M. Bartlett, and K. Lee. Faces of pain: automated measurement of spontaneous facial expressions of genuine and posed pain. *Proceedings of the 9th international conference on Multimodal interfaces*, 2007. 43

[146] E. F. Lock, K. A. Hoadley, J. Marron, and A. B. Nobel. Joint and individual variation explained (jive) for integrated analysis of multiple data types. *The annals of applied statistics*, 7(1):523, 2013. 182

[147] D. G. Lowe. Distinctive image features from scale-invariant keypoints. *International journal of computer vision*, 60(2):91–110, 2004. 51, 52

[148] B. Lucas and T. Kanade. An iterative image registration technique with an application to stereo vision. In *Proc. of the Int. Joint Conf. on Artificial Intelligence*, pages 121–130, 1981. 50

[149] S. Lucey, A. B. Ashraf, and J. Cohn. Investigating spontaneous facial action recognition through aam representations of the face. *Face recognition*, pages 275–286, 2007. 51

[150] M. Lyons, S. Akamatsu, M. Kamachi, and J. Gyoba. Coding facial expressions with gabor wavelets. In *Automatic Face and Gesture Recognition, 1998. Proceedings. Third IEEE International Conference on*, pages 200–205. IEEE, 1998. 51

[151] M. Wollmer, and F. Eyben, and S. Reiter, and B. Schuller, and C. Cox, and E. Douglas-Cowie, and R. Cowie . Abandoning emotion classes - towards continuous emotion recognition with modelling of long-range dependencies. In *Proc. of 9th Interspeech Conf.*, pages 597–600, 2008. 40, 70, 72, 89, 91, 96, 101, 117, 118

[152] S. Mariooryad and C. Busso. Analysis and compensation of the reaction lag of evaluators in continuous emotional annotations. In *in Affective Computing and Intelligent Interaction (ACII 2013)*, 2013. 59

[153] A. M. Martinez. The AR face database. *CVC Technical Report*, 24, 1998. 213, 214

[154] B. Martinez, M. F. Valstar, X. Binefa, and M. Pantic. Local evidence aggregation for regression-based facial point detection. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 35(5):1149–1163, 2013. 50

[155] W. M. Massaro and M. M. Cohen. Fuzzy logical model of bimodal emotion perception: Comment on "the perception of emotions by ear and by eye"; by de gelder and vroomen. *Cognition & Emotion*, 14(3):313–320, 2000. 47

[156] I. Matthews and S. Baker. Active appearance models revisited. *International Journal of Computer Vision*, 60(2):135–164, 2004. 50

[157] G. McKeown et al. The semaine database: Annotated multimodal records of emotionally colored conversations between a person and a limited agent. *IEEE TAC*, 2012. 14, 15, 39, 57, 58, 124, 125, 129, 184

[158] G. McKeown, M. F. Valstar, R. Cowie, and M. Pantic. The semaine corpus of emotionally coloured character interactions. In *2010 IEEE Int. Conf. on Multim. and Expo*, pages 1079 –1084, 2010. 164

[159] D. McNeill. *The conceptual basis of language / David McNeill.* Lawrence Erlbaum Associates ; distributed by the Halsted Press, Division of Wiley, Hillsdale, N.J. : New York :, 1979. 47

[160] D. McNeill. So you think gestures are nonverbal? *Psychological Review*, 92:350–371, 1985. 47

[161] D. Mcneill. *Language and Gesture (Language Culture and Cognition).* Cambridge University Press, August 2000. 48, 49

[162] H. K. Meeren, C. C. Van Heijnsbergen, and B. De Gelder. Rapid perceptual integration of facial expression and emotional body language. *Proc. of the National Academy of Sciences of the USA*, 102:1651816523, 2005. 47

[163] A. Mehrabian and J. A. Russell. *An Approach to Environmental Psychology.* MIT Press, 1980. 38

[164] A. Metallinou, A. Katsamanis, Y. Wang, and S. Narayanan. Tracking changes in continuous emotion states using body language and prosodic cues. In *Acoustics, Speech and Signal Processing (ICASSP), 2011 IEEE International Conference on*, pages 2288–2291. IEEE, 2011. 41, 58

[165] K. Mitra, A. Veeraraghavan, and R. Chellappa. Robust rvm regression using sparse outlier model. In *Proc. of IEEE Conf. on Computer Vision and Pattern Recognition*, pages 1887–1894, 2010. 105

[166] S. Mitra and T. Acharya. Gesture recognition: a survey. *Systems, Man, and Cybernetics, Part C: Applications and Reviews, IEEE Trans.*, 2007. 53

[167] J. Montepare, E. Koff, D. Zaitchik, and M. Albert. The use of body movement and gestures as cues to emotions in younger and older adults. *Journal of Nonverbal Behavior*, 23:133–152, 1999. 44

[168] K. P. Murphy. *Machine learning: a probabilistic perspective*. MIT press, 2012. 72

[169] R. R. N. Ambady. Thin slices of expressive behavior as predictors of interpersonal consequences : a meta-analysis. *Psychological Bulletin*, 111(2):256–274, 1992. 46

[170] B. K. Natarajan. Sparse approximate solutions to linear systems. *SIAM J. Comput.*, 24(2):227–234, 1995. 176

[171] M. H. Nguyen and F. De la Torre. Local minima free parameterized appearance models. In *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*, pages 1–8. IEEE, 2008. 50

[172] M. A. Nicolaou, H. Gunes, and M. Pantic. Audio-visual classification and fusion of spontaneous affective data in likelihood space. In *In Proc. of IEEE Int. Conf. on Pattern Recognition*, pages 3695–3699, 2010. 40, 116

[173] M. A. Nicolaou, H. Gunes, and M. Pantic. Automatic segmentation of spontaneous data using dimensional labels from multiple coders. In *Proc. of LREC Int. Workshop on Multimodal Corpora: Advances in Capturing, Coding and Analyzing Multimodality*, pages 43–48, 2010. 59, 88, 89, 95, 112

[174] M. A. Nicolaou, H. Gunes, and M. Pantic. Continuous prediction of spontaneous affect from multiple cues and modalities in valence-arousal space. *IEEE TAC*, 2011. 25, 34, 58, 81, 105, 128, 155, 185

[175] M. A. Nicolaou, H. Gunes, and M. Pantic. Continuous prediction of spontaneous affect from multiple cues and modalities in valence-arousal space. *Affective Computing, IEEE Transactions on*, 2(2):92–105, 2011. 85

[176] M. A. Nicolaou, H. Gunes, and M. Pantic. Output-associative rvm regression for dimensional and continuous emotion prediction. In *Proceedings of IEEE FG'11*, pages 16–23, Santa Barbara, CA, USA, March 2011. 124, 125

[177] M. A. Nicolaou, S. Zafeiriou, and M. Pantic. Correlated-spaces regression for learning continuous emotion dimensions. In *Proceedings of the 21st ACM international conference on Multimedia*, pages 773–776. ACM, 2013. 125

[178] H. Ning, T. Han, Y. Hu, Z. Zhang, Y. Fu, and T. Huang. a realtime shrug detector. *Proc. of the IEEE Int. Conf. on Automatic Face and Gesture Recognition*, 2006. 53

[179] X. Niyogi. Locality preserving projections. In *Advances in neural information processing systems 16: proceedings of the 2003 conference*, volume 16, page 153. The MIT Press, 2004. 196, 217, 218

[180] T. Ojala, M. Pietikainen, and T. Maenpaa. Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 24(7):971–987, 2002. 51, 52

[181] A. M. Oliveira, M. P. Teixeira, I. B. Fonseca, and M. Oliveira. Joint model-parameter validation of self-estimates of valence and arousal: Probing a differential-weighting model of affective intensity. In *Proc. of the 22nd Annual Meeting of the Int. Society for Psychophysics*, pages 245–250, 2006. 17, 86, 95, 104

[182] J. Orozco et al. Hierarchical on-line appearance-based tracking for 3d head pose, eyebrows, lips, eyelids and irises. *Image and Vision Computing*, February 2013. 126, 184

[183] C. E. Osgood, G. Suci, and P. Tannenbaum. *The measurement of meaning*. University of Illinois Press, Urbana, IL, 1957. 38

[184] M. Pantic and M. Bartlett. Machine analysis of facial expressions. In K. Delac and M. Grgic, editors, *Face Recognition*, pages 377–416. I-Tech Education and Publishing, Vienna, Austria, July 2007. 52

[185] M. Pantic, A. Nijholt, A. Pentland, and T. S. Huanag. Human-centred intelligent human? computer interaction (hci^2): how far are we from attaining it? *International Journal of Autonomous and Adaptive Communications Systems*, 1(2):168–187, 2008. 15

[186] M. Pantic and I. Patras. Detecting facial actions and their temporal segments in nearly frontal–view face image sequences. In *Proc. of the IEEE Int. Conf. on Systems, Man and Cybernetics*, volume 4, pages 3358–3363, 2005. 51

[187] M. Pantic and I. Patras. Dynamics of facial expression: Recognition of facial actions and their temporal segments from face profile image sequences. *IEEE Trans. on Systems, Man, and Cybernetics, Part B*, 2006. 51, 52

[188] M. Pantic and L. Rothkrantz. Facial action recognition for facial expression analysis from static face images. *IEEE Trans. on Systems, Man, and Cybernetics, Part B*, 2004. 51

[189] M. Pantic, M. Valstar, R. Rademaker, and L. Maat. Web-based database for facial expression analysis. *IEEE International Conference on Multimedia and Expo, 2005.*, 2005. 44, 55, 187

[190] I. Patras and M. Pantic. Particle filtering with factorized likelihoods for tracking facial features. In *Proc. of IEEE Int. Conf. on Automatic Face and Gesture Recognition*, pages 97–102, 2004. 87, 90, 112, 164, 188

[191] B. Paul. Accurate short-term analysis of the fundamental frequency and the harmonics-to-noise ratio of a sampled sound. In *In Proceedings of the Institute of Phonetic Sciences*, pages 97–110, 1993. 90, 113

[192] I. Pavlidis, J. Levine, and P. Baukol. Thermal imaging for anxiety detection. In *CVBVS '00: Proceedings of the IEEE Workshop on Computer Vision Beyond the Visible Spec-*

trum: Methods and Applications (CVBVS 2000), page 104, Washington, DC, USA, 2000. IEEE Computer Society. 47

[193] K. Pearson. On lines and planes of closest fit to systems of points in space. *Philosophical Magazine*, 2:559–572, 1901. 72

[194] Y. Peng, A. Ganesh, J. Wright, W. Xu, and Y. Ma. Rasl: Robust alignment by sparse and low-rank decomposition for linearly correlated images. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 34:2233–2246, 2012. 174, 182

[195] A. Pentland and A. Madan. Perception of social interest. In *Proc. IEEE Int. Conf. on Computer Vision, Workshop on Modeling People and Human Interaction (ICCV-PHI)*, 2005. 26, 41, 184

[196] S. Petridis, H. Gunes, S. Kaltwang, and M. Pantic. Static vs. dynamic modeling of human nonverbal behavior from multiple cues and modalities. In *Proc. of ACM Int. Conf. on Multimodal Interfaces*, pages 23–30, 2009. 90, 94, 113

[197] P. Petta, C. Pelachaud, and R. Cowie. Emotion-oriented systems. *The Humaine Handbook, ISBN*, 2011. 51

[198] R. W. Picard. Affective computing. *MIT Technical Report*, 1995. 12

[199] M. K. Pitt and N. Shephard. Filtering via simulation: auxiliary particle filters. *J. Am. Statistical Association*, 94(446):590–616, 1999. 87, 91, 113

[200] R. Plutchik and H. R. Conte. *Circumplex models of personality and emotions*. Washington, DC: American Psychological Association, 1997. 38

[201] R. Poppe. Vision-based human motion analysis: An overview. *Computer Vision and Image Understanding*, 2007. 53

[202] S. J. D. Prince and J. H. Elder. Probabilistic linear discriminant analysis for inferences about identity. In *ICCV*, 2007. 195, 214, 215, 216, 217

[203] R. Provine. Yawns, laughs, smiles, tickles, and talking. In J. A. Russell and J. M. Fernandez-Dols, editors, *The Psychology of Facial Expression*, pages 158–175. 1997. 46

[204] W. Qian and D. Titterington. Estimation of parameters in hidden markov models. *Phil. Trans. of the Royal Society of London. Series A: Physical and Engineering Sciences*, 337(1647):407–428, 1991. 203

[205] L. Rabiner and B. H. Juang. *Fundamentals of Speech Recognition*. Prentice Hall, united states ed edition, Apr. 1993. 153, 154, 161

[206] G. A. Ramirez, T. Baltrušaitis, and L.-P. Morency. Modeling latent discriminative dynamic of multi-dimensional affective signals. In *Affective Computing and Intelligent Interaction*, pages 396–406. Springer, 2011. 41

[207] J. O. Ramsay. *Functional data analysis*. Wiley Online Library, 2006. 77

[208] V. C. Raykar, S. Yu, L. H. Zhao, A. Jerebko, C. Florin, G. H. Valadez, L. Bogoni, and L. Moy. Supervised learning from multiple experts: whom to trust when everyone lies a bit. In *Proceedings of the 26th Annual International Conference on Machine Learning*, pages 889–896. ACM, 2009. 59

[209] V. C. Raykar, S. Yu, L. H. Zhao, G. H. Valadez, C. Florin, L. Bogoni, and L. Moy. Learning from crowds. *The Journal of Machine Learning Research*, 99:1297–1322, 2010. 59, 145, 146, 150, 166, 167

[210] R. Rosenthal, K. Scherer, and J. Harrigan. *Vocal expression of affect, In: The New Handbook of Methods in Nonverbal Behavior Research*. Harrigan, 2005. 46

[211] S. Roweis. EM algorithms for PCA and SPCA. *Advances in neural information processing systems*, pages 626–632, 1998. 73, 192, 194, 206, 216

[212] S. Roweis and Z. Ghahramani. A unifying review of linear gaussian models. *Neural Comput.*, 11(2):305–345, Feb. 1999. 150, 151, 152, 192, 205

[213] S. T. Roweis and L. K. Saul. Nonlinear dimensionality reduction by locally linear embedding. *Science*, 290(5500):2323–2326, 2000. 217

[214] O. Rudovic. *Machine Learning Techniques for Automated Analysis of Facial Expressions*. PhD thesis, December 2013. 44

[215] H. Rue and L. Held. *Gaussian Markov random fields: theory and applications*. CRC Press, 2004. 206

[216] J. A. Russell. A circumplex model of affect. *Journal of Personality and Social Psychology*, 39, 1980. 14, 38, 128

[217] J. A. Russell. Culture and the categorization of emotions. *Psychological bulletin*, 110(3):426, 1991. 11

[218] D. Ruta and B. Gabrys. Classifier selection for majority voting. *Information fusion*, 6(1):63–81, 2005. 59

[219] H. Sakoe and S. Chiba. Dynamic programming algorithm optimization for spoken word recognition. *IEEE Trans. Acoustics, Speech, and Signal Processing*, (1):43–49, 1978. 75, 174, 176

[220] K. Scherer. *The Neuropsychology of Emotion*, chapter Psychological models of emotion, pages 137–162. Oxford University Press, 2000. 39

[221] K. Schmidt and J. Cohn. Human facial expressions as adaptations: Evolutionary questions in facial expression research. *Yearbook of Physical Anthropology*, 2001. 48

[222] B. Schölkopf and A. J. Smola. *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond (Adaptive Computation and Machine Learning)*. The MIT Press, 2001. 72

[223] S. Schotz. Linguistic & paralinguistic phonetic variation in speaker recognition & text-to-speech synthesis. term paper for course. In *in Speech Technology, GSLT.*, 2002. 46

[224] C. Schuldt, I. Laptev, and B. C. Recognizing human actions: A local svm approach. In *Proc. 17th Int. Conf. Pattern Recognition*, pages 32–36, Washington, DC, USA, 2004. 186

[225] B. Schuller and A. Batliner. *Computational paralinguistics: emotion, affect and personality in speech and language processing*. John Wiley & Sons, 2013. 46, 54

[226] B. Schuller, N. Köhler, R. Müller, and G. Rigoll. Recognition of interest in human conversational speech. In *INTERSPEECH*, 2006. 42

[227] B. Schuller, R. Müller, F. Eyben, J. Gast, B. Hörnler, M. Wöllmer, G. Rigoll, A. Höthker, and H. Konosu. Being bored? recognising natural interest by extensive audiovisual integration for real-life application. *Image and Vision Computing*, 27(12):1760–1774, 2009. 26, 41, 42, 126, 184, 185

[228] B. Schuller and G. Rigoll. Recognising interest in conversational speech-comparing bag of frames and supra-segmental features. In *INTERSPEECH*, pages 1999–2002, 2009. 26, 41, 42, 184

[229] B. Schuller, M. Valstar, et al. Avec 2012: the continuous audio/visual emotion challenge - an introduction. In *ICMI*, pages 361–362, 2012. 127

[230] M. Schuster and K. K. Paliwal. Bidirectional recurrent neural networks. *IEEE Trans. on Signal Processing*, 45:2673–2681, November 1997. 69

[231] C. Shan, S. Gong, and P. W. McOwan. Beyond facial expressions: Learning human emotion from body gestures. In *BMVC*, pages 1–10, 2007. 175

[232] S. Shariat and V. Pavlovic. Isotonic CCA for sequence alignment and activity recognition. pages 2572–2578, 2011. 175

[233] J. Shi and J. Malik. Normalized cuts and image segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.*, 22(8):888–905, Aug. 2000. 160

[234] H. T. Siegelmann. *Neural Networks and Analog Computation: Beyond the Turing Limit (Progress in Theoretical Computer Science)*. Birkhäuser Boston, 1 edition, December 1998. 67

[235] C. Sigg and J. Buhmann. Expectation-maximization for sparse and non-negative pca. In *Proceedings of the 25th international conference on Machine learning*, pages 960–967. ACM, 2008. 74

[236] P. J. Silvia. *Exploring the psychology of interest*. Oxford University Press, 2006. 41

[237] T. Sim, S. Baker, and M. Bsat. The cmu pose, illumination, and expression (pie) database. In *Proc. of the IEEE Int. Conf. on Automatic Face and Gesture Recognition*, May 2002. 213

[238] B. Spinoza. Ethics, 1677. 11

[239] K. N. Spreckelmeyer, M. Kutas, T. P. Urbach, E. Altenmller, and T. F. Mnte. Combined perception of emotion in pictures and musical sounds. *Brain Research*, 1070(1):160 – 170, 2006. 47

[240] L. Sun, S. Ji, and J. Ye. A least squares formulation for canonical correlation analysis. In *Proceedings of the 25th international conference on Machine learning*, pages 1024–1031. ACM, 2008. 74, 182

[241] L. Sun, S. Ji, and J. Ye. A least squares formulation for a class of generalized eigenvalue problems in machine learning. In *Proceedings of the 26th Annual International Conference on Machine Learning*, pages 977–984. ACM, 2009. 24, 76, 192

[242] A. Tellegen and D. Watson. Toward a consensual structure of mood. *Psychological Bulletin*, 1985. 38, 39

[243] A. Thayananthan, R. Navaratnam, B. Stenger, P. H. S. Torr, and R. Cipolla. Multivariate relevance vector machines for tracking. In *In ECCV*, pages 124–138. Springer-Verlag, 2006. 105

[244] R. Thayer. *The biopsychology of mood and arousal*. Oxford University Press US, 1989. 38, 39

[245] Y. Tian, T. Kanade, and J. F. Cohn. Evaluation of gabor-wavelet-based facial action unit recognition in image sequences of increasing complexity. In *FGR '02: Proceedings*

of the Fifth IEEE International Conference on Automatic Face and Gesture Recognition, page 229, Washington, DC, USA, 2002. IEEE Computer Society. 52

[246] M. E. Tipping. Sparse bayesian learning and the relevance vector machine. *JMLR*, 1:211–244, 2001. 21, 63, 70, 71, 72, 107, 108, 109, 127, 166, 184

[247] M. E. Tipping and C. M. Bishop. Mixtures of probabilistic principal component analyzers. *Neural Comput.*, 11(2):443–482, Feb. 1999. 6

[248] M. E. Tipping and C. M. Bishop. Probabilistic principal component analysis. *Journal of the Royal Statistical Society, Series B*, 61:611–622, 1999. 73, 74, 192, 194, 206

[249] M. E. Tipping and A. Faul. Fast marginal likelihood maximisation for sparse bayesian models. In *Proc. of Int. Workshop on Artificial Intelligence and Statistics*, pages 3–6, 2003. 21, 71, 72, 109, 116

[250] S. S. Tomkins. Affect, imagery, consciousness: Vol. i. the positive affects. 1962. 41

[251] K. P. Truong, D. A. Leeuwen van, M. A. Neerincx, and F. M. Jong de. Arousal and valence prediction in spontaneous emotional speech: Felt versus perceived emotion. In *Proc. of Interspeech*, pages 2027–2030, 2009. 97, 101, 118

[252] P. Tsiamyrtzis, J. Dowdall, D. Shastri, I. Pavlidis, M. Frank, and P. Ekman. Imaging facial physiology for the detection of deceit. *International Journal of Computer Vision*, 2007. 47

[253] L. R. Tucker. An inter-battery method of factor analysis. *Psychometrika*, 23(2):111–136, 1958. 73, 75, 147

[254] R. Turner and M. Sahani. A maximum-likelihood interpretation for slow feature analysis. *Neural computation*, 19(4):1022–1038, 2007. 197, 198, 200, 202

[255] M. Valstar, B. Martinez, X. Binefa, and M. Pantic. Facial point detection using boosted regression and graph models. In *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*, pages 2729–2736. IEEE, 2010. 50

[256] M. Valstar and M. Pantic. Fully automatic facial action unit detection and temporal analysis. In *CVPRW '06: Proceedings of the 2006 Conference on Computer Vision and Pattern Recognition Workshop*, page 149, Washington, DC, USA, 2006. IEEE Computer Society. 42, 51, 52

[257] M. Valstar, B. Schuller, K. Smith, F. Eyben, B. Jiang, S. Bilakhia, S. Schnieder, R. Cowie, and M. Pantic. Avec 2013: the continuous audio/visual emotion and depression recognition challenge. In *Proceedings of the 3rd ACM international workshop on Audio/visual emotion challenge*, pages 3–10. ACM, 2013. 41, 51

[258] M. F. Valstar, H. Gunes, and M. Pantic. How to distinguish posed from spontaneous smiles using geometric features. In *Proc. of the ACM Int. Conf. on Multimodal Interfaces*, pages 38–45, 2007. 42, 43, 48, 51, 53

[259] J. Van den Stock, R. Righart, and B. de Gelder. Body expressions influence recognition of emotions in the face and voice. *Emotion*, 7(3):487–494, August 2007. 44, 47

[260] R. Van der Merwe and E. Wan. The square-root unscented kalman filter for state and parameter-estimation. In *In Proc. of IEEE Int. Conf. on Acoustics, Speech, and Signal Processing, 2001.*, volume 6, pages 3461 –3464 vol.6, 2001. 151

[261] L. Vandenberghe and S. Boyd. Semidefinite programming. *SIAM Review*, 38(1):49–95, 1996. 176

[262] A. Vedaldi and A. Zisserman. Efficient additive kernels via explicit feature maps. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 34(3):480–492, 2012. 64

[263] P. Viola and M. Jones. Robust real-time face detection. *International Journal of Computer Vision*, 2004. 50

[264] D. Vrakas and I. P. Vlahavas. *Artificial Intelligence for Advanced Problem Solving Techniques*. Information Science Reference - Imprint of: IGI Publishing, Hershey, PA, 2008. 72

[265] H. Vu, C. Carey, and S. Mahadevan. Manifold warping: Manifold alignment over time. In *Proc. 26th Conference on Artificial Intelligence*, 2012. 75, 77, 174

[266] D. Vukadinovic and M. Pantic. Fully automatic facial feature point detection using gabor feature based boosted classifiers. In *Proc. of the IEEE Int. Conf. on Systems, Man and Cybernetics*, volume 2, pages 1692–1698, 2005. 50, 90

[267] J. Wagner, J. Kim, and E. Andre. From physiological signals to emotions: Implementing and comparing selected methods for feature extraction and classification. In *Proc. of IEEE Int. Conf. on Multimedia and Expo*, pages 940–943, 2005. 40

[268] X. Wang and X. Tang. Face photo-sketch synthesis and recognition. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 31(11):1955–1967, 2009. 186

[269] J. P. Weston et al. Kernel dependency estimation. Technical Report 98, Germany, August 2002. 105

[270] C. M. Whissell. The dictionary of affect in language. *Emotion: Theory, research and experience. The measurement of emotions*, 4:113–131, 1989. 46

[271] J. Whitehill and C. W. Omlin. Haar features for facs au recognition. In *Automatic Face and Gesture Recognition, 2006. FGR 2006. 7th International Conference on*, pages 5–pp. IEEE, 2006. 51, 52

[272] A. Wilson, A. Bobick, and J. Cassell. Temporal classification of natural gesture and application to videocoding. *Conference on Computer Vision and Pattern Recognition, 1997*, 1997. 49

[273] L. Wiskott and T. Sejnowski. Slow feature analysis: Unsupervised learning of invariances. *Neural computation*, 14(4):715–770, 2002. 197

[274] M. Wöllmer, F. Eyben, S. Reiter, B. Schuller, C. Cox, E. Douglas-Cowie, and R. Cowie. Abandoning emotion classes-towards continuous emotion recognition with modelling of long-range dependencies. In *INTERSPEECH*, pages 597–600, 2008. 58

[275] M. Wollmer, B. Schuller, F. Eyben, and G. Rigoll. Combining long short-term memory and dynamic bayesian networks for incremental emotion-sensitive artificial listening. *IEEE Journal of Selected Topics in Signal Processing*, 4(5):867 –881, Oct. 2010. 40, 70

[276] D. H. Wolpert. Stacked generalization. *Neural Networks*, 5:241–259, 1992. 111

[277] J. Xiao, T. Kanade, and J. Cohn. Robust full-motion recovery of head by dynamic templates andre-registration techniques. *Fifth IEEE International Conference on Automatic Face and Gesture Recognition*, 2002. 53

[278] X. Xiong and F. De la Torre. Supervised descent method and its applications to face alignment. In *Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on*, pages 532–539. IEEE, 2013. 50

[279] D. Xu, S. Yan, D. Tao, S. Lin, and H.-J. Zhang. Marginal fisher analysis and its variants for human gait recognition and content- based image retrieval. *Trans. Img. Proc.*, 16(11):2811–2821, Nov. 2007. 209

[280] S. Yan, D. Xu, B. Zhang, H.-J. Zhang, Q. Yang, and S. Lin. Graph embedding and extensions: A general framework for dimensionality reduction. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 29(1):40–51, 2007. 195

[281] Y.-H. Yang, Y.-C. Lin, Y.-F. Su, and H. H. Chen. Music emotion classification: A regression approach. In *Proc. of IEEE Int. Conf. on Multimedia and Expo*, pages 208–211, 2007. 95

[282] A. Yilmaz, O. Javed, and M. Shah. Object tracking: A survey. *ACM Computing Surveys (CSUR)*, 2006. 53

[283] C. Yu, P. M. Aoki, and A. Woodruff. Detecting user engagement in everyday conversations. In *Proc. of 8th Int. Conf. on Spoken Language Processing*, 2004. 40

[284] S. Yu, K. Yu, V. Tresp, H.-P. Kriegel, and M. Wu. Supervised probabilistic principal component analysis. In *Proc. of the 12th ACM SIGKDD Int. Conf. on Knowledge discovery and data mining*, KDD '06, pages 464–473, 2006. 74, 157

[285] Z. Zeng, M. Pantic, G. Roisman, and T. Huang. A survey of affect recognition methods: Audio, visual, and spontaneous expressions. *IEEE Tran. on Pattern Analysis and Machine Intelligence*, 31:39–58, 2009. 38, 43, 46, 47, 51, 54, 90, 104, 126, 149

[286] C. Zhang and Z. Zhang. A survey of recent advances in face detection. Technical report, Tech. rep., Microsoft Research, 2010. 50

[287] J. Zhang. The mean field theory in EM procedures for Markov random fields. *IEEE Transactions on Signal Processing*, 40(10):2570–2583, 1992. 203

[288] S. Zhang and T. Sim. Discriminant subspace analysis: A Fukunaga-Koontz approach. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 29(10):1732–1745, 2007. 214, 215, 216, 217

[289] Y. Zhang and Q. Ji. Active and dynamic information fusion for facial expression understanding from image sequences. *IEEE Trans. on pattern analysis and machine intelligence*, 2005. 52

[290] Y. Zhang and D.-Y. Yeung. Heteroscedastic probabilistic linear discriminant analysis with semi-supervised extension. In *Proceedings of the ECML-PKDD: Part II*, ECML PKDD '09, pages 602–616, Berlin, Heidelberg, 2009. Springer-Verlag. 195, 196

[291] Z. Zhang. Microsoft kinect sensor and its effect. *MultiMedia, IEEE*, 19(2):4–10, 2012. 54

[292] Z. Zhang, M. Lyons, M. Schuster, and S. Akamatsu. Comparison between geometry-based and gabor-wavelets-based facial expression recognition using multi-layer perceptron. In *Automatic Face and Gesture Recognition, 1998. Proceedings. Third IEEE International Conference on*, pages 454–459. IEEE, 1998. 51

[293] G. Zhao and M. Pietikainen. Dynamic texture recognition using local binary patterns with an application to facial expressions. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 29(6):915–928, 2007. 51

[294] W. Zhao, R. Chellappa, P. J. Phillips, and A. Rosenfeld. Face recognition: A literature survey. *Acm Computing Surveys (CSUR)*, 35(4):399–458, 2003. 50

[295] W. Zheng, H. Tang, Z. Lin, and T. S. Huang. Emotion recognition from arbitrary view facial images. In *Computer Vision–ECCV 2010*, pages 490–503. Springer, 2010. 51

[296] F. Zhou and F. De la Torre. Generalized time warping for multi-modal alignment of human motion. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2012. 75, 77, 154, 162, 174, 175, 182, 183

[297] F. Zhou, F. De la Torre, and J. K. Hodgins. Aligned cluster analysis for temporal segmentation of human motion. In *Proc. 2008 IEEE Int. Conf. Automatic Face and Gestures Recognition*, 2008. 75, 174

[298] F. Zhou and F. D. la Torre. Canonical time warping for alignment of human behavior. In *Advances in Neural Information Processing Systems 22*, pages 2286–2294, 2009. 75, 77, 146, 152, 153, 162, 174, 175, 182, 183

[299] G. Zhou, A. Cichocki, and S. Xie. Common and individual features analysis: beyond canonical correlation analysis. *arXiv preprint arXiv:1212.3913*, 2012. 182

[300] M. Zhu and A. M. Martinez. Subclass discriminant analysis. *IEEE Trans. Pattern Anal. Mach. Intell.*, 28(8):1274–1286, Aug. 2006. 209

[301] X. Zhu and D. Ramanan. Face detection, pose estimation, and landmark localization in the wild. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, pages 2879–2886. IEEE, 2012. 50

Appendix A

Unified Framework for Probabilistic Component Analysis

In what follows, we detail a set of derivations regarding the proposed Unified Framework on Probabilistic Component Analysis (Chapter 11). In more detail, we go through the full derivation of EM-PCA, as well as present an extension of the family of models presented in Chapter 11 to mixtures of component analysers.

A.1 EM for PCA

Firstly, we define a fully connected prior on the latent variables:

$$P(\mathbf{Y}|\beta) = \frac{1}{Z}\exp\left\{-\sum_{n=1}^{N}\frac{1}{2\sigma_n^2}\sum_{i=1}^{T}\frac{1}{T-1}\sum_{j=1,j\neq i}^{T}(y_{n,i}-\lambda_n y_{n,j})^2\right\} \tag{A.1}$$

where $\beta = \{\sigma_{1:N}, \lambda_{1:N}\}$. By expanding the normalising integral we obtain[1]

$$P(\mathbf{y}_i|\mathbb{E}[y_j]_{j\neq i},\beta) = \frac{\exp\left\{-\frac{1}{2}\sum_{n=1}^{N}\frac{1}{\sigma_n^2}\sum_{i=1}^{T}\frac{1}{T-1}\sum_{j=1,j\neq i}^{T}(y_{n,i}^2 - 2\lambda_n y_{n,i}\mathbb{E}[y_{n,j}] + \lambda_n^2\mathbb{E}[y_{n,j}^2])\right\}}{\int_{y_i}\exp\left\{-\frac{1}{2}\sum_{n=1}^{N}\frac{1}{\sigma_n^2}\frac{1}{T-1}\sum_{j=1,j\neq i}^{T}(y_{n,i}^2 - 2\lambda_n y_{n,i}\mathbb{E}[y_{n,j}] + \lambda_n^2\mathbb{E}[y_{n,j}^2])\right\}dy_i} \tag{A.2}$$

[1] In Chapter 11 we use the general $P(\mathbf{y}_i|\mathbf{m}_i^{(R)},\beta)$ for $P(\mathbf{y}_i|\mathbb{E}[y_j]_{j\neq i},\beta)$, where here $\mathbf{m}_i^{(R)} = \mathbf{m}_i^{(\text{PCA})}$

where the internal part of the exponent becomes:

$$\begin{aligned} & -\frac{1}{2}\sum_{n=1}^{N}\frac{1}{\sigma_n^2}\sum_{i=1}^{T}\frac{1}{T-1}\sum_{j=1,j\neq i}^{T}(y_{n,i}^2 - 2\lambda_n y_{n,i}\mathbb{E}[y_{n,j}] + \lambda_n^2\mathbb{E}[y_{n,j}^2]) \\ = & -\frac{1}{2}\sum_{n=1}^{N}\frac{1}{\sigma_n^2}\sum_{i=1}^{T}\left(y_{n,i}^2 - 2y_{n,i}\left[\lambda_n\frac{1}{T-1}\sum_{j=1,j\neq i}^{T}\mathbb{E}[y_{n,j}]\right] + \left[\lambda_n^2\frac{1}{T-1}\mathbb{E}[y_{n,j}^2]\right]\right) \\ = & -\frac{1}{2}\sum_{n=1}^{N}\frac{1}{\sigma_n^2}\sum_{i=1}^{T}\left(y_{n,i} - \lambda_n\frac{1}{T-1}\sum_{j\neq i}\mathbb{E}[y_{n,j}]\right)^2 + c \end{aligned} \tag{A.3}$$

hence,

$$P(\mathbf{y}_i|\mathbb{E}[\mathbf{y}_j]_{j\neq i},\beta) \sim \mathcal{N}(\mathbf{y}_i|\mathbf{m}_i,\mathbf{\Sigma}) \tag{A.4}$$

where $\mathbf{m}_i = \mathbf{\Lambda}\frac{1}{T-1}\sum_{j\neq i}^{T}\mathbb{E}[\mathbf{y}_j]$, $\mathbf{\Sigma} = diag(\sigma_1^2,\ldots,\sigma_N^2)$, and $\mathbf{\Lambda} = diag(\lambda_1,\ldots,\lambda_n)$.

Therefore:

$$P(\mathbf{y}_i|\mathbb{E}[y_j]_{j\neq i},\Psi) = \frac{P(\mathbf{x}_i|\mathbf{y}_i,\theta)P(\mathbf{y}_i|\mathbb{E}[y_j]_{j\neq i},\beta)}{\int_{\mathbf{y}_i}P(\mathbf{x}_i|\mathbf{y}_i,\theta)P(\mathbf{y}_i|\mathbb{E}[\mathbf{y}_j]_{j\neq i},\beta)d\mathbf{y}_i} \tag{A.5}$$

where $\theta = \{\mathbf{W},\sigma_x\}$ and $\Psi = \{\theta,\beta\}$. The observed probability is defined as

$$P(\mathbf{x}_i|\mathbf{y}_i,\theta) = \mathcal{N}(\mathbf{x}_i|\mathbf{W}\mathbf{y}_i,\sigma_x^2\mathbf{I}) \tag{A.6}$$

and the mean of the posterior is found as

$$\mathbb{E}[\mathbf{y}_i] = \frac{\int_{\mathbf{y}_i}\mathbf{y}_i P(\mathbf{x}_i|\mathbf{y}_i,\theta)P(\mathbf{y}_i|\mathbb{E}[\mathbf{y}_j]_{j\neq i},\beta)d\mathbf{y}_i}{\int_{\mathbf{y}_i}P(\mathbf{x}_i|\mathbf{y}_i,\theta)P(\mathbf{y}_i|\mathbb{E}[\mathbf{y}_j]_{j\neq i},\beta)d\mathbf{y}_i} \tag{A.7}$$

where by considering Eq. A.4 and A.6

$$\mathbb{E}[\mathbf{y}_i] = \mathbb{E}\left[\mathcal{N}(\mathbf{x}_i|\mathbf{W}\mathbf{y}_i,\sigma_x^2\mathbf{I})\mathcal{N}(\mathbf{y}_i|\mathbf{m}_i,\mathbf{\Sigma})\right] \tag{A.8}$$

where we have a product of Gaussians whose expected value (mean) we are interested in. By completing the square for $\mathbf{y}_i$:

$$\mathbb{E}[\mathbf{y}_i] = \left(\mathbf{W}^T\mathbf{W} + \hat{\mathbf{\Sigma}}^{-1}\right)^{-1}\left(\mathbf{W}^T\mathbf{x}_i + \hat{\mathbf{\Sigma}}^{-1}\mathbf{m}_i\right) \tag{A.9}$$

where now $\hat{\mathbf{\Sigma}}_{mn} = \delta_{mn}\left[\frac{\Sigma_{mn}}{\sigma_x^2}\right]$. Similarly for the variance,

$$\mathbb{E}[\mathbf{y}_i\mathbf{y}_i^T] = \sigma_x^2\left(\mathbf{W}^T\mathbf{W} + \hat{\mathbf{\Sigma}}^{-1}\right)^{-1} + \mathbb{E}[\mathbf{y}_i]\mathbb{E}[\mathbf{y}_i]^T \tag{A.10}$$

Having recovered the first order moments, we move on to the maximisation step. By denoting Ψ as the complete set of parameters, we optimise:

$$\begin{aligned}
\theta = \{\mathbf{W}, \sigma_x\} &= argmax \sum_{i=1}^{T}\int_{\mathbf{y}_i} P(\mathbf{y}_i|\mathbb{E}[y_j]_{j\neq i}, \mathbf{x}_i, \Psi) ln P(\mathbf{x}_i|\mathbf{y}_i, \theta) d\mathbf{y}_i \\
&= \sum_{i=1}^{T}\int_{\mathbf{y}_i} ln\frac{1}{(2\pi)^{\frac{F}{2}}\sigma_x^F} P(\mathbf{y}_i|\mathbb{E}[\mathbf{y}_j]_{j\neq i}, \mathbf{x}_i, \Psi) d\mathbf{y}_i \\
&+ \sum_{i=1}^{T}\int_{\mathbf{y}_i}\left[-\frac{1}{2\sigma_x^2}(\mathbf{x}_i - \mathbf{W}\mathbf{y}_i)^T(\mathbf{x}_i - \mathbf{W}\mathbf{y}_i)\right] P(\mathbf{y}_i|\mathbb{E}[\mathbf{y}_j]_{j\neq i}, \mathbf{x}_i, \Psi) d\mathbf{y}_i \\
&= Tln\frac{1}{2\pi^{\frac{F}{2}}\sigma_x^F} - \frac{1}{2\sigma_x^2}\sum_{i=1}^{T}\left(\mathrm{Tr}(\mathbf{x}_i^T\mathbf{x}_i) - 2(\mathbf{W}^{-T}x_i)^T\mathbb{E}[\mathbf{y}_i] + \mathrm{Tr}[\mathbb{E}[\mathbf{y}_i\mathbf{y}_i^T]\mathbf{W}^T W]\right)
\end{aligned} \tag{A.11}$$

Subsequently, we maximise the log-likelihood wrt the parameters, recovering the update equations:

$$\frac{\partial L(\mathbf{W}, \sigma_x)}{\partial \mathbf{W}} = 0 \;\Rightarrow\; -\frac{1}{2\sigma_x^2}\sum_{i=1}^{T}(-2\mathbf{x}_i\mathbb{E}[\mathbf{y}_i]^T + 2\mathbf{W}\mathbb{E}[\mathbf{y}_i\mathbf{y}_i^T]) = 0 \tag{A.12}$$

$$\Rightarrow\; \mathbf{W} = \left(\sum_{i=1}^{T}\mathbf{x}_i\mathbb{E}[\mathbf{y}_i]^T\right)\left(\sum_{i=1}^{T}\mathbb{E}[\mathbf{y}_i\mathbf{y}_i^T]\right)^{-1} \tag{A.13}$$

$$\frac{\partial L(\mathbf{W}, \sigma_x)}{\partial \sigma_x} = 0 \;\Rightarrow\; \sigma_x^2 = \frac{1}{FT}\sum_{i=1}^{T}\left\{\mathrm{Tr}[\mathbf{x}_i\mathbf{x}_i^T] - 2\mathbb{E}[\mathbf{y}_i]^T\mathbf{W}^T\mathbf{x}_i + \mathrm{Tr}[\mathbb{E}[\mathbf{y}_i\mathbf{y}_i^T]\mathbf{W}^T\mathbf{W}]\right\} \tag{A.14}$$

When maximising the σ_n and λ_n parameters, the maximisation step is as follows:

$$\beta = \{\sigma_n, \lambda_n\} \quad = \quad argmax \sum_{i=1}^{T} \int_{\mathbf{y}_i} P(\mathbf{y}_i|\mathbb{E}[\mathbf{y}_j]_{j\neq i}, \mathbf{x}_i, \Psi) ln P(\mathbf{y}_i|\mathbb{E}[\mathbf{y}_j]_{j\neq i}, \beta) d\mathbf{y}_i \tag{A.15}$$

where

$$P(\mathbf{y}_i|\mathbb{E}[\mathbf{y}_j]_{j\neq i}, \beta) = \frac{1}{(2\pi)^{\frac{N}{2}}|\mathbf{\Sigma}|^{\frac{1}{2}}} \exp\left\{-\frac{1}{2}(\mathbf{y}_i - \mathbf{m}_i)^T \mathbf{\Sigma}^{-1}(\mathbf{y}_i - \mathbf{m}_i)\right\} \tag{A.16}$$

$$= \frac{1}{(2\pi)^{\frac{N}{2}} \prod_{n=1}^{N} \sigma_n} \exp\left\{-\frac{1}{2}(\mathbf{y}_i - \mathbf{m}_i)^T \mathbf{\Sigma}^{-1}(\mathbf{y}_i - \mathbf{m}_i)\right\} \tag{A.17}$$

$$ln P(\mathbf{y}_i|\mathbb{E}[\mathbf{y}_j]_{j\neq i}, \beta) = ln \frac{1}{(2\pi)^{\frac{N}{2}} \prod_{n=1}^{N} \sigma_n} - \frac{1}{2}(\mathbf{y}_i - \mathbf{m}_i)^T \mathbf{\Sigma}^{-1}(\mathbf{y}_i - \mathbf{m}_i) \tag{A.18}$$

therefore,

$$\begin{aligned} &= \sum_{i=1}^{T} \int_{\mathbf{y}_i} P(\mathbf{y}_i|\mathbb{E}[\mathbf{y}_j]_{j\neq i}, \mathbf{x}_i, \Psi)(ln \frac{1}{(2\pi)^{\frac{N}{2}}} - \sum_{n=1}^{N} ln\sigma_n - \frac{1}{2}(\mathbf{y}_i - \mathbf{m}_i)^T \mathbf{\Sigma}^{-1}(\mathbf{y}_i - \mathbf{m}_i)) d\mathbf{y}_i \\ &= T ln \frac{(T-1)^{\frac{N}{2}}}{(2\pi)^{\frac{N}{2}}} - T \sum_{n=1}^{N} ln\sigma_n - \frac{1}{2} \sum_{i=1}^{T} \Bigg[\int_{\mathbf{y}_i} \mathbf{y}_i^T \mathbf{\Sigma}^{-1} \mathbf{y}_i P(\mathbf{y}_i|\mathbb{E}[\mathbf{y}_j]_{j\neq i}, \mathbf{x}_i, \Psi) - 2(\mathbf{m}_i)^T \mathbf{\Sigma}^{-1} \\ &\quad \underbrace{\int_{\mathbf{y}_i} \mathbf{y}_i P(\mathbf{y}_i|\mathbb{E}[\mathbf{y}_j]_{j\neq i}, x_i, \Psi) d_{\mathbf{y}_i}}_{\mathbb{E}_{\mathbf{y}_i}} + (\mathbf{m}_i)^T \mathbf{\Sigma}^{-1} \mathbf{m}_i \Bigg] \end{aligned} \tag{A.19}$$

where

$$\int_{\mathbf{y}_i} \mathbf{y}_i^T \mathbf{\Sigma}^{-1} \mathbf{y}_i P(\mathbf{y}_i|\mathbb{E}[\mathbf{y}_j]_{j\neq i}, \mathbf{x}_i, \Psi) d_{\mathbf{y}_i} = \mathrm{Tr}[\mathbb{E}[\mathbf{y}_i \mathbf{y}_i^T] \mathbf{\Sigma}^{-1}] \tag{A.20}$$

the log-likelihood now becomes:

$$= Tln\frac{1}{(2\pi)^{N/2}} - T\sum_{n=1}^{N} ln\sigma_n - \frac{1}{2}\sum_{i=1}^{T}\Big[\mathrm{Tr}[\mathbb{E}[\mathbf{y}_i\mathbf{y}_i^T]\boldsymbol{\Sigma}^{-1} \tag{A.21}$$

$$-2\mathbf{m}_i^T\boldsymbol{\Sigma}^{-1}\mathbb{E}[\mathbf{y}_i] + \mathbf{m}_i^T\boldsymbol{\Sigma}^{-1}\mathbf{m}_i\Big] \tag{A.22}$$

$$= C - T\sum_{n=1}^{N}\ln\sigma_n - \frac{1}{2}\sum_{i=1}^{T}\sum_{n=1}^{N}\left(\frac{1}{\sigma_n^2}\mathbb{E}[y_{n,i}^2] - 2\frac{1}{\sigma_n^2}\mathbb{E}[y_{n,i}]m_{n,i} + \frac{1}{\sigma_n^2}m_{n,i}^2\right) \tag{A.23}$$

and finally, by taking the derivatives:

$$\frac{\partial L}{\partial\sigma_n} = 0 \quad \Rightarrow \quad -T\frac{1}{\sigma_n} + \sum_{i=1}^{T}\frac{1}{\sigma_n^3}(\mathbb{E}[y_{n,i}^2] - 2\mathbb{E}[y_{n,i}]m_{n,i} + m_{n,i}^2) = 0 \tag{A.24}$$

$$\Rightarrow \quad \sigma_n^2 = \frac{1}{T}\sum_{i=1}^{T}(\mathbb{E}[y_{n,i}^2] - 2\mathbb{E}[y_{n,i}]m_{n,i} + m_{n,i}^2) \tag{A.25}$$

A.2 Mixtures of Component Analysers

In several applications, fitting a single Gaussian to the data is unrealistic and proves to be suboptimal. In this section, we formulate a mixture model for our unified framework for probabilistic component analysis, in effect providing mixture models for PCA, LDA, LPP and SFA. Note that in this section, we drop the dependence on the MRF model in order to avoid cluttering the notation. Assuming the linear model utilised in Chapter 11, the joint-data likelihood for a mixture of M component analysers with T the number of data samples can be defined as

$$P(\mathbf{x}, \mathbf{y}_1, \ldots, \mathbf{y}_M) = \prod_{i=1}^{T}\left[\prod_{m=1}^{M}\pi_m^{z_{im}}P(\mathbf{x}_i|\mathbf{y}_i^m)P(\mathbf{y}_i^m)\right] \tag{A.26}$$

where π_i is the corresponding mixing proportion with $\pi_i \geq 0$ and $\sum_i^M \pi_j = 1$, while z_{ij} is a binary vector labelling which mixture model is responsible for each data point i. The

log-likelihood is defined as

$$\mathcal{L}_c = \sum_{i=1}^{T}\sum_{m=1}^{M} \mathbf{z}_{im} ln\left[\pi_m P(\mathbf{x}_i|\mathbf{y}_i^m,\theta)P(\mathbf{y}_i^m|\mathbf{m}_i^m,\beta)\right] \tag{A.27}$$

$$= \sum_{i=1}^{T}\sum_{m=1}^{M} \mathbf{z}_{im} ln\pi_m + \sum_{i=1}^{T}\sum_{m=1}^{M} \mathbf{z}_{im} lnP(\mathbf{x}_i|\mathbf{y}_i^m,\theta)$$

$$+ \sum_{i=1}^{T}\sum_{m=1}^{M} \mathbf{z}_{im} lnP(\mathbf{y}_i^m|\mathbf{m}_i^m,\beta) \tag{A.28}$$

In the expectation step, we recover the first order moments for each mixture component m as

$$\mathbb{E}[\mathbf{y}_i^m] = \int_{\mathbf{y}_i^m} \mathbf{y}_i^m P(\mathbf{y}_i^m|\mathbf{x}_i,\mathbf{m}_i^m,\Psi)d\mathbf{y}_i^m \tag{A.29}$$

$$= \left[\mathbf{W}_m^T\mathbf{W}_m + \left(\hat{\mathbf{\Sigma}}_m\right)^{-1}\right]^{-1}\left[\mathbf{W}_m^T\mathbf{x}_i + \left(\hat{\mathbf{\Sigma}}_m\right)^{-1}\mathbf{m}_i^m\right] \tag{A.30}$$

$$\mathbb{E}[\mathbf{y}_i^m(\mathbf{y}_i^m)^T] = (\sigma_x^m)^2\left[\mathbf{W}_m^T\mathbf{W}_m + \left(\hat{\mathbf{\Sigma}}_m\right)^{-1}\right]^{-1} + \mathbb{E}[\mathbf{y}_i^m]\mathbb{E}[\mathbf{y}_i^m]^T. \tag{A.31}$$

while the expected value of $\mathbf{z}_{im}$, which represents the responsibility of mixture m for data point x_i is found as

$$\mathbb{E}[\mathbf{z}_{im}] = \frac{\pi_m \int_{\mathbf{y}_i^m} P(\mathbf{x}_i|\mathbf{y}_i^m,\theta^m)P(\mathbf{y}_i^m|\mathbf{m}_i^m,\beta)d\mathbf{y}_i^m}{\sum_n^M \pi_n \int_{\mathbf{y}_i^n} P(\mathbf{x}_i|\mathbf{y}_i^n,\theta^m)P(\mathbf{y}_i^n|\mathbf{m}_i^n,\beta)d\mathbf{y}_i^n}$$

$$= \frac{\pi_m P(\mathbf{x}_i|\Psi^m)}{\sum_n^M \pi_n P(\mathbf{x}_i|\Psi^n)}$$

Where $P(\mathbf{x}_i|\Psi^m) = \mathcal{N}(\mathbf{B}_m\mathbf{W}_m\mathbf{A}\hat{\mathbf{\Sigma}}_{\mathbf{m}}^{-1}\mathbf{m}_i, \sigma_x^2\mathbf{B}_m)$, with $\mathbf{A} = (\mathbf{W}_m^T\mathbf{W}_m + \hat{\mathbf{\Sigma}}^{-1})^{-1}$ and $\mathbf{B}_m = (\mathbf{I} - \mathbf{W}_m\mathbf{A}\mathbf{W}_m^T)^{-1}$. Taking the expectation of the likelihood with respect to posterior distributions and obtaining the derivative for each parameter, we obtain the maximisation step updates. Note that the constraint $\sum_i \pi_i = 1$ needs to be incorporated, and can be achieved using a Lagrange multiplier, similarly to [247]. This leads to the update

$$\pi_m = \frac{1}{T}\sum_{i=1}^{T}\mathbb{E}[\mathbf{z}_{im}]. \tag{A.32}$$

The weights for each mixture model m are updated as

$$\mathbf{W}_m = \left[\sum_{i=1}^{T} \mathbb{E}[\mathbf{z}_{im}]\mathbb{E}[\mathbf{y}_i^m]^T\right]\left[\sum_{i=1}^{T} \mathbb{E}[\mathbf{z}_{im}]\mathbb{E}[\mathbf{y}_i^m\mathbf{y}_i^m]^T]\right]^{-1} \tag{A.33}$$

while the variances are found as

$$\begin{aligned}(\sigma_{\mathbf{x}}{}^m)^2 &= \tfrac{1}{FT}\sum_{i=1}^{T}\mathbb{E}[\mathbf{z}_{im}]\Big\{||\mathbf{x}_i||^2 - 2\sum_{i=1}^{T}\mathbb{E}[\mathbf{y}_i^m]^T\mathbf{W}_m^T\mathbf{x}_i + Tr(\mathbb{E}[\mathbf{y}_i^m(\mathbf{y}_i^m)^T]\mathbf{W}_m{}^T\mathbf{W}_m)\Big\}\\ (\sigma_n)^2 &= \tfrac{F(\lambda_n)}{T}\sum_{m=1}^{M}\sum_{i=1}^{T}\Big(\mathbb{E}[(y_{n,i}^2)^m] - 2\mathbb{E}[y_{n,i}^m]m_{n,i} + m_{n,i}^2\Big)\end{aligned} \tag{A.34}$$