Robust Correlated and Individual Component Analysis

Yannis Panagakis, *Member, IEEE,* Mihalis A. Nicolaou, *Member, IEEE,*
Stefanos Zafeiriou, *Member, IEEE,* and Maja Pantic, *Fellow, IEEE*

Abstract—Recovering correlated and individual components of two, possibly temporally misaligned, sets of data is a fundamental task in disciplines such as image, vision, and behavior computing, with application to problems such as multi-modal fusion (via correlated components), predictive analysis, and clustering (via the individual ones). Here, we study the extraction of correlated and individual components under real-world conditions, namely i) the presence of gross non-Gaussian noise and ii) temporally misaligned data. In this light, we propose a method for the Robust Correlated and Individual Component Analysis (RCICA) of two sets of data in the presence of gross, sparse errors. We furthermore extend RCICA in order to handle temporal incongruities arising in the data. To this end, two suitable optimization problems are solved. The generality of the proposed methods is demonstrated by applying them onto 4 applications, namely i) heterogeneous face recognition, ii) multi-modal feature fusion for human behavior analysis (i.e., audio-visual prediction of interest and conflict), iii) face clustering, and iv) the temporal alignment of facial expressions. Experimental results on 2 synthetic and 7 real world datasets indicate the robustness and effectiveness of the proposed methods on these application domains, outperforming other state-of-the-art methods in the field.

Index Terms—Multi-modal analysis, Canonical correlation analysis, Individual components, Time warping, Low-rank, Sparsity.

1 Introduction

The analysis of two sets of high-dimensional data arising from different modalities and distinct feature sets is inherent to many tasks and applications pertaining to image, vision, and behaviour computing, among other disciplines. For instance, an image may be represented via a variety of visual descriptors such as SIFTs, HoGs, IGOs [1], [2], [3] etc., which can be seen as distinct feature sets corresponding to the same object. Another prominent example of such a scenario lies in the task of face recognition: a face can be recognized by employing the normal image as captured in the visible spectrum, as well as infrared captures or even forensic sketches [4], [5]. Similarly, a particular human behaviour can be identified by certain vocal, gestural, and facial features extracted from both the audio and visual modalities [6], [7].

Since such sets of multimodal data compromising of distinct feature sets refer to the same object or behaviour, it is anticipated that part of the conveyed information is shared amongst all observation sets (i.e., correlated components), while the remaining information consists of individual information (individual components) which are particular only to a specific observation set. The correlation amongst the low-level features extracted from two different modalities provide useful information for tasks such as feature fusion [8], [9], multiview learning [10], multi-label prediction [11], and multimodal behaviour analysis [6], [7], [12]. On the other hand, the individual components are deemed important for tasks such as clustering and signal separation [13]. These individual features may interfere with finding the correlated components, just as the correlated components are likely to obscure the individual ones. Consequently, it is very important to *simultaneously* and *accurately* extract the *correlated* and the *individual components* among two datasets.

The problem becomes rather challenging when dealing with data contaminated by *gross errors*, which are also *temporally misaligned*, i.e., temporal discrepancies manifest amongst the observation sequences. In practice, gross errors [14] arise from either device artifacts (e.g., pixel corruptions, sonic artifacts), missing and incomplete data (e.g., partial image texture occlusions), or feature extraction failure (e.g., incorrect object localization, tracking errors). These errors *rarely* follow a Gaussian distribution [15]. Furthermore, asynchronous sensor measurements (e.g., lag between audio and visual sensors), view point changes, network lags, speech rate differences, and the speed of an action, behaviour, or event result into temporally misaligned sets of data. Clearly, the accurate temporal alignment of noisy, temporally misaligned sets of data is a cornerstone in many computer vision [16], [17], behaviour analysis [18], [12], and speech processing [19] problems, to name but a few.

Several methods have been proposed for the analysis of two sets of data. A subset of them is briefly described in Section 2. The Canonical Correlation Analysis (CCA) [20] is a widely used method for finding linear correlated

- *This work has been funded by the European Community Horizon 2020 [H2020/2014-2020] under grant agreement no. 645094 (SEWA). Yannis Panagakis was also partially funded by the ERC under the FP7 Marie Curie Intra-European Fellowship. The work of Mihalis Nicolaou was also funded in part by the European Community 7th Framework Programme [FP7/2007-2013] under grant agreement no. 611153 (TERESA). Stefanos Zafeiriou was partially supported by the EPSRC project EP/J017787/1 (4D-FAB).*

Y. Panagakis, S. Zafeiriou, and M. Pantic are with the Dept. of Computing, Imperial College London, UK. M. Nicolaou was with the with the Dept. of Computing, Imperial College London, UK. He is now with the Dept. of Computing at Goldsmiths, University of London, UK. M. Pantic is also with the EEMCS, University of Twente, the Netherlands.
Corresponding author: Y. Panagakis, email: i.panagakis@imperial.ac.uk.

components among two data sets. Notable extensions of the CCA are the sparse CCA [11], [21], the kernel- [22] and deep-CCA [23], as well as its probabilistic [24], [12] and Bayesian variants [25]. The Canonical Time Warping (CTW) [17], extents the CCA to handle time warping in data. In order to extract correlated components among multiple data sets, generalizations of the CCA can be employed [26], [4]. However, the aforementioned methods ignore the individual components of the data sets; a drawback which alleviated by the Joint and Individual Variation Explained (JIVE) [27] and Common Orthogonal Basis Extraction (COBE) [13]. Since most of the methods mentioned above rely on least squares error minimization, they are prone to gross errors and outliers [14], making the estimated components to be arbitrarily away from the true ones. This drawback is alleviated to some extend by the robust methods in [18], [28], which are the preliminary works of this paper.

Here, distinct from the previous methods, the Robust Correlated and Individual Component Analysis (RCICA) is proposed, enabling the recovery of the correlated and individual components of two (possibly temporally misaligned) data sets in the presence of gross (but sparse) errors. The contributions of the paper are organized as follows.

1) Inspired by recent advances in learning using low-rank and sparse models e.g., [15], [29], [30], [31], we propose a general framework for the robust recovery of correlated and individual components. In particular, the RCICA decomposes each dataset into a sum of three terms: a *low-rank* matrix capturing the correlated components, a *low-rank* matrix accounting for individual ones, and a *sparse* term modelling the gross errors. To this end, a suitable model, involving the minimization of weighted sums of nuclear- and ℓ_1-norms is proposed in Section 3. The Robust CCA (RCCA) [28] which recovers the reconstruction of the correlated components, and the sparse corruptions but ignores the individual components is a special case of the RCICA, as shown in this paper.
2) The RCICA is extended to handle temporally misaligned, noisy data in Section 4. To achieve this, the Dynamic Time Warping (DTW) [19] is incorporated into the RICA, allowing the temporal alignment of the data sets onto the subspace spanned by the robustly estimated correlated components. By ignoring the individual components the RCICA with time warping capabilities is reduced to the Robust Canonical Time Warping (RCTW) [18].
3) Two efficient algorithms for the RCICA and its extension are developed based on the Alternating Direction Method of Multipliers [32], and presented in Sections 3 and 4, respectively.

To demonstrate the generality of the proposed models and their algorithmic framework, in Section 5 experiments are performed on four application domains, namely i) heterogeneous face recognition, where images are obtained via multiple sensors, ii) multimodal fusion for human behaviour analysis (i.e., predictive analysis of the level of interest and conflict from audio-visual cues), iii) face clustering, and iv) the temporal alignment of actions units. Experimental results on 2 synthetic and 7 real world datasets, contaminated by non-Gaussian gross errors, indicate the robustness and effectiveness of the proposed methods on these application domains, outperforming compared methods. Conclusions are drawn in Section 6. Finally, we note that technical details are deferred to the supplementary material.

Notations. Throughout the paper, matrices (vectors) are denoted by uppercase (lowercase) boldface letters e.g., $\mathbf{X}, \mathbf{Y}$, $(\mathbf{x}, \mathbf{y})$. $\mathbf{I}$ $(\mathbf{1})$ denotes the identity matrix (vector of ones) of compatible dimensions. $\mathbf{0}$ is the zero matrix. The ith column of $\mathbf{X}$ is denoted as $\mathbf{x}_i$. The set of real numbers is denoted by $\mathbb{R}$. A set of N real matrices of varying dimensions is denoted by $\{\mathbf{X}^{(n)} \in \mathbb{R}^{I_n \times J_n}\}_{n=1}^{N}$. Regarding matrix norms, $\|\mathbf{X}\|_*$ denotes the nuclear norm and it is defined as the sum of its singular values; the matrix ℓ_1-norm is denoted by $\|\mathbf{X}\|_1 \doteq \sum_i \sum_j |x_{ij}|$, $\|\mathbf{X}\|$ is the spectral norm, and $\|\mathbf{X}\|_F \doteq \sqrt{\sum_i \sum_j x_{ij}^2} = \sqrt{\mathrm{tr}(\mathbf{X}^T\mathbf{X})}$ is the Frobenius norm, where $\mathrm{tr}(\cdot)$ denotes the trace of a square matrix.

2 Backround

To make the paper self-contained, this section includes a brief review of the CCA [20], the JIVE [27], the DTW [19], and the CTW [17].

2.1 Canonical Correlation Analysis

The CCA extracts correlated features from a pair of multivariate data. In particular, given two data sets $\{\mathbf{X}^{(n)} = [\mathbf{x}_1^{(n)}|\mathbf{x}_2^{(n)}|\ldots|\mathbf{x}_J^{(n)}] \in \mathbb{R}^{I_n \times J}\}_{n=1}^{2}$, the CCA finds two matrices $\mathbf{V}^{(1)} \in \mathbb{R}^{I_1 \times K}$ and $\mathbf{V}^{(2)} \in \mathbb{R}^{I_2 \times K}$, with $K \leq \min(I_1, I_2)$. These matrices define a common, low-dimensional latent subspace such that the linear combination of the variables in $\mathbf{X}^{(1)}$, i.e., ${\mathbf{V}^{(1)}}^T\mathbf{X}^{(1)}$ are highly correlated with a linear combination of the variables in $\mathbf{X}^{(2)}$, i.e., ${\mathbf{V}^{(2)}}^T\mathbf{X}^{(2)}$. The CCA corresponds to the solution of the constrained least-squares minimization problem [11], [33]:

$$\begin{aligned} &\underset{\{\mathbf{V}^{(n)}\}_{n=1}^{2}}{\operatorname{argmin}} \ \frac{1}{2}\|{\mathbf{V}^{(1)}}^T\mathbf{X}^{(1)} - {\mathbf{V}^{(2)}}^T\mathbf{X}^{(2)}\|_F^2 \\ &\text{s.t.} \ \ {\mathbf{V}^{(n)}}^T\mathbf{X}^{(n)}{\mathbf{X}^{(n)}}^T\mathbf{V}^{(n)} = \mathbf{I}, n = 1, 2. \end{aligned} \tag{1}$$

2.2 Joint and Individual Variation Explained

The JIVE recovers the joint and individual components among $N \geq 2$ data sets $\{\mathbf{X}^{(n)} \in \mathbb{R}^{I_n \times J}, n = 1, 2, \ldots, N\}$. In particular, each matrix is decomposed into three terms: a low-rank matrix $\mathbf{J}^{(n)} \in \mathbb{R}^{I_n \times J}$ capturing joint structure between data sets, a low-rank matrix capturing individual structure $\mathbf{A}^{(n)} \in \mathbb{R}^{I_n \times J}$ to each data set, and a matrix $\mathbf{R}^{(n)} \in \mathbb{R}^{I_n \times J}$ accounting for i.i.d. residual noise. That is,

$$\mathbf{X}^{(n)} = \mathbf{J}^{(n)} + \mathbf{A}^{(n)} + \mathbf{R}^{(n)}, n = 1, 2, \ldots, N. \tag{2}$$

Let $\mathbf{X}, \mathbf{J}$, and $\mathbf{R}$ be $\sum_{n=1}^{N} I_n \times J$ matrices constructed by concatenation of the corresponding matrices[1], the JIVE solves the rank-constrained least-squares problem [27]:

$$\begin{aligned} &\underset{\{\mathbf{J}, \{\mathbf{A}^{(n)}\}_{n=1}^{N}, \mathbf{R}\}}{\operatorname{argmin}} \quad \frac{1}{2}\|\mathbf{R}\|_F^2 \\ &\text{s.t. } \mathbf{R} = \mathbf{X} - \mathbf{J} - [\mathbf{A}^{(1)^T}, \mathbf{A}^{(2)^T}, \ldots, \mathbf{A}^{(n)^T}]^T, \\ &\quad \operatorname{rank}(\mathbf{J}) = K, \operatorname{rank}(\mathbf{A}^{(n)}) = K^{(n)}, \\ &\quad \mathbf{J}\mathbf{A}^{(n)^T} = \mathbf{0}, \;\; n = 1, 2, \ldots, N. \end{aligned} \tag{3}$$

Problem (3) imposes rank constraints on joint and individual components and requires the rows of $\mathbf{J}$ and $\{\mathbf{A}^{(n)}\}_{n=1}^{N}$ to be orthogonal. The intuition behind the orthogonality constraint is that, sample patterns responsible for joint structure between data types are unrelated to sample patterns responsible for individual structure [27].

A closely related method to the JIVE is the COBE which extract the common and the individual components among N data sets of the same dimensions by solving a set of least-squares minimization problems [13].

2.3 Dynamic and Canonical Time Warping

Given two temporally misaligned data sets, namely $\{\mathbf{X}^{(n)} \in \mathbb{R}^{I \times J_n}, n = 1, 2.\}$ the DTW aligns them along the time axis by solving [19]:

$$\begin{aligned} &\underset{\{\mathbf{\Delta}^{(n)}\}_{n=1}^{2}}{\operatorname{argmin}} \quad \frac{1}{2}\|\mathbf{X}^{(1)}\mathbf{\Delta}^{(1)} - \mathbf{X}^{(2)}\mathbf{\Delta}^{(2)}\|_F^2, \\ &\text{s.t. } \mathbf{\Delta}^{(n)} \in \{0,1\}^{J_n \times J}, n = 1, 2, \end{aligned} \tag{4}$$

where $\mathbf{\Delta}^{(n)}$, $n = 1, 2$ are binary selection matrices encoding the alignment path. Although the number of possible alignments is exponential in $J_1 \cdot J_2$, the DTW recovers the optimal alignment path in $\mathcal{O}(J_1 \cdot J_2)$ by employing dynamic programming. Clearly, the DTW can handle only data of the same dimensions. The CTW [17] incorporates CCA into the DTW, allowing the alignment of data sequences of different dimensions by projecting them into a common latent subspace found by the CCA [34]. Furthermore, the CCA-based projections perform feature selection by reducing the dimensionality of the data to that of the common latent subspace, handling the irrelevant or possibly noisy attributes.

More formally, let $\{\mathbf{X}^{(n)} \in \mathbb{R}^{I_n \times J_n}\}_{n=1}^{2}$ be a set of temporally misaligned data of different dimensionality (i.e., $I_1 \neq I_2$), the CCA is incorporated into the DTW by solving [17]:

$$\begin{aligned} &\underset{\{\mathbf{V}^{(n)}, \mathbf{\Delta}^{(n)}\}_{n=1}^{2}}{\operatorname{argmin}} \quad \frac{1}{2}\|\mathbf{V}^{(1)^T}\mathbf{X}^{(1)}\mathbf{\Delta}^{(1)} - \mathbf{V}^{(2)^T}\mathbf{X}^{(2)}\mathbf{\Delta}^{(2)}\|_F^2, \\ &\text{s.t. } \mathbf{V}^{(n)^T}\mathbf{X}^{(n)}\mathbf{X}^{(n)^T}\mathbf{V}^{(n)} = \mathbf{I}, \\ &\quad \mathbf{V}^{(1)^T}\mathbf{X}^{(1)}\mathbf{\Delta}^{(1)}\mathbf{\Delta}^{(2)^T}\mathbf{X}^{(2)^T}\mathbf{V}^{(2)} = \mathbf{D}, \\ &\quad \mathbf{X}^{(n)}\mathbf{\Delta}^{(n)}\mathbf{1} = \mathbf{0}, \;\; \mathbf{\Delta}^{(n)} \in \{0,1\}^{J_n \times J}, n = 1, 2. \end{aligned} \tag{5}$$

$\mathbf{V}^{(1)} \in \mathbb{R}^{I_1 \times K}$ and $\mathbf{V}^{(2)} \in \mathbb{R}^{I_2 \times K}$ project $\mathbf{X}^{(1)}$ and $\mathbf{X}^{(2)}$, respectively onto a common latent subspace of $K \leq \min(I_1, I_2)$ dimensions, where the correlation between the data sequences is maximized. $\mathbf{D}$ is a diagonal matrix of compatible dimensions. The set of constraints in (5) is imposed in order to make the CTW translation, rotation, and scaling invariant.

Remark. By adopting the least squares error, the aforementioned methods assume Gaussian distributions with small variance [14]. Such an assumption rarely holds in real word multi-modal data, where gross non-Gaussian corruptions are in abundance (cf. Section 1). Consequently, the components obtained by employing the CCA, the JIVE, the DTW, and the CTW in the analysis of grossly corrupted data may be arbitrarily away from the true ones, degenerating their performance.

To alleviate the aforementioned limitation and recover both the correlated and individual components a general framework is detailed next.

3 Robust Correlated and Individual Components Analysis

3.1 Problem Statement

Consider two data sets from different modalities or feature sets possibly contaminated by gross but sparse errors (cf. Section 1). Without loss of generality these datasets are represented by two zero-mean matrices, namely $\{\mathbf{X}^{(n)} \in \mathbb{R}^{I_n \times J}\}_{n=1}^{2}$ of different dimensions, i.e., $I_1 \neq I_2$. The RCICA recovers the *correlated* and *individual* components of the data sets as well as the *sparse corruptions* by seeking a decomposition of each matrix into three terms:

$$\mathbf{X}^{(n)} = \mathbf{C}^{(n)} + \mathbf{A}^{(n)} + \mathbf{E}^{(n)}, \;\; n = 1, 2. \tag{6}$$

$\mathbf{C}^{(n)} \in \mathbb{R}^{I_n \times J}$ and $\mathbf{A}^{(n)} \in \mathbb{R}^{I_n \times J}$ are *low-rank* matrices with mutually independent column spaces, capturing the correlated and individual components, respectively and $\mathbf{E}^{(n)} \in \mathbb{R}^{I_n \times J}$ is a *sparse* matrix accounting for sparse non-Gaussian errors.

To ensure that the fundamental identifiability of the recovered components is guaranteed, the column spaces of $\{\mathbf{A}^{(n)}\}_{n=1}^{2}$ must be orthogonal to those of $\{\mathbf{C}^{(n)}\}_{n=1}^{2}$. To facilitate this, the components are decomposed as:

$$\mathbf{C}^{(n)} = \mathbf{U}^{(n)}\,\mathbf{V}^{(n)^T}\,\mathbf{X}^{(n)}, \tag{7}$$

$$\mathbf{A}^{(n)} = \mathbf{Q}^{(n)}\,\mathbf{H}^{(n)}, \tag{8}$$

where $\{\mathbf{U}^{(n)} \in \mathbb{R}^{I_n \times K}\}_{n=1}^{2}$ and $\{\mathbf{Q}^{(n)} \in \mathbb{R}^{I_n \times K^{(n)}}\}_{n=1}^{2}$ are column orthonormal matrices spanning the columns of $\{\mathbf{C}^{(n)}\}_{n=1}^{2}$ and $\{\mathbf{A}^{(n)}\}_{n=1}^{2}$, respectively. K denotes the upper bound of unknown rank of $\{\mathbf{C}^{(n)}\}_{n=1}^{2}$ and $\{K^{(n)}\}_{n=1}^{2}$ are the upper bounds of unknown rank of $\{\mathbf{A}^{(n)}\}_{n=1}^{2}$. The mutual orthogonality of the column spaces is established by requiring $\{\mathbf{Q}^{(n)^T}\mathbf{U}^{(n)} = \mathbf{0}\}_{n=1}^{2}$. In analogy to the CCA, $\{\mathbf{V}^{(n)^T}\mathbf{X}^{(n)} \in \mathbb{R}^{K \times J}\}_{n=1}^{2}$ are required to be maximally correlated.

A natural estimator accounting for the low-rank of the correlated and independent components and the sparsity of $\{\mathbf{E}^{(n)}\}_{n=1}^{2}$ is to minimize the objective function of CCA, i.e., $\frac{1}{2}\|\mathbf{V}^{(1)^T}\mathbf{X}^{(1)} - \mathbf{V}^{(2)^T}\mathbf{X}^{(2)}\|_F^2$ as well as the rank of $\{\mathbf{C}^{(n)}, \mathbf{A}^{(n)}\}_{n=1}^{2}$ and the number of nonzero entries of $\{\mathbf{E}^{(n)}\}_{n=1}^{2}$ measured by the ℓ_0-(quasi) norm, e.g., [15], [29],

[1] $\mathbf{X} \doteq [\mathbf{X}^{(1)^T}, \mathbf{X}^{(2)^T}, \ldots, \mathbf{X}^{(N)^T}]^T$, $\mathbf{J} \doteq [\mathbf{J}^{(1)^T}, \mathbf{J}^{(2)^T}, \ldots, \mathbf{J}^{(N)^T}]^T$, $\mathbf{R} \doteq [\mathbf{R}^{(1)^T}, \mathbf{R}^{(2)^T}, \ldots, \mathbf{R}^{(N)^T}]^T$.

[35], [18]. Unfortunately, both rank and ℓ_0-norm minimization is NP-hard [36], [37]. The nuclear- and the ℓ_1- norms are typically adopted as convex surrogates to rank and ℓ_0- norm, respectively [38], [39]. Accordingly, the objective function for the RCICA is defined as:

$$\mathcal{F}(\mathcal{V}) \doteq \sum_{n=1}^{2} \Big[\|\mathbf{U}^{(n)}\mathbf{V}^{(n)^T}\|_* + \lambda_*^{(n)} \|\mathbf{Q}^{(n)}\mathbf{H}^{(n)}\|_* + \lambda_1^{(n)} \|\mathbf{E}^{(n)}\|_1 \Big] + \frac{\lambda_c}{2} \|\mathbf{V}^{(1)^T}\mathbf{X}^{(1)} - \mathbf{V}^{(2)^T}\mathbf{X}^{(2)}\|_F^2, \tag{9}$$

where the unknown variables are collected in $\mathcal{V} \doteq \{\mathbf{U}^{(n)}, \mathbf{V}^{(n)}, \mathbf{Q}^{(n)}, \mathbf{H}^{(n)}, \mathbf{E}^{(n)}\}_{n=1}^2$ and λ_c, $\{\lambda_*^{(n)}\}_{n=1}^2$, $\{\lambda_1^{(n)}\}_{n=1}^2$, are positive parameters controlling the correlation, rank, and sparsity of the derived spaces.

Due to the unitary invariance of the nuclear-norm, e.g., $\|\mathbf{Q}^{(n)}\mathbf{V}^{(n)^T}\|_* = \|\mathbf{V}^{(n)^T}\|_*$, (9) is simplified and thus the RCICA solves the constrained non-linear optimization problem:

$$\begin{aligned} \underset{\mathcal{V}}{\text{argmin}} \ & \sum_{n=1}^{2} \Big[\|\mathbf{V}^{(n)^T}\|_* + \lambda_*^{(n)} \|\mathbf{H}^{(n)}\|_* + \lambda_1^{(n)} \|\mathbf{E}^{(n)}\|_1 \Big] + \frac{\lambda_c}{2} \|\mathbf{V}^{(1)^T}\mathbf{X}^{(1)} - \mathbf{V}^{(2)^T}\mathbf{X}^{(2)}\|_F^2, \\ \text{s.t.} \ & \text{(i)} \ \mathbf{X}^{(n)} = \mathbf{U}^{(n)}\mathbf{V}^{(n)^T}\mathbf{X}^{(n)} + \mathbf{Q}^{(n)}\mathbf{H}^{(n)} + \mathbf{E}^{(n)} \\ & \text{(ii)} \ \mathbf{V}^{(n)^T}\mathbf{X}^{(n)}\mathbf{X}^{(n)^T}\mathbf{V}^{(n)} = \mathbf{I}, \\ & \text{(iii)} \ \mathbf{U}^{(n)^T}\mathbf{U}^{(n)} = \mathbf{I}, \ \ \mathbf{Q}^{(n)^T}\mathbf{Q}^{(n)} = \mathbf{I}, \\ & \text{(iv)} \ \mathbf{Q}^{(n)^T}\mathbf{U}^{(n)} = \mathbf{0}, \ \ n = 1, 2. \end{aligned} \tag{10}$$

Recall that the constraints (i) decompose each matrix into three terms capturing the correlated and the individual components as well as the sparse corruptions. The constraints (ii) are inherited by the CCA (cf. (1)) and are imposed in order to normalize the variance of the correlated components thus making them invariant to translation, rotation, and scaling (i.e., since data may have large scale differences, this constraint normalizes them in order to facilitate the identification of correlated/shared components). The third set of constraints (iii) deem RCICA to be a *projective* method, a point which will be further clarified shortly in what follows. The constraints (iv) are imposed in order to ensure the identifiability of the model. That is, in order to perfectly disentangle the low-rank correlated and individual components, their column spaces should be mutually orthogonal. Otherwise, it would be impossible to guarantee the feasibility of the decomposition.

If we assume that there are no individual components (i.e., by setting $\{\lambda_*^{(n)} \to \infty\}_{n=1}^2$), and the dimensionality of the data is the same i.e., $I_1 = I_2$, and by setting $\bar{\mathbf{C}}^{(n)} = \mathbf{U}^{(n)}\mathbf{V}^{(n)^T}$, then the RCICA is reduced to the RCCA [18]:

$$\begin{aligned} \underset{\{\bar{\mathbf{C}}^{(n)}, \mathbf{E}^{(n)}\}_{n=1}^2}{\text{argmin}} \ & \sum_{n=1}^{2} \Big[\|\bar{\mathbf{C}}^{(n)}\|_* + \lambda_1^{(n)} \|\mathbf{E}^{(n)}\|_1 \Big] + \frac{\lambda_c}{2} \|\bar{\mathbf{C}}^{(n)}\mathbf{X}^{(1)} - \bar{\mathbf{C}}^{(n)}\mathbf{X}^{(2)}\|_F^2, \\ \text{s.t.} \ & \mathbf{X}^{(n)} = \bar{\mathbf{C}}^{(n)}\mathbf{X}^{(n)} + \mathbf{E}^{(n)}, \ \ n = 1, 2, \end{aligned} \tag{11}$$

where $\{\bar{\mathbf{C}}^{(n)} \in \mathbb{R}^{I_n \times I_n}\}_{n=1}^2$ are low-rank matrices reconstructing correlated components and $\{\lambda_1^{(n)}\}_{n=1}^2$ are positive parameters controlling the sparsity in the error matrices.

Clearly, the RCICA has several appealing properties, deeming the technique advantageous in comparison to relevant methods. They are listed in what follows. 1) The RCICA is a more general approach, meaning that the CCA is also a special case of the RCICA. Indeed, if we assume that there are no gross errors in the data (i.e., $\{\mathbf{E}^{(n)} = \mathbf{0}\}_{n=1}^2$ and by letting $\{\lambda_*^{(n)} \to \infty\}_{n=1}^2$, i.e., there are no individual components, it is easy to verify that the solution of (10) is identical to that of (1), while $\{\mathbf{U}^{(n)} = \mathbf{V}^{(n)}\}$. 2) The RCICA can inherently handle data sets of different dimensionality. 3) The RCICA is projective in the sense that the correlated and individual features of unseen (test) vectors can be extracted via the projection matrices $\{\mathbf{U}^{(n)}\}_{n=1}^2$ and $\{\mathbf{Q}^{(n)}\}_{n=1}^2$, respectively. Obviously, this is not the case for the RCCA in (11) where the reconstruction of the correlated components is recovered. 4) The exact number of correlated and individual components needs not be known in advance. Instead an upper bound of the components' number is sufficient. The minimization of the nuclear-norms in (10) and (11) enable the actual number (i.e., rank) of the components to be determined automatically. Clearly, this is not the case in the CCA and the JIVE where the number of components should be exactly determined. We finally note that the RCICA and the RCCA can handle data contaminated by Gaussian noise by vanishing the error term, that is by setting $\{\lambda_1^{(n)} \to \infty\}_{n=1}^2$. Experimental results on synthetic data contaminated by Gaussian noise can be found in the supplementary material.

3.2 Alternating-Direction Method-Based Algorithm

The optimization problem (10) is difficult to be solved, mainly due to the presence of the nuclear- and ℓ_1-norms which are non-differentiable but convex functions and the set of non-linear equality constraints, i.e., the generalized orthogonality constraints (ii) and the orthogonality constraints (iii). In this paper, to solve (10) an algorithm based on the Alternating-Directions Method of Multipliers (ADMM) [32] is developed. The ADMM a simple but powerful method that is well suited to large-scale problems. It takes the form of a decomposition-coordination procedure, in which the solutions to small local subproblems are coordinated to find a solution to a large global problem.

To solve (10) via the ADMM, the generalized orthogonality constraints in (10) are tackled by introducing the splitting variables $\{\mathbf{P}^{(n)} = \mathbf{X}^{(n)^T}\mathbf{V}^{(n)}\}_{n=1}^2$. That is, the set of the generalized orthogonality constraints in (10) i.e., $\{\mathbf{V}^{(n)^T}\mathbf{X}^{(n)}\mathbf{X}^{(n)^T}\mathbf{V}^{(n)}\}_{n=1}^2 = \mathbf{I}$ is equivalently written as $\{\mathbf{X}^{(n)^T}\mathbf{V}^{(n)} = \mathbf{P}^{(n)}, \quad \mathbf{P}^{(n)^T}\mathbf{P}^{(n)} = \mathbf{I}\}_{n=1}^2$. Consequently, by collecting the set of primal variables in $\mathcal{V}' \doteq \{\mathbf{U}^{(n)}, \mathbf{V}^{(n)}, \mathbf{Q}^{(n)}, \mathbf{H}^{(n)}, \mathbf{P}^{(n)}, \mathbf{E}^{(n)}\}_{n=1}^2$, (10) is equivalent to the following optimization problem:

$$
\begin{aligned}
\underset{\mathcal{V}'}{\operatorname{argmin}} \quad & \sum_{n=1}^{2}\Big[\|\mathbf{V}^{(n)^T}\|_* + \lambda_*^{(n)}\|\mathbf{H}^{(n)}\|_* + \lambda_1^{(n)}\ \|\mathbf{E}^{(n)}\|_1\Big] \\
& + \frac{\lambda_c}{2}\|\mathbf{V}^{(1)^T}\mathbf{X}^{(1)} - \mathbf{V}^{(2)^T}\mathbf{X}^{(2)}\|_F^2, \\
\text{s.t. } & \mathbf{X}^{(n)} = \mathbf{U}^{(n)}\mathbf{V}^{(n)^T}\mathbf{X}^{(n)} + \mathbf{Q}^{(n)}\mathbf{H}^{(n)} + \mathbf{E}^{(n)} \\
& \mathbf{X}^{(n)^T}\mathbf{V}^{(n)} = \mathbf{P}^{(n)},\ \ \mathbf{P}^{(n)^T}\mathbf{P}^{(n)} = \mathbf{I}, \\
& \mathbf{U}^{(n)^T}\mathbf{U}^{(n)} = \mathbf{I},\ \ \mathbf{Q}^{(n)^T}\mathbf{Q}^{(n)} = \mathbf{I}, \\
& \mathbf{Q}^{(n)^T}\mathbf{U}^{(n)} = \mathbf{0},\ \ n = 1, 2.
\end{aligned} \tag{12}
$$

Next, (12) is solved by developing a natural variant of the ADMM where a partially augmented Lagrangian function is minimized. Here, partial refers to when some of the constraints are not included in the augmentation process but kept explicitly in order to exploit their structure. Specifically, the partially *augmented* Lagrangian function for the linear constraints in (12) is introduced:

$$
\begin{aligned}
& \mathcal{L}(\mathcal{V}', \mathcal{M}) = \\
& \sum_{n=1}^{2}\Big[\|\mathbf{V}^{(n)^T}\|_* + \lambda_*^{(n)}\|\mathbf{H}^{(n)}\|_* + \lambda_1^{(n)}\ \|\mathbf{E}^{(n)}\|_1\Big] \\
& + \frac{\lambda_c}{2}\|\mathbf{V}^{(1)^T}\mathbf{X}^{(1)} - \mathbf{V}^{(2)^T}\mathbf{X}^{(2)}\|_F^2\ + \\
& \sum_{n=1}^{2}\Big[\operatorname{tr}\Big(\mathbf{\Lambda}^{(n)^T}\Big(\mathbf{X}^{(n)} - \mathbf{U}^{(n)}\mathbf{V}^{(n)^T}\mathbf{X}^{(n)} - \mathbf{Q}^{(n)}\mathbf{H}^{(n)} - \mathbf{E}^{(n)}\Big)\Big) \\
& + \frac{\mu^{(n)}}{2}\|\mathbf{X}^{(n)} - \mathbf{U}^{(n)}\mathbf{V}^{(n)^T}\mathbf{X}^{(n)} - \mathbf{Q}^{(n)}\mathbf{H}^{(n)} - \mathbf{E}^{(n)}\|_F^2 \\
& + \operatorname{tr}\Big(\mathbf{M}^{(n)^T}\Big(\mathbf{X}^{(n)^T}\mathbf{V}^{(n)} - \mathbf{P}^{(n)}\Big)\Big) \\
& + \frac{\mu^{(n)}}{2}\|\mathbf{X}^{(n)^T}\mathbf{V}^{(n)} - \mathbf{P}^{(n)}\|_F^2\Big],
\end{aligned} \tag{13}
$$

where $\{\mu^{(n)}\}_{n=1}^2$ are positive parameters and $\mathcal{M} \doteq \{\mathbf{\Lambda}^{(n)}, \mathbf{M}^{(n)}\}_{n=1}^2$ gathers the Lagrange multipliers associated with the sets of linear constraints $\{\mathbf{X}^{(n)} = \mathbf{U}^{(n)}\mathbf{V}^{(n)^T}\mathbf{X}^{(n)} + \mathbf{Q}^{(n)}\mathbf{H}^{(n)} + \mathbf{E}^{(n)}\}_{n=1}^2$ and $\{\mathbf{X}^{(n)^T}\mathbf{V}^{(n)} = \mathbf{P}^{(n)}\}_{n=1}^2$ in (12).

Therefore, (12) is equivalent to solving

$$
\begin{aligned}
\underset{\mathcal{V}'}{\operatorname{argmin}}\, \mathcal{L}(\mathcal{V}', \mathcal{M})\ \ \text{s.t. } & \mathbf{P}^{(n)^T}\mathbf{P}^{(n)} = \mathbf{I}, \\
& \mathbf{U}^{(n)^T}\mathbf{U}^{(n)} = \mathbf{I},\ \ \mathbf{Q}^{(n)^T}\mathbf{Q}^{(n)} = \mathbf{I}, \\
& \mathbf{Q}^{(n)^T}\mathbf{U}^{(n)} = \mathbf{0},\ \ n = 1, 2.
\end{aligned} \tag{14}
$$

The proposed ADMM-based solver minimizes (14), where the objective function is described in (13), with respect to each variable in an alternating fashion and finally the Lagrange multipliers are updated at each iteration. Let t denotes the iteration index, given $\{\mathcal{V}'[t], \mathcal{M}[t]\}$ and $\{\mu^{(n)}\}_{n=1}^2$ the iteration of the ADMM solver reads as follows:

Update the primal variables:

$$
\begin{aligned}
& \mathbf{U}^{(n)}[t+1] = \underset{\mathbf{U}^{(n)}}{\operatorname{argmin}}\, \mathcal{L}\left(\mathcal{V}'[t], \mathcal{M}[t]\right) \\
& \quad \text{s.t. } \mathbf{U}^{(n)^T}\mathbf{U}^{(n)} = \mathbf{I}, \mathbf{Q}^{(n)^T}\mathbf{U}^{(n)} = \mathbf{0},\ \ n = 1, 2. \\
& \quad = \underset{\mathbf{U}^{(n)}}{\operatorname{argmin}} \frac{\mu^{(n)}}{2}\|\mathbf{X}^{(n)} - \mathbf{U}^{(n)}\mathbf{V}^{(n)^T}\mathbf{X}^{(n)} - \mathbf{Q}^{(n)}\mathbf{H}^{(n)} \\
& \qquad\qquad - \mathbf{E}^{(n)} + {\mu^{(n)}}^{-1}\mathbf{\Lambda}^{(n)}\|_F^2 \\
& \quad \text{s.t. } \mathbf{U}^{(n)^T}\mathbf{U}^{(n)} = \mathbf{I},\ \ \mathbf{Q}^{(n)^T}\mathbf{U}^{(n)} = \mathbf{0},\ \ n = 1, 2.
\end{aligned} \tag{15}
$$

$$
\begin{aligned}
& \mathbf{V}^{(n)}[t+1] = \underset{\mathbf{V}^{(n)}}{\operatorname{argmin}}\, \mathcal{L}\left(\mathcal{V}'[t], \mathcal{M}[t]\right) \\
& \quad = \underset{\mathbf{V}^{(n)}}{\operatorname{argmin}} \sum_{n=1}^{2}\|\mathbf{V}^{(n)^T}\|_* + \frac{\lambda_c}{2}\|\mathbf{V}^{(1)^T}\mathbf{X}^{(1)} - \mathbf{V}^{(2)^T}\mathbf{X}^{(2)}\|_F^2 \\
& \quad + \frac{\mu^{(n)}}{2}\|\mathbf{X}^{(n)} - \mathbf{U}^{(n)}\mathbf{V}^{(n)^T}\mathbf{X}^{(n)} - \mathbf{Q}^{(n)}\mathbf{H}^{(n)} - \mathbf{E}^{(n)} \\
& \qquad + {\mu^{(n)}}^{-1}\mathbf{\Lambda}^{(n)}\|_F^2 \\
& \quad + \frac{\mu^{(n)}}{2}\|\mathbf{X}^{(n)^T}\mathbf{V}^{(n)} - \mathbf{P}^{(n)} + {\mu^{(n)}}^{-1}\mathbf{M}^{(n)}\|_F^2,\ \ n = 1, 2.
\end{aligned} \tag{16}
$$

$$
\begin{aligned}
& \mathbf{Q}^{(n)}[t+1] = \underset{\mathbf{Q}^{(n)}}{\operatorname{argmin}}\, \mathcal{L}\left(\mathcal{V}'[t], \mathcal{M}[t]\right) \\
& \quad \text{s.t. } \mathbf{Q}^{(n)^T}\mathbf{Q}^{(n)} = \mathbf{I}, \mathbf{Q}^{(n)^T}\mathbf{U}^{(n)} = \mathbf{0},\ \ n = 1, 2. \\
& \quad = \underset{\mathbf{Q}^{(n)}}{\operatorname{argmin}} \frac{\mu^{(n)}}{2}\|\mathbf{X}^{(n)} - \mathbf{U}^{(n)}\mathbf{V}^{(n)^T}\mathbf{X}^{(n)} - \mathbf{Q}^{(n)}\mathbf{H}^{(n)} \\
& \qquad\qquad - \mathbf{E}^{(n)} + {\mu^{(n)}}^{-1}\mathbf{\Lambda}^{(n)}\|_F^2 \\
& \quad \text{s.t. } \mathbf{Q}^{(n)^T}\mathbf{Q}^{(n)} = \mathbf{I},\ \ \mathbf{Q}^{(n)^T}\mathbf{U}^{(n)} = \mathbf{0},\ \ n = 1, 2.
\end{aligned} \tag{17}
$$

$$
\begin{aligned}
& \mathbf{H}^{(n)}[t+1] = \underset{\mathbf{H}^{(n)}}{\operatorname{argmin}}\, \mathcal{L}\left(\mathcal{V}'[t], \mathcal{M}[t]\right) \\
& \quad = \underset{\mathbf{H}^{(n)}}{\operatorname{argmin}} \sum_{n=1}^{2}\lambda_*^{(n)}\|\mathbf{H}^{(n)}\|_* + \frac{\mu^{(n)}}{2}\|\mathbf{X}^{(n)} - \mathbf{U}^{(n)}\mathbf{V}^{(n)^T}\mathbf{X}^{(n)} \\
& \qquad - \mathbf{Q}^{(n)}\mathbf{H}^{(n)} - \mathbf{E}^{(n)} + {\mu^{(n)}}^{-1}\mathbf{\Lambda}^{(n)}\|_F^2,\ \ n = 1, 2.
\end{aligned} \tag{18}
$$

$$
\begin{aligned}
& \mathbf{P}^{(n)}[t+1] = \underset{\mathbf{P}^{(n)}}{\operatorname{argmin}}\, \mathcal{L}\left(\mathcal{V}'[t], \mathcal{M}[t]\right) \\
& \quad \text{s.t. } \mathbf{P}^{(n)^T}\mathbf{P}^{(n)} = \mathbf{I},\ \ n = 1, 2. \\
& \quad = \underset{\mathbf{P}^{(n)}}{\operatorname{argmin}} \frac{\mu^{(n)}}{2}\|\mathbf{X}^{(n)^T}\mathbf{V}^{(n)} - \mathbf{P}^{(n)} + {\mu^{(n)}}^{-1}\mathbf{M}^{(n)}\|_F^2 \\
& \quad \text{s.t. } \mathbf{P}^{(n)^T}\mathbf{P}^{(n)} = \mathbf{I},\ \ n = 1, 2.
\end{aligned} \tag{19}
$$

$$
\begin{aligned}
& \mathbf{E}^{(n)}[t+1] = \underset{\mathbf{E}^{(n)}}{\operatorname{argmin}}\, \mathcal{L}\left(\mathcal{V}'[t], \mathcal{M}[t]\right) \\
& \quad = \underset{\mathbf{E}^{(n)}}{\operatorname{argmin}} \sum_{n=1}^{2}\lambda_1^{(n)}\|\mathbf{E}^{(n)}\|_1 + \frac{\mu^{(n)}}{2}\|\mathbf{X}^{(n)} - \mathbf{U}^{(n)}\mathbf{V}^{(n)^T}\mathbf{X}^{(n)} \\
& \qquad - \mathbf{Q}^{(n)}\mathbf{H}^{(n)} - \mathbf{E}^{(n)} + {\mu^{(n)}}^{-1}\mathbf{\Lambda}^{(n)}\|_F^2,\ \ n = 1, 2.
\end{aligned} \tag{20}
$$

Update the Lagrange Multipliers:

$$
\begin{aligned}
\mathbf{\Lambda}[t+1] = \mathbf{\Lambda}[t] + \mu^{(n)}\big(& \mathbf{X}^{(n)} - \mathbf{U}^{(n)}\mathbf{V}^{(n)^T}\mathbf{X}^{(n)} \\
& - \mathbf{Q}^{(n)}\mathbf{H}^{(n)} - \mathbf{E}^{(n)}\big),\ \ n = 1, 2.
\end{aligned} \tag{21}
$$

$$
\mathbf{M}[t+1] = \mathbf{M}[t] + \mu^{(n)}\big(\mathbf{X}^{(n)^T}\mathbf{V}^{(n)} - \mathbf{P}^{(n)}\big),\ \ n = 1, 2. \tag{22}
$$

Algorithm 1: ADMM solver for (12).

```
1 Input: Data: ${\mathbf{X}^{(n)} \in \mathbb{R}^{I_n \times J}}_{n=1}^2$. Parameters: $\lambda_c, \{\lambda_*^{(n)}, \lambda_1^{(n)}\}_{n=1}^2$. The number (upper bound) of $K$ correlated and $\{K^{(n)}\}_{n=1}^2$ individual components.
2 Output: Correlated components $\{\mathbf{U}^{(n)}, \mathbf{V}^{(n)}\}_{n=1}^2$, individual component $\{\mathbf{Q}^{(n)}, \mathbf{H}^{(n)}\}_{n=1}^2$, and sparse errors $\{\mathbf{E}^{(n)}\}_{n=1}^2$.
1:  Initialize: Set $\{\mathbf{U}^{(n)}[0], \mathbf{V}^{(n)}[0], \mathbf{Q}^{(n)}[0], \mathbf{H}^{(n)}[0], \mathbf{P}^{(n)}[0], \mathbf{E}^{(n)}[0], \mathbf{\Lambda}^{(n)}[0], \mathbf{M}^{(n)}[0]\}_{n=1}^2$ to zero matrices,
    $\{\mu^{(n)} = 1.25/\|\mathbf{X}^{(n)}\|\}_{n=1}^2$, $\rho > 0$, $\epsilon > 0$.
2:  while not converged do
3:    for $n = 1$ to $2$ do
4:      $\mathbf{U}^{(n)}[t+1] \leftarrow \mathcal{Q}\big[\big(\mathbf{I} - \mathbf{Q}^{(n)}[t]\mathbf{Q}^{(n)}[t]^T\big)\big(\mathbf{X}^{(n)} - \mathbf{Q}^{(n)}[t]\mathbf{H}^{(n)}[t] - \mathbf{E}^{(n)}[t] + {\mu^{(n)}}^{-1}\mathbf{\Lambda}^{(n)}[t]\big)\big({\mathbf{X}^{(n)}}^T\mathbf{V}^{(n)}[t]\big)\big]$.
5:      $\mathbf{V}^{(n)}[t+1] \leftarrow \mathcal{D}_{\frac{1}{\eta^{(n)}}}\big[\mathbf{V}^{(n)}[t] - {\eta^{(n)}}^{-1}\nabla f(\mathbf{V}^{(n)}[t])\big]$.
6:      $\mathbf{Q}^{(n)}[t+1] \leftarrow \mathcal{Q}\big[\big(\mathbf{I} - \mathbf{U}^{(n)}[t+1]{\mathbf{U}^{(n)}}^T[t+1]\big)\big(\mathbf{X}^{(n)} - \mathbf{U}^{(n)}[t+1]\mathbf{V}^{(n)}[t+1]^T\mathbf{X}^{(n)} - \mathbf{E}^{(n)}[t] + {\mu^{(n)}}^{-1}\mathbf{\Lambda}^{(n)}[t]\big)\mathbf{H}^{(n)}[t]^T\big]$.
7:      $\mathbf{H}^{(n)}[t+1] \leftarrow \mathcal{D}_{\frac{\lambda_*^{(n)}}{\mu^{(n)}}}\big[(\mathbf{Q}^{(n)}[t+1]^T\big({\mu^{(n)}}^{-1}\mathbf{\Lambda}^{(n)}[t] + \mathbf{X}^{(n)} - \mathbf{U}^{(n)}[t+1]\mathbf{V}^{(n)}[t+1]^T\mathbf{X}^{(n)} - \mathbf{E}^{(n)}[t]\big)\big]$.
8:      $\mathbf{P}^{(n)}[t+1] \leftarrow \mathcal{Q}\big[{\mathbf{X}^{(n)}}^T\mathbf{V}^{(n)}[t+1] + {\mu^{(n)}}^{-1}\mathbf{M}^{(n)}[t]\big]$.
9:      $\mathbf{E}^{(n)}[t+1] \leftarrow \mathcal{S}_{\frac{\lambda_1^{(n)}}{\mu^{(n)}}}\big[\mathbf{X}^{(n)} - \mathbf{U}^{(n)}[t+1]\mathbf{V}^{(n)}[t+1]^T\mathbf{X}^{(n)} - \mathbf{Q}^{(n)}[t+1]\mathbf{H}^{(n)}[t+1] + {\mu^{(n)}}^{-1}\mathbf{\Lambda}^{(n)}[t]\big]$.
10:     Update the Lagrange multipliers by (21) and (22).
11:     Update $\mu^{(n)}$ by $\mu^{(n)} \leftarrow \min(\rho \cdot \mu^{(n)}, 10^{18})$, when the maximum relative chance in the variables is smaller than $\epsilon$.
12:     Update $\eta^{(n)}$ by $\eta^{(n)} \leftarrow \|(\mu^{(n)} + \lambda_c)\mathbf{X}^{(n)}{\mathbf{X}^{(n)}}^T\|_F$.
13:   end for
14:   Update $\lambda_c$ by $\lambda_c \leftarrow \frac{\sum_{n=1}^2 \text{rank}(\mathbf{x}^{(n)})}{\|{\mathbf{V}^{(1)}}^T[t+1]\mathbf{X}^{(1)} - {\mathbf{V}^{(2)}}^T[t+1]\mathbf{X}^{(2)}\|_F}$.
15:   $t \leftarrow t + 1$.
16: end while
```

The solutions of (15)-(20) rely on the operators and Lemmas introduced next. The shrinkage operator e.g., [15] is defined as $\mathcal{S}_\tau[q] \doteq \text{sgn}(q)\text{max}(|q|-\tau, 0)$ which can be extended to matrices by applying it element-wise. The singular value thresholding operator (SVT) is defined for any matrix $\mathbf{Y}$ as [40]: $\mathcal{D}_\tau[\mathbf{Y}] \doteq \mathbf{B}\mathcal{S}_\tau[\mathbf{\Sigma}]\mathbf{W}^T$ with $\mathbf{Y} = \mathbf{B\Sigma W}^T$ being the singular value decomposition (SVD). Furthermore, based on the SVD of $\mathbf{Y}$, the the Procrustes operator is defined as $\mathcal{Q}[\mathbf{Y}] \doteq \mathbf{BW}^T$.

Lemma 1 [41]: The constraint minimization problem:

$$\begin{aligned} &\underset{\mathbf{U}}{\text{argmin}} \ \|\mathbf{Y} - \mathbf{UD} - \mathbf{QS}\|_F^2 \\ &\text{s.t.} \ \ \mathbf{U}^T\mathbf{U} = \mathbf{I}, \mathbf{Q}^T\mathbf{U} = \mathbf{0} \end{aligned} \tag{23}$$

has a closed-form solution given by $\mathbf{U} = \mathcal{Q}[(\mathbf{I}-\mathbf{QQ}^T)\mathbf{YD}^T]$.

Lemma 2 [42]: The constraint minimization problem:

$$\underset{\mathbf{P}}{\text{argmin}} \ \|\mathbf{Y} - \mathbf{P}\|_F^2 \ \ \text{s.t.} \ \ \mathbf{P}^T\mathbf{P} = \mathbf{I}. \tag{24}$$

has a closed-form solution given by $\mathbf{P} = \mathcal{Q}[\mathbf{Y}]$.

In particular, based on Lemma 1 the solution of (15) and (17) is obtained via the Procustes operator. The solution of (16) is derived in the supplementary material and is obtained by applying the SVT operator. (18) is a nuclear norm regularized least squares minimization problem and its closed form solution is given by the SVT operator [40]. Problem (19) is solved as in Lemma 2 by the Procrustes operator. The minimizer of (20) is given by the soft thresholding operator [15]. The ADMM for solving (12) is outlined in Algorithm 1.

Computational Complexity and Convergence. The dominant cost of each iteration in Algorithm 1 is the computation the SVT operator in Step 5. Thus, the complexity of each iteration is $\mathcal{O}(\max(I_1^2 \cdot J, I_2^2 \cdot J)$. Regarding the convergence of Algorithm 1, there is no established convergence proof of the ADMM to local minima when employed to solve non-convex problems [32]. While a formal convergence proof goes beyond the scope of this paper, the weak convergence of Algorithm 1 can be established following [29]. In practice, the extensive experiments in Section 5, indicate that the convergence of Algorithm 1 is empirically guaranteed.

4 RCICA WITH TIME WARPINGS (RCITW)

Accurate temporal alignment of noisy data sequences is essential in several problems such as the alignment and the temporal segmentation of human motion [43], the alignment of facial and motion capture data [17], [18], the alignment of multiple continuous annotations [12] etc. The problem is defined as finding the temporal coordinate transformation that brings two given data sequences into alignment in time. To handle temporally misaligned, grossly corrupted data, the DTW is incorporated into the RCICA. Formally, given two sets $\{\mathbf{X}^{(n)} \in \mathbb{R}^{I_n \times J_n}\}_{n=1}^2$ of different dimensionality and length, i.e., $I_1 \neq I_2$, $J_1 \neq J_2$, the RCITW enables their temporal alignment onto the subspace spanned by the robustly estimated correlated components. To this end, the RCITW solves:

$$\begin{aligned} \underset{\{\mathcal{V}, \{\mathbf{\Delta}^{(n)}\}_{n=1}^2\}}{\text{argmin}} \ & \sum_{n=1}^{2} \Big[\|{\mathbf{V}^{(n)}}^T\|_* + \lambda_*^{(n)}\|\mathbf{H}^{(n)}\|_* + \lambda_1^{(n)} \ \|\mathbf{E}^{(n)}\|_1\Big] \\ & + \frac{\lambda_c}{2}\|{\mathbf{V}^{(1)}}^T\mathbf{X}^{(1)}\mathbf{\Delta}^{(1)} - {\mathbf{V}^{(2)}}^T\mathbf{X}^{(2)}\mathbf{\Delta}^{(2)}\|_F^2, \\ \text{s.t.} \ \ & \mathbf{X}^{(n)} = \mathbf{U}^{(n)}{\mathbf{V}^{(n)}}^T\mathbf{X}^{(n)} + \mathbf{Q}^{(n)}\mathbf{H}^{(n)} + \mathbf{E}^{(n)} \\ & {\mathbf{X}^{(n)}}^T\mathbf{V}^{(n)} = \mathbf{P}^{(n)}, \ \ {\mathbf{P}^{(n)}}^T\mathbf{P}^{(n)} = \mathbf{I}, \\ & {\mathbf{U}^{(n)}}^T\mathbf{U}^{(n)} = \mathbf{I}, \ \ {\mathbf{Q}^{(n)}}^T\mathbf{Q}^{(n)} = \mathbf{I}, \ \ {\mathbf{Q}^{(n)}}^T\mathbf{U}^{(n)} = \mathbf{0}, \\ & \mathbf{X}^{(n)}\mathbf{\Delta}^{(n)}\mathbf{1} = \mathbf{0}, \ \ \mathbf{\Delta}^{(n)} \in \{0,1\}^{J_n \times J} \ \ n = 1, 2, \end{aligned} \tag{25}$$

where $\mathbf{\Delta}^{(n)} \in \{0,1\}^{J_n \times J}, n = 1,2$ are binary selection matrices encoding the warping path as in the CTW. The constraint $\mathbf{X}^{(n)}\mathbf{\Delta}^{(n)}\mathbf{1} = \mathbf{0}, n = 1,2$ ensures that the temporally aligned data are zero-mean. By solving (25), the temporally aligned correlated components of reduced dimensions are given by $\{\mathbf{V}^{(n)^T}\mathbf{X}^{(n)}\mathbf{\Delta}^{(n)} \in \mathbb{R}^{K \times J}\}_{n=1}^2$. Moreover, one can obtain a reconstruction of the temporally aligned data in the original space by $\{\mathbf{U}^{(n)}\mathbf{V}^{(n)^T}\mathbf{X}^{(n)}\mathbf{\Delta}^{(n)} \in \mathbb{R}^{I_n \times J}\}_{n=1}^2$. Since the RCCA is a special case of the RCICA, as discussed in Section 3.1, the RCTW can be straightforwardly seen as a special case of RCITW.

An ADMM-based solver for (25) is outlined in Algorithm 2. Its derivation is similar to that of Algorithm 1 in Section 3.

Algorithm 2: ADMM solver for (25).

```
Input: As in Algorithm 1.
Output: $\{\mathbf{U}^{(n)}, \mathbf{V}^{(n)}, \mathbf{Q}^{(n)}, \mathbf{H}^{(n)}, \mathbf{E}^{(n)}\}_{n=1}^2$ as in Algorithm 1 and the warping paths $\{\mathbf{\Delta}^{(n)}\}_{n=1}^2$.
Initialize: Set all the variables to zero matrices as in Algorithm 1 and initialize $\{\mathbf{\Delta}^{(n)}[0]\}_{n=1}^2$ by the DTW.
while not converged do
    for $n = 1$ to 2 do
        Update the optimization variables $\mathbf{U}^{(n)}[t+1]$, $\mathbf{V}^{(n)}[t+1], \mathbf{Q}^{(n)}[t+1], \mathbf{H}^{(n)}[t+1], \mathbf{P}^{(n)}[t+1]$ and $\mathbf{E}^{(n)}[t+1]$ by employing the corresponding operators as in Algorithm 1.
    end for
    $\{\mathbf{\Delta}^{(n)}[t+1]\}_{n=1}^2 \leftarrow$ DTW$(\mathbf{V}^{(1)}[t+1]^T\mathbf{X}^{(1)}, \mathbf{V}^{(2)}[t+1]^T\mathbf{X}^{(2)})$.
    Make the matrices $\{\mathbf{X}^{(n)}\mathbf{\Delta}^{(n)}[t+1]\}_{n=1}^2$ zero mean.
    for $n = 1$ to 2 do
        Update the Lagrange multipliers by (21) and (22).
        Update $\mu^{(n)}$ and $\eta^{(n)}$ in Algorithm 1.
    end for
    Update $\lambda_c$ by $\lambda_c \leftarrow \frac{\sum_{n=1}^2 \text{rank}(\mathbf{X}^{(n)})}{\|\mathbf{V}^{(1)}[t+1]^T\mathbf{X}^{(1)}\mathbf{\Delta}^{(1)}[t+1] - \mathbf{V}^{(2)}[t+1]^T\mathbf{X}^{(2)}\mathbf{\Delta}^{(2)}[t+1]\|_F}$.
    $t \leftarrow t + 1$.
end while
```

5 Experimental Evaluation

This section outlines four applications that can benefit from the RCICA/RCITW, summarised in what follows.

- The performance of the RCICA is assessed in the context of multi-modal feature fusion with applications to: i) *heterogeneous face recognition and matching* (Section 5.3) and ii) *human behaviour analysis* in terms of *interest* level prediction and *conflict* detection from audio-visual cues (Section 5.4 and 5.5, respectively). In these tasks, the performance of the correlated components extracted by the RCICA and its special case, namely the RCCA, is compared against that obtained by the correlated/common components extracted by state-of-the-art methods, namely the JIVE [27], the COBE [13], the CCA [20], as well as the least-squares formulations of CCA with ℓ_1- and ℓ_2-, norm regularization [11] by conducting experiments on 4 datasets (2 for each task).
- The individual components among two distinct feature sets (i.e., pixel intensities and the Image Gradient Orientations (IGOs) [2]) are exploited for *face clustering*, constituting a third application of the RCICA (Section 5.6). By conducting experiments on 2 datasets, the performance of the individual features extracted by the RCICA is compared against that of the individual features extracted by the JIVE as well as the state-of-the-art subspace clustering methods, namely the sparse subspace clustering (SSC) [44], the low-rank representation-based subspace clustering (LRR) [45], and least-squares regression subspace clustering (LSR) [46], [47].
- As a fourth application, the RCITW is evaluated in *temporal alignment of facial expressions* by means of action units (5.7). Comparisons are made against state-of-the-art temporal alignment methods, namely the CTW [17], the GTW [43] and the RCTW [18], as previously noted is a special case of the RCITW.

Apart from the aforementioned applications, the effectiveness of the proposed robust methods is corroborated by conducting experiments with synthetic data (Section 5.1 and 5.2). These are important since they provide a ground truth, against which performance can be assessed by evaluating suitable figures of merit.

Unless otherwise specified, the parameters of the methods compared in this section were found via cross-validation in a set disjoint from the test one. All the experiments were conducted using Matlab 2013b on a i7 3.20Ghz PC, with 32GB of RAM, running Windows 7.

5.1 RCICA on Synthetic Data

Given two matrices corrupted by sparse noise, the goal of the RCICA is to correct the noise and recover the correlated and individual terms. To simulate this task, synthetic data are generated as follows. Each set of matrices $\{\mathbf{X}^{(n)} = \mathbf{C}_0^{(n)} + \mathbf{A}_0^{(n)} + \mathbf{E}_0^{(n)} \in \mathbb{R}^{I \times J}\}_{n=1}^2$ is parametrized by $(I, J, K, K^{(1)}, K^{(2)})$, where I, J are the matrices' dimensions, K is the number of correlated components, and $K^{(1)}, K^{(2)}$ are the number of individual components. To generate low-rank correlated matrices $\{\mathbf{C}_0^{(n)}\}_{n=1}^2$ the method in [48] is employed. Each individual matrix $\mathbf{A}_0^{(n)}$ with rank $K^{(n)}$ is generated as $\mathbf{A}_0^{(n)} = \mathbf{L}^{(n)}\mathbf{N}^{(n)}$, with $\mathbf{L}^{(n)} \in \mathbb{R}^{I \times K^{(n)}}$ and $\mathbf{N}^{(n)} \in \mathbb{R}^{J \times K^{(n)}}$. The entries of $\mathbf{L}, \mathbf{N}$ are independently sampled from $\mathcal{N}(0,1)$. $\mathbf{E}_0^{(n)}$ is a sparse matrix with 70% of its entries being zero. The nonzero entries are independent $\mathcal{N}(0,2)$ values.

The average recovery accuracy of the correlated components as well as the individual and sparse terms obtained by the RCICA is reported in Table 1. For comparison purposes the average recovery accuracy of the individual features obtained by the JIVE and the COBE as well as the correlated components obtained by the CCA and our preliminary method, namely the RCCA, is also presented in Table 1. It is worth mentioning that, the JIVE and the COBE extracts joint but not correlated components and thus we cannot evaluate their performance in correlated

TABLE 1: Comparison among the RCICA, the RCCA, the JIVE, the COBE, and the CCA on the synthetic data. For each quintuple $\left(I, J, K, K^{(1)}, K^{(2)}\right)$, each method is applied on the same data. The average recovery accuracy and the average running time in CPU seconds) of each method were obtained by repeating the experiments 10 times.

Size $\left(I, J, K, K^{(1)}, K^{(2)}\right)$	**Method**	$\|\mathbf{V}^{(1)^T}\mathbf{X}^{(1)} - \mathbf{V}^{(2)^T}\mathbf{X}^{(2)}\|_F$	$\left\{\frac{\|\mathbf{A}^{(n)}-\mathbf{A}_0^{(n)}\|_F}{\|\mathbf{A}_0^{(n)}\|_F}\right\}_{n=1}^2$	$\left\{\frac{\|\mathbf{E}^{(n)}-\mathbf{E}_0^{(n)}\|_F}{\|\mathbf{E}_0^{(n)}\|_F}\right\}_{n=1}^2$	**Time**
$(100, 100, 5, 10, 20)$	**RCICA**	**2.99**	**{0.10, 0.48}**	{0.05, 0.19}	0.88
	RCCA	3.25	N/A	{ 0.70, 0.71}	6.36
	JIVE	N/A	{0.59, 0.59}	N/A	25.54
	COBE	N/A	{1.71, 2.48}	N/A	0.009
	CCA	5.62	N/A	N/A	0.017
$(500, 500, 5, 10, 20)$	**RCICA**	2.17	**{0.01, 0.02}**	{0.02, 0.01}	25.55
	RCCA	1.99	N/A	{0.38,0.38}	284.78
	JIVE	N/A	{0.57, 0.40}	N/A	226.65
	COBE	N/A	{1.00, 1.00}	N/A	0.16
	CCA	**1.67**	N/A	N/A	0.32
$(1000, 1000, 50, 100, 200)$	**RCICA**	**10.17**	**{0.14, 0.48}**	{0.03, 0.06}	1.25×10^3
	RCCA	29.54	N/A	{0.36,0.34}	445.08
	JIVE	N/A	{0.60, 0.68}	N/A	9.58×10^3
	COBE	N/A	{5.34,7.74}	N/A	3.61
	CCA	10.88	N/A	N/A	3.01
$(5000, 5000, 50, 100, 200)$	**RCICA**	**7.16**	**{0.02,0.03}**	{0.01,0.01}	38.67×10^3
	RCCA	27.25	N/A	{0.36, 0.35}	4.89×10^4
	JIVE	N/A	{0.55,0.64}	N/A	205×10^3
	COBE	N/A	{5.45,7.74}	N/A	15
	CCA	94.48	N/A	N/A	29

components extraction. In these experiments, the number of the correlated/joint and individual components in the RCICA, the JIVE, and the COBE are set as $K, K^{(1)}, K^{(2)}$. The rank and sparsity controlling parameters of the RCICA are set as $\{\lambda_*^{(n)} = 1\}_{n=1}^2$ and $\{\lambda_1^{(n)} = 1/\sqrt{I}\}_{n=1}^2$, respectively and the sparsity controlling parameters of the RCCA as $\{\lambda_1^{(n)} = 1/\sqrt{I}\}_{n=1}^2$. By inspecting Table 1, we observe that the RCICA recovers accurately the correlated, the individual components, and the sparse errors than the compared methods. Also, it is faster than the JIVE and the RCCA. The COBE fails to recover the individual compotes at all. The experimental results indicate that by treating the correlated and individual components simultaneously, the actual correlated components are more accurately recovered.

5.2 Temporal Alignment of Synthetic Data

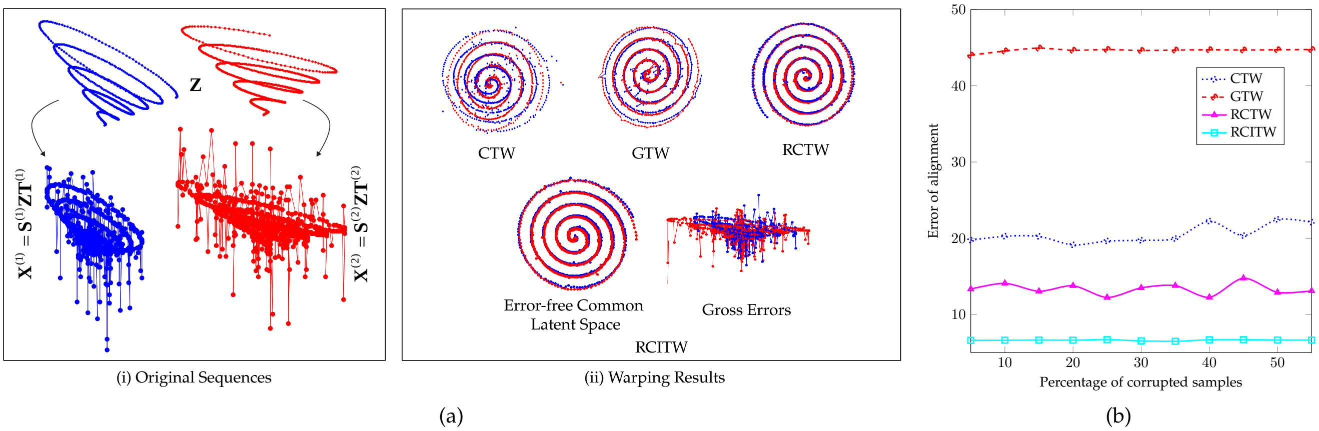
The performance of the RCITW in temporal alignment of grossly corrupted data is assessed here by conducting experiment on synthetic 3D spirals [17]. In more detail, sets of 3D spirals are generated as follows: $\mathbf{X}^{(1)} = \mathbf{S}^{(1)}\mathbf{Z}\mathbf{T}^{(1)} \in \mathbb{R}^{3\times J_1}$, $\mathbf{X}^{(2)} = \mathbf{S}^{(2)}\mathbf{Z}\mathbf{T}^{(2)} \in \mathbb{R}^{3\times J_2}$, where $\mathbf{Z} \in \mathbb{R}^{3\times J}$ is the true latent data sequence. $\mathbf{S}^{(1)}, \mathbf{S}^{(2)} \in \mathbb{R}^{3\times 3}$ and $\mathbf{T}^{(1)} \in \mathbb{R}^{J_1\times J}, \mathbf{T}^{(2)} \in \mathbb{R}^{J_2\times J}$ are random spatial and temporal warping matrices, respectively. Next, both $\mathbf{X}^{(1)}$ and $\mathbf{X}^{(2)}$ are corrupted by adding gross non-gaussian noise to a percentage of samples (i.e., columns of $\mathbf{X}^{(1)}$ and $\mathbf{X}^{(2)}$) ranging from 5 to 55%. The alignment error, a metric defined in [43], is employed for the evaluation of temporal alignment. The performance of the RCITW in compared against that of the CTW, the GTW, and the RCTW. In all experiments, the number of the correlated components in the CTW, and GTW is set to 3. The number of correlated in the RCITW is set to 3 while the number of individual components of each sequence is set to 1. The rank and sparsity controlling parameters of the RCITW are set as $\{\lambda_*^{(n)} = 1\}_{n=1}^2$ and $\{\lambda_1^{(n)} = 1/\sqrt{J_n}\}_{n=1}^2$, respectively. The same values are used for the sparsity controlling parameters of the RCTW.

In Fig. 1(a), temporal alignment results of the compared methods are illustrated. In particular, in Fig. 1(a)(i), the original 3D spirals are shown, along with the perturbation by sparse, gross noise. In Fig. 1(a)(ii), we show the resulting, temporally-aligned latent spaces derived by each competing method. It is clear that, the RCICA is able to isolate the gross errors and infer the clean, temporally aligned correlated latent space. Quantitative results are presented in Fig. 1(b), where the error is presented as a function of the percentage of corrupted samples in each synthetic sequence. The results demonstrate that the RCITW outperforms the compared methods, exhibiting a low alignment error.

5.3 Heterogeneous Face Recognition

Heterogeneous face recognition consists in matching between heterogeneous image modalities, depicting the face of the same person. The RCICA is applied to this task by conducting experiments on the CASIA Heterogeneous Face Biometrics [4] and the CUHK [5] databases. Samples from both databases are depicted in Fig. 2. The performance of the competing methods in heterogeneous face recognition is assessed using the recognition error.

The CASIA Heterogeneous Face Biometrics database [4] 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 100 subjects, with 4 VIS and 4 NIR face images per subject, while for 3D faces, 2 images per subject are included for 92 subjects, and 1 image for the remaining 8 subjects. A subset of the data for which all VIS, NIR and 3D spectrum images are available, consisting of 100 subjects and 600 images in total is used in the experiments next. We perform two sets of



(i) Original Sequences

(ii) Warping Results

(a)

(b)

Fig. 1: Application of RCITW and compared techinques to synthetic data. (a) Result visualisation, where the input spirals have been corrupted by sparse spike noise. The original sequences corrupted by noise are shown in (i), while in (ii) the time-warped latent space obtained via each method is shown. (b) Mean alignment error obtained by the CTW, the GTW, the RCTW and the RCITW, as a function of the percentage of corrupted samples on synthetic data.

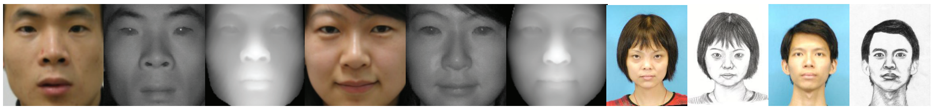


Fig. 2: Example data included in the CASIA HFB [4] (male and female subject, visual, infra-red and 3d) and CUHK [5] (female and male subject, visual and sketch) databases.

distinct experiments, considering two modalities each time in each one. In the first experiment, the data matrices $\mathbf{X}^{(1)}$ and $\mathbf{X}^{(2)}$ contain VIS spectrum and 3D images, respectively. In the second one, the data matrices are constructed using VIS spectrum and NIR images. In both experiments, the correlated components are inferred during training. During testing, only one modality is present; therefore, the correlated components are recovered by projecting the queried modality onto the correlated space, via the learnt projections. Next, we utilise the CUHK [5] database. A portion of the database containing 188 subjects is employed. For each subject, a visual image along with a sketch is provided (See Fig. 2). We use 100 subjects for training and 88 for test. Since the subjects' identities in training and test sets are disjoint, we perform correlation-based matching on the test set in order to match the sketches to the visual images and vice-versa, using the correlated space learnt during training.

Furthermore, the compared methods are evaluated in the presence of noise, by adopting six noise levels. In each level, a percentage of image's pixel is corrupted in a percentage images from each dataset. To this end, we uniformly select a number of images from each dataset, which are subsequently corrupted by superimposing black patches on a certain percentage of the image area.

In Fig. 3, the recognition error obtained by the competing methods is plotted as a function of the noise level. Clearly, the RCICA and the RCCA outperforms all other compared methods when the data are contaminated by noise.

5.4 Audio-Visual Fusion for Interest Prediction

The automatic detection of the level of interest in audio-visual sequences is a problem which has been gaining rising attention in the field of machine learning and pattern recognition [49], [50], [51], 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 the RCICA on the problem of fusion multi-modal signals for the automatic estimation of the level of interest.

Data and Annotations. The SEMAINE database [52], which contains a set of audio-visual recordings focusing on dyadic interaction scenarios, is employed. In more detail, each subject is 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. SEMAINE has been annotated in terms of emotion dimensions, particularly in terms of valence, arousal, power, expectation, and intensity. The interaction scenario employed in SEMAINE is though highly appropriate for analysing interest: since the behaviour of operators elicits naturalistic conversation, the subject can be interested in the conversation regarding some personal issue that the subject might be facing, or can become either annoyed or bored (i.e., disinterested) and e.g., request the conversation to finish or switch to another operator with different behaviour. We use a portion of the database running approximately 85 minutes, which has been annotated for emotion dimensions. We utilise 5 annotators, from which we use the averaged annotation. Furthermore, we obtained interest annotations from 8 annotators. The annotations where provided continuously over time, ranging from -1 to 1. The instructions given to the annotators were based on earlier work [50], and

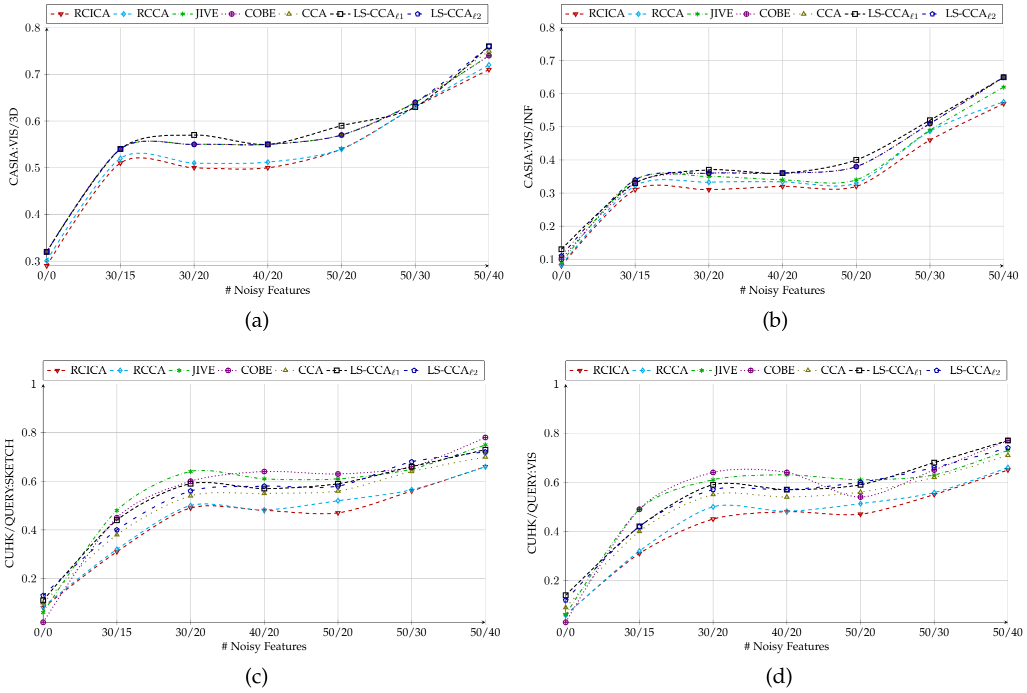


Fig. 3: Recognition error obtained by the compared methods on the CASIA HFB and CUHK databases.

have been readjusted in order to fit 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).

Feature Extraction & Experimental Setting. For extracting facial expression features, we employ an Active Appearance Model (AAM) based tracker [53], designed for simultaneous tracking of 3D head pose, lips, eyebrows, eyelids and irises in videos. For each 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 audio features, we utilise MFCCs and Delta-MFCCs 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 [7], obtaining a 29-dimensional feature vector. Cross-validation is performed given the features and annotations. Regression was performed via a Relevance Vector Machine (RVM) [54]. 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 Root Mean Squared Error (RMSE) and the Correlation Coefficient (COR).

Results and Discussion. Results are presented in Tab. 2. 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 [55], interest is more correlated with arousal, which is the primary dimension for which audio cues are known to perform better [56], [57], while this has also been confirmed by other works on interest recognition (c.f., [50]). 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 the RCICA and the RCCA outperforms all compared methods, 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 2: 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$_f$), RCICA fusion (RCICA$_f$) and other compared techniques.

	Face	Audio	F$_l$	RCICA$_f$	RCCA$_f$	JIVE$_f$	COBE$_f$	CCA$_f$	LS-CCA$_{\ell 1,f}$	LS-CCA$_{\ell 2,f}$
RMSE	0.182	0.176	0.176	0.169	0.171	0.173	0.173	0.176	0.176	0.179
COR	0.432	0.460	0.443	**0.490**	**0.490**	0.460	0.463	0.458	0.480	0.464

5.5 Audio-visual Fusion for Conflict Detection

In this section we address the problem of the automatic detection of *conflict* based on both the audio and visual modalities. Conflict is usually defined as a high level of disagreement, where at least one of the involved interlocutors feels emotionally offended. While conflict has been extensively investigated in human sciences and recognized as one of the main dimensions along which an interaction is perceived and assessed, machine analysis of conflict is limited to the works of Kim *et al.* [58], [59], where the degree of conflict in audio recording is investigated by employing various prosodic/conversational features.

The detection of conflict episodes in audio-visual recordings is inherently a difficult task, since it incorporates the simultaneous analysis of more than one subjects at the same time. The latter renders the problem even more difficult, in terms of both computer vision (tracking, localization etc.) as well as machine learning effort. In effect, these difficulties are likely to result in more noisy data for the task at hand. In this light, we apply the proposed RCICA for the robust, multi-modal fusion of audio-visual cues, aiming towards the accurate detection of conflict in interactive scenarios.

Data and Annotations. In the presented experiments we focus on a set of video excerpts containing live political debates, where conflict between participant arises naturally. The recordings consist of more than 60 hours of live political debates which have been televised in Greece between 2011 and 2012. It is important to clarify that in contrast to most datasets pertaining to multi-party conversations and interactions, these political debates are entirely unscripted and unposed, while participants have conflict of interests and are highly motivated to lead to a real conflict. We extract 160 video excerpts with a total duration of 2 hours and 50 minutes, consisting of dyadic interactions. The presence of conflict has been annotated by 10 experts in terms of conflict intensity. Discrete labels indicating the presence of conflict have been obtained by segmenting each video in non-overlapping conflict/non-conflict segments by applying an indicator function on the average annotations. This results in a total of 300 episodes, where conflict and non-conflict episodes compromise 50% of the entire number of episodes each.

Feature Extraction. As aforementioned, we extract features from both the audio and visual modalities. In particular, we utilise prosodic and cepstral features for analysing the audio content of each excerpt, namely the pitch, mean and RMS energy as well as MFCCs and differential (delta) MFCCs. This process results to a 49-dimensional audio feature vector. Regarding visual cues, we aim to capture facial behavioural cues which are deemed to be highly correlated to conflict, such as head nodding, blinking, fidgeting and frowning [60]. To this end, we utilise the recently proposed person independent Active Appearance Model (AAM) tracker, the Active Orientation Model (AOM) [2] for facial tracking. In more detail, the faces of all interactants are detected in the first frame of each video utilising the Viola-Jones face detector [61], while subsequently the AOM is applied for tracking 68 2-dimensional facial points for each of the debate participants. This process results results in a 272-dimensional feature vector obtained by stacking the tracked points for each of the participants.

Experiments and Results. We perform cross-validation to investigate the problem of audio-visual fusion for the detection of conflict both on i) a frame-based level, where frames are treated independently and a classification is performed for each frame separately, and ii) clip-based, where a single video label is assigned to each clip via majority voting. The results are presented in Table 3, where as can be clearly seen, the RCICA and its preliminary version, namely the RCCA outperforms compared methods. We note that since the RCCA cannot inherently handle datasets of different dimensions, dimensionality reduction via PCA has been applied to data before the the extraction of correlated components by the RCCA.

5.6 Face Clustering

Given face images of multiple subjects, acquired with a fixed pose and varying illumination and occlusions, we consider the problem of clustering images according to their subjects identities. To this end, the Extended Yale B [62] and the AR databases [63] are employed, where illuminations and natural pixel collusions occurred. The images of all datasets were downsampled to 48×42. Each database is represented by two matrices $\mathbf{X}^{(1)}$ and $\mathbf{X}^{(2)}$. $\mathbf{X}^{(1)}$ contains in its columns the pixel intensities of the facial images while $\mathbf{X}^{(2)}$ contains the corresponding IGOs. In this context, the individual components among the pixel intensities and IGOs are expected to carry discriminative information, suitable for accurate clustering. Consequently, the RCICA and the JIVE are employed to extract the individual components which are clustered next via the k-means algorithm. For comparison purposes, the SSC, the LRR, and the LRS are applied on the pixel intensities of the images. In all the experiments the number of the individual components as well as the number of clusters which required as inputs in the aforementioned methods are set equal to the actual number of clusters from the ground-truth. The performance of the aforementioned methods in face clustering is evaluated in terms of clustering accuracy (AC) and normalized mutual information (NMI) [64].

Two sets of experiments are conducted on each database. The Extended Yale B database consists of frontal face images of 38 individuals (64 images from each person) acquired under various lighting conditions. The face images for the

TABLE 3: Detection accuracy of conflict on the political debate data, utilising facial trackings (Face), audio cues (Audio), feature-level fusion (F_l), CCA-based fusion (CCA$_f$), Robust CCA fusion (R CICA$_f$) and other compared techniques. Both clip-based and frame based results are presented, based on a nearest neighbour classifier (NN).

	Face	Audio	F$_l$	RCICA$_f$	RCCA$_f$	JIVE$_f$	COBE$_f$	CCA$_f$	LS-CCA$_{\ell 1,f}$	LS-CCA$_{\ell 2,f}$
Clip	0.79	0.75	0.80	0.84	0.83	0.80	0.62	0.75	0.82	0.79
Frame	0.70	0.64	0.74	0.74	0.74	0.74	0.51	0.65	0.71	0.72

first 5 and 10 individuals are used in the first and second experiment, respectively.

Furthermore, clustering of face images from people wearing sunglasses and scarf are investigated by employing the AR dataset. The AR dataset contains two separate sessions. In each session, each subject has 7 face images with different facial variations, 3 face images with sunglasses occlusion and 3 face images with scarf occlusion. For sunglasses and scarf occlusions the first session was used.

The performance of the competing methods in the aforementioned experiments is reported in Table 4. The experimental results indicate that the individual features extracted by the RCICA are more discriminative than those extracted by the JIVE. The RCICA outperforms with respect to MNI all the methods that is compared to. With respect to AC, the performance of the RCICA is comparable with that obtained by the subspace clustering methods on the AR database, while it is inferior on the YALE B.

TABLE 4: Performance comparison of various methods in face clustering. The metrics are presented in %. The reported results are obtained by the averaging the results of 10 runs.

Method:	RCICA	JIVE	SSC	LRR	LSR
5 Subjects - Yale					
AC	81.38	75.13	86.56	**87.81**	84.06
NMI	**83.00**	62.10	77.86	76.46	70.97
10 Subjects - Yale					
AC	63.59	53.91	75.00	**80.07**	63.08
NMI	**73.21**	60.49	67.19	67.39	57.75
AR-Glasses					
AC	**74.76**	36.80	72.30	72.00	74.20
NMI	**88.85**	65.37	84.42	84.77	85.73
AR-Scarf					
AC	70.64	36.88	**73.70**	70.70	73.20
NMI	**86.92**	65.18	85.02	84.48	85.51

5.7 Temporal Action Unit Alignment

To asses the performance of the RCITW on the temporal alignment of facial expressions, the MMI database [65] is employed. The MMI database 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 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 [66]. 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 each video. This type of noise is common when using detection-based trackers, in which case a point can be misdirected for several frames.

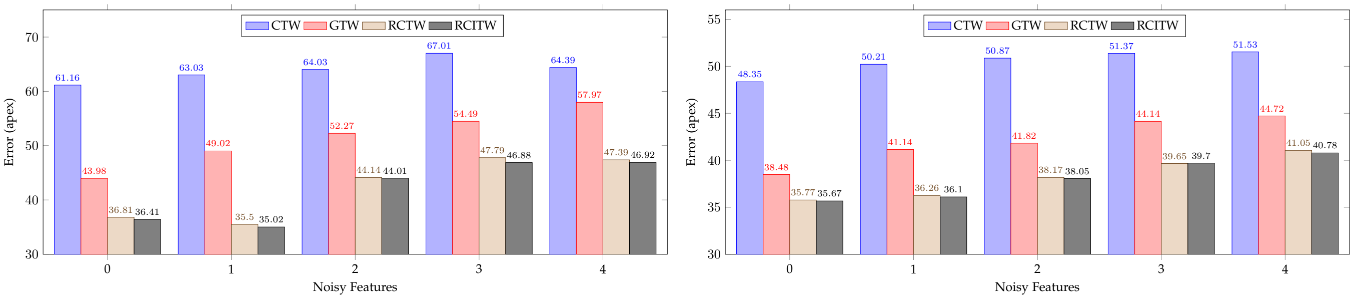
Results are presented in Fig. 4. The error we used is the percentage of misaligned frames for each pair of videos, normalised per frame (i.e., divided by the aligned video length). We present results on average (for the entire video, and results regarding the apex (which is the 'peak' of the expression. 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 RCITW can outperform both the CTW and the GTW in this scenario, maintaining relatively low error even when heavily increasing the presence of noise. The results of RCITW are comparable to those of RCCA.

6 Conclusions

A framework for simultaneous correlated and individual features extraction from two possibly temporally misaligned, noisy sets of data has been developed in this paper. By resorting to the ADMM, two novel algorithms have been proposed for solving suitable sparsity regularized rank-minimization problems for the RCICA and the RCITW, and their special cases RCCA and RCTW. Regarding applications, focus placed on muti-modal data analysis. We applied the prosed methods in multi-modal feature fusion for heterogeneous face recognition, interest prediction from videos, conflict detection in televised political debates, face clustering by employing the individual features among different visual descriptors, and temporal alignment of facial expressions. Extensive experiments on synthetic and real word data drawn from these applications domains demonstrate the robustness and the effectiveness of the proposed framework. A possible future research direction lies in exploiting discriminant information into the proposed framework for discriminant correlated and individual component analysis. Furthermore, to capture non-linear correlations among different modalities, kernel version of the RCICA and its extension will be investigated. A third line of future research includes the extension of the proposed methods so that to handle multiple (more than two) datasets.

References

[1] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *Int. J. Comput. Vision*, vol. 60, no. 2, pp. 91–110, 2004.



(a) (b)

Fig. 4: Action Unit alignment results comparing the RCITW, the RCTW, the CTW, and the GTW. (a) Error for apex phase. (b) Average error.

[2] G. Tzimiropoulos, S. Zafeiriou, and M. Pantic, “Subspace learning from image gradient orientations,” *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 34, no. 12, pp. 2454–2466, 2012.

[3] K. Simonyan, A. Vedaldi, and A. Zisserman, “Learning local feature descriptors using convex optimisation,” *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 36, no. 8, pp. 1573–1585, 2014.

[4] S. Z. Li, Z. Lei, and M. Ao, “The HFB face database for heterogeneous face biometrics research,” in *Proc. 2009 IEEE Computer Vision and Pattern Recognition Workshops*, 2009, pp. 1–8.

[5] X. Wang and X. Tang, “Face photo-sketch synthesis and recognition,” *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 31, no. 11, pp. 1955–1967, 2009.

[6] C. Shan, S. Gong, and P. W. McOwan, “Beyond facial expressions: Learning human emotion from body gestures,” in *Proc. 2007 British Machine Vision Conf.*, 2007, pp. 1–10.

[7] Z. Zhihong, M. Pantic, G. I. Roisman, and T. S. Huang, “A survey of affect recognition methods: Audio, visual, and spontaneous expressions,” *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 31, no. 1, pp. 39–58, 2009.

[8] N. M. Correa, Y.-O. Li, T. Adali, and V. D. Calhoun, “Fusion of fMRI, sMRI, and EEG data using canonical correlation analysis,” in *Proc. 2009 IEEE Int. Conf. Acoustics, Speech and Signal Processing*, 2009, pp. 385–388.

[9] P. K. Atrey, M. A. Hossain, A. El-Saddik, and M. S. Kankanhalli, “Multimodal fusion for multimedia analysis: a survey,” *Multimedia Syst.*, vol. 16, no. 6, pp. 345–379, 2010.

[10] S. Sun, “A survey of multi-view machine learning,” *Neural Computing and Applications*, vol. 23, no. 7-8, 2013.

[11] L. Sun, S. Ji, and J. Ye, “Canonical correlation analysis for multilabel classification: A least-squares formulation, extensions, and analysis,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 33, no. 1, pp. 194–200, 2011.

[12] M. A. Nicolaou, V. Pavlovic, and M. Pantic, “Dynamic probabilistic cca for analysis of affective behavior and fusion of continuous annotations,” *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 36, no. 7, pp. 1299–1311, 2014.

[13] G. Zhou, A. Cichocki, and S. Xie, “Common and individual features analysis: beyond canonical correlation analysis,” *arXiv preprint arXiv:1212.3913*, 2012.

[14] P. J. Huber, *Robust Statistics*. Wiley, 1981.

[15] E. J. Candes, X. Li, Y. Ma, and J. Wright, “Robust principal component analysis?” *Journal of ACM*, vol. 58, pp. 1–37, 2011.

[16] 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*, vol. 33, pp. 172–185, 2011.

[17] F. Zhou and F. D. la Torre, “Canonical time warping for alignment of human behavior,” in *Advances in Neural Information Processing Systems*, 2009, pp. 2286–2294.

[18] 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 Conference on Computer Vision and Pattern Recognition*, Portland, Oregon, USA, June 2013.

[19] H. Sakoe and S. Chiba, “Dynamic programming algorithm optimization for spoken word recognition,” *IEEE Trans. Acoustics, Speech, and Signal Processing*, no. 1, pp. 43–49, 1978.

[20] H. Hotteling, “Relations between two sets of variates,” *Biometrika*, vol. 28, no. 3-4, pp. 321–377, 1936.

[21] D. Chu, L.-Z. Liao, M. K. Ng, and X. Zhang, “Sparse canonical correlation analysis: New formulation and algorithm,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 12, pp. 3050–3065, 2013.

[22] S. Akaho, “A kernel method for canonical correlation analysis,” in *Proc. 2011 Int. Meeting of the Psychometric Society*. Springer-Verlag, 2001.

[23] G. Andrew, R. Arora, K. Livescu, and J. Bilmes, “Deep canonical correlation analysis,” in *Proc. 2013 Int. Conf. Machine Learning*, Atlanta, Georgia, 2013.

[24] F. R. Bach and M. I. Jordan, “A probabilistic interpretation of canonical correlation analysis,” Department of Statistics University of California, Berkeley, Tech. Rep., 2005.

[25] A. Klami, S. Virtanen, and S. Kaski, “Bayesian canonical correlation analysis,” *Journal of Machine Learning Research*, vol. 14, pp. 965–1003, 2013.

[26] J. R. Kettenring, “Canonical analysis of several sets of variables,” *Biometrika*, vol. 58, no. 3, pp. 433–451, 1971.

[27] 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*, vol. 7, no. 1, p. 523, 2013.

[28] M. A. Nicolaou, Y. Panagakis, S. Zafeiriou, and M. Pantic, “Robust canonical correlation analysis: Audio-visual fusion for learning continuous interest,” in *Proc. 2014 IEEE Int. Conf. Acoustics, Speech and Signal Processing*.

[29] G. Liu and S. Yan, “Active subspace: Toward scalable low-rank learning,” *Neural Comput.*, vol. 24, no. 12, pp. 3371–3394, 2012.

[30] Y. Panagakis, C. Kotropoulos, and G. Arce, “Music genre classification via joint sparse low-rank representation of audio features,” *IEEE/ACM Trans. Audio, Speech, and Language Processing*, vol. 22, no. 12, pp. 1905–1917, Dec 2014.

[31] G. Papamakarios, Y. Panagakis, and S. Zafeiriou, “Generalised scalable robust principal component analysis,” in *Proc. 2014 British Machine Vision Conf.*, pp. 11–37.

[32] D. P. Bertsekas, *Constrained Optimization and Lagrange Multiplier Methods*, 2nd ed. Belmont, MA: Athena Scientific, 1996.

[33] F. D. la Torre, “A least-squares framework for component analysis,” *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 34, no. 6, pp. 1041–1055, 2012.

[34] H. Hotelling, “Relations between two sets of variates,” *Biometrika*, vol. 28, no. 3/4, pp. 321–377, 1936.

[35] D. Huang, R. S. Cabral, and F. De la Torre, “Robust regression,” in *European Conference on Computer Vision (ECCV)*, 2012.

[36] L. Vandenberghe and S. Boyd, “Semidefinite programming,” *SIAM Review*, vol. 38, no. 1, pp. 49–95, 1996.

[37] B. K. Natarajan, “Sparse approximate solutions to linear systems,” *SIAM J. Comput.*, vol. 24, no. 2, pp. 227–234, 1995.

[38] M. Fazel, “Matrix rank minimization with applications,” Ph.D. dissertation, Dept. Electrical Engineering, Stanford University, CA, USA, 2002.

[39] D. Donoho, “For most large underdetermined systems of equations, the minimal ℓ_1-norm near-solution approximates the sparsest near-solution,” *Communications on Pure and Applied Mathematics*, vol. 59, no. 7, pp. 907–934, 2006.

[40] J. F. Cai, E. J. Candes, and Z. Shen, “A singular value thresholding algorithm for matrix completion,” *SIAM Journal Optimization*, vol. 2, no. 2, pp. 569–592, 2009.

[41] C. Bao, J.-F. Cai, and H. Ji, "Fast sparsity-based orthogonal dictionary learning for image restoration," in *2013 IEEE Int. Conf. Computer Vision*, 2013, pp. 3384–3391.

[42] H. Zou, T. Hastie, and R. Tibshirani, "Sparse principal component analysis," *Journal of Computational and Graphical Statistics*, vol. 15, p. 2006, 2004.

[43] F. Zhou and F. De la Torre Frade, "Generalized time warping for multi-modal alignment of human motion," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2012.

[44] E. Elhamifar and R. Vidal, "Sparse subspace clustering: Algorithm, theory, and applications," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 11, pp. 2765–2781, 2013.

[45] G. Liu, Z. Lin, S. Yan, J. Sun, and Y. Ma, "Robust recovery of subspace structures by low-rank representation," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 35, no. 1, pp. 171–184, 2013.

[46] C. Lu, H. Min, Z. Zhao, L. Zhu, D. Huang, and S. Yan, "Robust and efficient subspace segmentation via least squares regression," in *Proc. 12th European Conf. Computer Vision*, 2012, pp. 347–360.

[47] Y. Panagakis and C. Kotropoulos, "Music structure analysis by ridge regression of beat-synchronous audio features," in *Proc. 2012 Int. Conf. Music Information Retrieval*, 2012, pp. 271–276.

[48] E. Schumann, "Creating rank-correlated triangular variates," Tech. Rep., 2010.

[49] A. Pentland and A. Madan, "Perception of social interest," in *Proc. 2005 IEEE Int. Conf. on Computer Vision, Workshop on Modeling People and Human Interaction*, 2005.

[50] 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*, vol. 27, no. 12, pp. 1760–1774, 2009.

[51] B. Schuller and G. Rigoll, "Recognising interest in conversational speech-comparing bag of frames and supra-segmental features." in *Proc. 2009 Interspeech*, 2009, pp. 1999–2002.

[52] G. McKeown, M. Valstar, R. Cowie, M. Pantic, and M. Schroder, "The SEMAINE database: Annotated multimodal records of emotionally colored conversations between a person and a limited agent," *IEEE Trans. Affect. Comput.*, vol. 3, no. 1, pp. 5–17, 2012.

[53] J. Orozco, O. Rudovic, J. Gonzalez Garcia, and M. Pantic, "Hierarchical online appearance-based tracking for 3D head pose, eyebrows, lips, eyelids, and irises," *Image and vision computing*, vol. 31, no. 4, pp. 322–340, April 2013.

[54] M. E. Tipping, "Sparse bayesian learning and the relevance vector machine," *J. Mach. Learn. Res*, vol. 1, pp. 211–244, 2001.

[55] P. J. Lang, M. K. Greenwald, M. M. Bradley, and A. O. Hamm, "Looking at pictures: Affective, facial, visceral, and behavioral reactions," *Psychophysiology*, vol. 30, no. 3, pp. 261–273, 1993.

[56] M. A. Nicolaou, H. Gunes, and M. Pantic, "Continuous prediction of spontaneous affect from multiple cues and modalities in valence-arousal space," *IEEE Trans. Affective Computing*, vol. 2, no. 2, pp. 92–105, 2011.

[57] H. Gunes, B. Schuller, M. Pantic, and R. Cowie, "Emotion representation, analysis and synthesis in continuous space: A survey," in *Porc. 2011 IEEE Int. Conf. Automatic Face & Gesture Recognition and Workshops*, 2011, pp. 827–834.

[58] S. Kim, F. Valente, and A. Vinciarelli, "Automatic detection of conflicts in spoken conversations: Ratings and analysis of broadcast political debates," in *Proc. 2012 IEEE Int. Conf. Audio, Speech and Signal Processing*, 2012.

[59] S. Kim, S. H. Yella, and F. Valente, "Automatic detection of conflict escalation in spoken conversation," in *Proc. 13th Annual Conf. International Speech Communication Association*, 2012.

[60] V. W. Cooper, "Participant and observer attribution of affect in interpersonal conflict: an examination of noncontent verbal behavior," *J. Nonverbal Behavior*, vol. 10, no. 2, pp. 134–144, 1986.

[61] P. Viola and M. J. Jones, "Robust real-time face detection," *Int. J. Computer Vision*, vol. 57, no. 2, pp. 137–154, 2004.

[62] K.-C. Lee, J. Ho, and D. Kriegman, "Acquiring linear subspaces for face recognition under variable lighting," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 27, no. 5, pp. 684–698, 2005.

[63] A. Martínez and R. Benavente, "The AR face database," Computer Vision Center, Tech. Rep. 24, Jun 1998.

[64] W. Xu, X. Liu, and Y. Gong, "Document clustering based on nonnegative matrix factorization," in *Proc. 26th ACM SIGIR Int. Conf. Research and Development Informaion Retrieval*, 2003, pp. 267–273.

[65] M. Pantic, M. Valstar, R. Rademaker, and L. Maat, "Web-based database for facial expression analysis," in *Proc. of IEEE Int. Conference on Multimedia and Expo*, Amsterdam, The Netherlands, July 2005, pp. 317–321.

[66] I. Patras and M. Pantic, "Particle filtering with factorized likelihoods for tracking facial features," in *Proc. 2004 IEEE Int. Conf. on Automatic Face and Gesture Recognition*, 2004, pp. 97–102.

Yannis Panagakis is a Research Fellow in the Department of Computing, Imperial College London. He received his PhD and MSc degrees from the Department of Informatics, Aristotle University of Thessaloniki and his B.Sc. degree in Informatics and Telecommunication from the National and Kapodistrian University of Athens, Greece. Yannis received various scholarships and awards for his studies and research, including the prestigious Marie-Curie Fellowship in 2013. His current research interests include machine learning, signal processing, and mathematical optimization with applications to computer vision, human behaviour analysis, and music information research.

Mihalis A. Nicolaou is a Lecturer at the Department of Computing at Goldsmiths, University of London. Before that, he received the Ptychion (BSc, Hons.) in Informatics and Telecommunications from the University of Athens, Greece in 2008. Subsequently, Mihalis received the MSc (Distinction) and PhD (Distinction) from the Department of Computing, Imperial College London, UK in 2009 and 2014 respectively, while Mihalis was a post-doc at the same department up until August 2015. Current research interests in machine learning include component analysis and time-series analysis, with a particular focus on the machine analysis of human behavior, specifically with respect to continuous and dimensional emotion descriptions.

Stefanos Zafeiriou is a Senior Lecturer (equivalent to Associate Professor) in Pattern Recognition/Statistical Machine Learning for Computer Vision in the Department of Computing, Imperial College London. He has been awarded one of the prestigious Junior Research Fellowships (JRF) from Imperial College London in 2011 to start his own independent research group. He is/has participated in more than 12 EU, British and Greek research projects. Dr. Zafeiriou currently serves as an Associate Editor in IEEE Transactions on Cybernetics and Image and Vision Computing journal. He has been guest editor in more than four special issues and co-organized more than seven workshops/ special sessions in top venues such as CVPR/FG/ICCV/ECCV. He has co-authored more than 40 journal papers mainly on novel statistical machine learning methodologies applied to computer vision problems such as 2D/3D face and facial expression recognition, deformable object tracking, human behaviour analysis published in the most prestigious journals in his field of research (such as IEEE T-PAMI, IJCV, IEEE T-IP, IEEE T-NNLS, IEEE T-VCG, IEEE T-IFS etc). His students are frequent recipients of very prestigious and highly competitive fellowships such as Google Fellowship, Intel Fellowship and the Qualcomm fellowship.

Maja Pantic is a professor in affective and behavioral computing in the Department of Computing at Imperial College London, United Kingdom, and in the Department of Computer Science at the University of Twente, the Netherlands. She currently serves as the editor in chief of Image and Vision Computing Journal and as an associate editor for both the IEEE Transactions on Pattern Analysis and Machine Intelligence and the IEEE Transactions on Affective Computing. She has received various awards for her work on automatic analysis of human behavior, including the European Research Council Starting Grant Fellowship 2008 and the Roger Needham Award 2011. She is a fellow of the IEEE.