Subjectively Interesting Component Analysis:

Data Projections that Contrast with Prior Expectations

Bo Kang[1], Jefrey Lijffijt[1], Raúl Santos-Rodríguez[2], Tijl De Bie[1,2]

[1] Ghent University, [2] University of Bristol

What is SICA?

- Displays **subjectively interesting structure** in the data
- By means of **linear projections** of the data

Subjectively interesting projection?

- Dimensionality reduction research:
  - a. Prediction vs. exploration
  - b. Linear vs. non-linear
  - c. Objective vs. **subjective**
- Projections that are interesting to the **end-user**

How to measure?

- Given dataset $\hat{\mathbf{X}} \in \mathbb{R}^{n\times d}$ with n data points in d-dimensional Euclidean space $\hat{x} \in \mathbb{R}^d$
- Model user's belief state as **background distribution**:

$$p_{\mathbf{X}} : \mathbb{R}^{n\times d} \to \mathbb{R}$$

- Allow user to specify **prior expectations**:

- **Scale** of data:

$$\mathbb{E}_{p_{\mathbf{X}}}\left[\frac{1}{n}\sum_{i}^{n} ||x_i||^2\right] = b$$

- **Pairwise similarities** defined by **graph** $G(\hat{\mathbf{X}}, E)$, pairs in E are expected to be similar:

$$\mathbb{E}_{p_{\mathbf{X}}}\left[\frac{1}{|E|}\sum_{(i,j)\in E} ||x_i - x_j||^2\right] = c$$

- In 1-dimensional case, find projection $\hat{\mathbf{X}}\mathbf{w} \in \mathbb{R}^{n\times 1}$ that maximize the information content of **projection pattern** $\mathbf{X}\mathbf{w} \in \left[\hat{\mathbf{X}}\mathbf{w}, \hat{\mathbf{X}}\mathbf{w}+\Delta\right]$ with respect to $p_{\mathbf{X}}$:

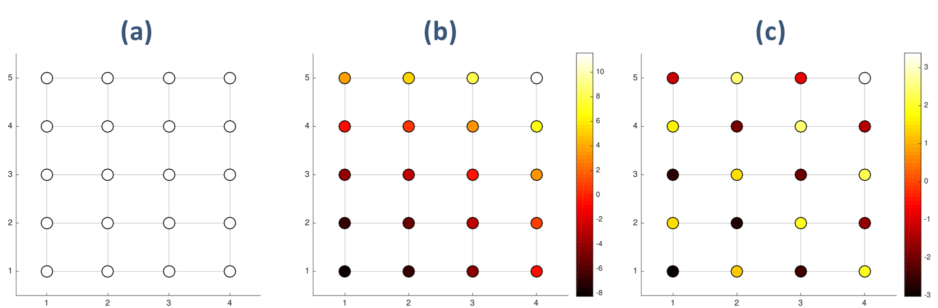
$$\max_{\mathbf{w}} \; -\log\left(\Pr_{\mathbf{x}}\left(\mathbf{X}\mathbf{w} \in \left[\hat{\mathbf{X}}\mathbf{w}, \hat{\mathbf{X}}\mathbf{w}+\Delta\right]\right)\right)$$

where Δ is a **resolution parameter**, e.g., smallest distance that is visually resolvable

- Multi-dimensional case can be extended straightforwardly

Case study: Synthetic grid

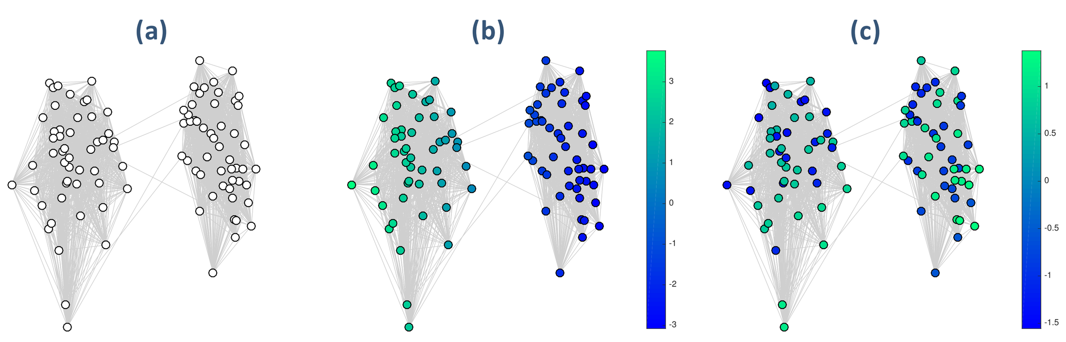
- **Dataset**: $\hat{\mathbf{X}} \in \mathbb{R}^{20\times 10}$, 20 data points on a rectangular grid. The 1st attribute varies strongly along one diagonal direction of the grid. The 2nd attribute alternates between -1 and +1. Remaining attributes are standard Gaussian noise
- **Prior expectation:** A smooth variance along the grid, encoded by a graph connecting adjacent nodes (Fig. a)



- **Results:** Top component of PCA (Fig. b) confirms user's expectation, not informative. Top SICA component reveals another underlying property (Fig. c), complementing the prior belief

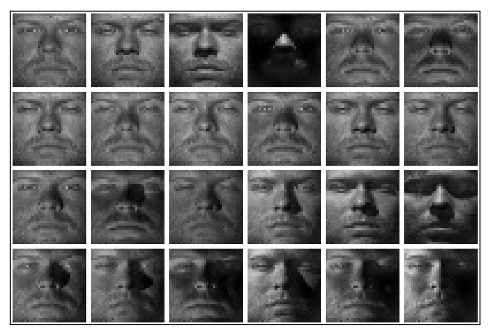
Case study: Synthetic social graph

- **Dataset**: $\hat{\mathbf{X}} \in \mathbb{R}^{100\times 10}$, a social network of 100 people. The 1st attribute separates the data into two communities. The 2nd attribute uniformly assigns -1 and +1 to data, reflecting, e.g., sentiments towards topics. Remaining attributes are standard Gaussian noise
- **Prior expectation**: People are alike if they are connected (Fig. a)



- **Results**: PCA confirms our expectation (Fig. b). SICA finds alternative 'community' structure corresponding to 2nd feature (Fig. b)

Case study: Images and lighting



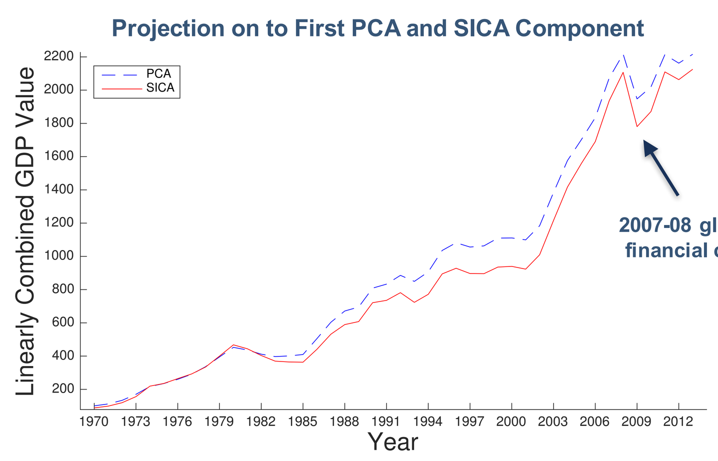
- **Dataset**: $\hat{\mathbf{X}} \in \mathbb{R}^{1684\times 1024}$, 1684 gray-scale frontal images (32 x 32 pixels) of 31 human subjects under 64 lighting conditions, compiled from the Extended Yale Face Database B
- **Prior expectation**: Images with same lighting are similar; $G(\hat{\mathbf{X}}, E)$ consists of 64 cliques
- **Results**: The top PCA eigenfaces reflect lighting conditions, while the top SICA ones reflect facial features

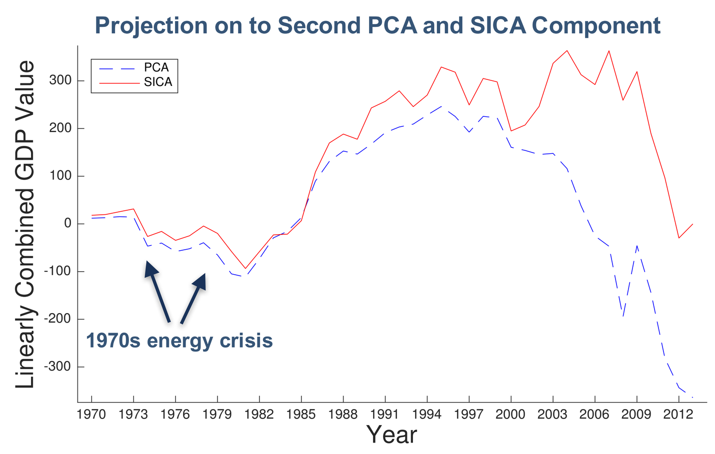
Top Five Eigenfaces of PCA

Top Five Eigenfaces of SICA

Case study: World economy

- **Dataset**: $\hat{\mathbf{X}} \in \mathbb{R}^{44\times 110}$, 44 years of GDP per capita of 110 countries between year 1970 and 2013. Countries are categorized into seven regions
- **Prior expectation**: GDP value between adjacent years are unlikely to have drastic changes
- **Results**: Projections on to first PCA and SICA components show a similar and smooth increase over the years. The projection on to second SICA component shows more local fluctuations

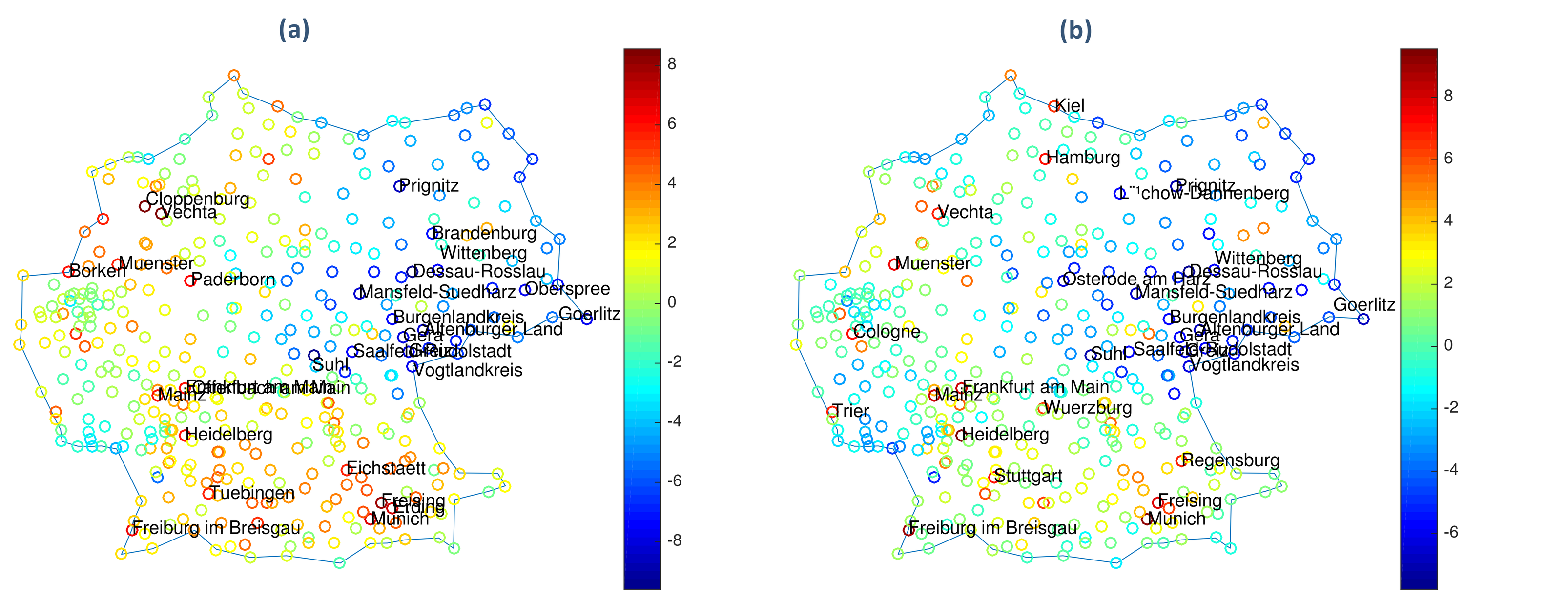




- PCA second component distributes weights equally to different regions
- SICA second component mainly gives positive weights to European countries, and negative weights to Middle-Eastern countries

Case study: Spatial socio-economics

- **Dataset**: $\hat{\mathbf{X}} \in \mathbb{R}^{412\times 5}$, age demographics of 412 districts (Landkreise) in Germany. It contains five categories: Elder (age > 64), Old (between 45 and 64), Middle Aged (between 25 and 44), Young (between 18 and 24), and Children (age < 18)
- **Prior expectation**: Historically, population density and birth rate in eastern Germany are lower than the rest of the country. This information corresponds to a graph constraint with two cliques
- **Results**: PCA confirms prior expectation (Fig. a). SICA instead highlights the large cities, whose demographics are different from less urban areas: still some contrast between East and West Germany remains (Fig. b)



- Interpret results by inspecting the elements of the first PCA and SICA component

	Elder	Old	Mid-Age	Young	Children
PCA 1st Component	-0.61	-0.42	0.43	0.09	0.51
SICA 1st Component	-0.62	-0.32	0.69	0.19	0.06