# A Graph Based Approach for Formalizing Subjective Interestingness of Data Projections

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

## 1. Motivation

- Principal Component Analysis (PCA) is effective: variance in data is often dominated by relevant structure.
- However, not all structure is interesting to every user.
- Subjectively Interesting Component Analysis (SICA) aims to find subjectively interesting projection of the data.

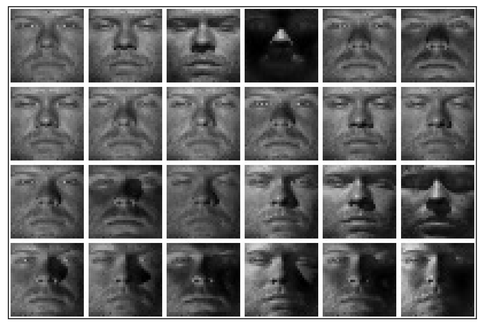
## 2. Method

- **Dataset:** $n$ data points in $d$ - dimensional real space $\hat{\boldsymbol{x}} \in \mathbb{R}^d$, represented by means of matrix $\hat{\boldsymbol{X}} \in \mathbb{R}^{n \times d}$.
- **Prior expectation:** $p_{\boldsymbol{X}} : \mathbb{R}^{n \times d} \to \mathbb{R}$ represents prior expectations of user
  - Scale of data:
$$\mathbb{E}_{p_{\boldsymbol{X}}}\left[\frac{1}{n}\sum_{i}^{n}||\boldsymbol{x}_i||^2\right] = b.$$
  - Similarities between pairs from $E$:
$$\mathbb{E}_{p_{\boldsymbol{X}}}\left[\frac{1}{|E|}\sum_{(i,j)\in E}||\boldsymbol{x}_i - \boldsymbol{x}_j||^2\right] = c.$$
  Set $E$ contains all pairs of points expected to be similar, which defines a graph $G([1 \dots n], E)$ over all data points.
- $p_{\boldsymbol{X}}$ is inferred as MaxEnt distribution subject to these constraints[1,2].
- **Subjectively interesting projection:** maximizes information content of event $\hat{\boldsymbol{X}}\boldsymbol{W} \in [\hat{\boldsymbol{\Pi}}_{\boldsymbol{W}}, \hat{\boldsymbol{\Pi}}_{\boldsymbol{W}} + \Delta\boldsymbol{1}]$ with projection $\hat{\boldsymbol{\Pi}}_{\boldsymbol{W}} \triangleq \hat{\boldsymbol{X}}\boldsymbol{W} \in \mathbb{R}^{n \times k}$.
$$\max_{\boldsymbol{W}} \ \mathrm{Tr}(\boldsymbol{W}'\hat{\boldsymbol{X}}'[\lambda\boldsymbol{I} + \mu\boldsymbol{L}]\hat{\boldsymbol{X}}\boldsymbol{W}),$$
$$\text{s.t. } \boldsymbol{W}'\boldsymbol{W} = \boldsymbol{I}.$$

[1] Tijl De Bie. An information theoretic framework for data mining. *In Proc. of KDD*, pages 564–572, 2011.
[2] Tijl De Bie. Subjective interestingness in exploratory data mining. *In Proc. of IDA*, pages 19–31, 2013.

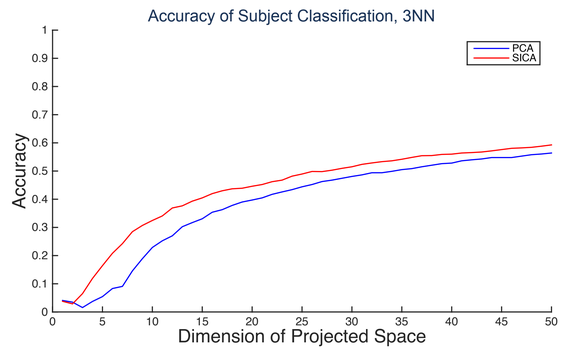
## 3.1 Case Study: Images

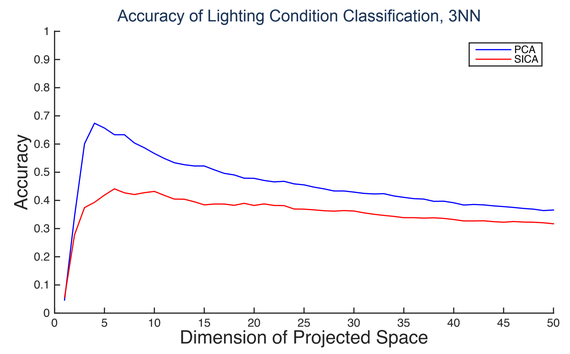
- **Dataset:** 1684 gray-scale frontal images (32x32 pixels) of 31 human subjects under 64 lighting conditions, compiled from the Extended Yale Face Database B[3,4], i.e., $\hat{\boldsymbol{X}} \in \mathbb{R}^{1684 \times 1024}$.
- **Prior expectation:** images with same lighting are similar. Resulting graph constraint consists of 64 cliques.



Subject One, First 24 Lighting Conditions

- **Results:** apply 3-NN on subject classification in projected feature space, SICA achieved better classification accuracy, while on lighting condition classification, PCA achieved better accuracy.




- The top PCA eigenfaces reflect lighting conditions, while the top SICA ones reflect facial features.

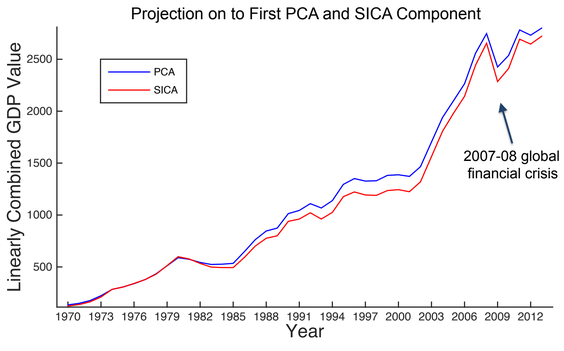
Top Five Eigenfaces of PCA

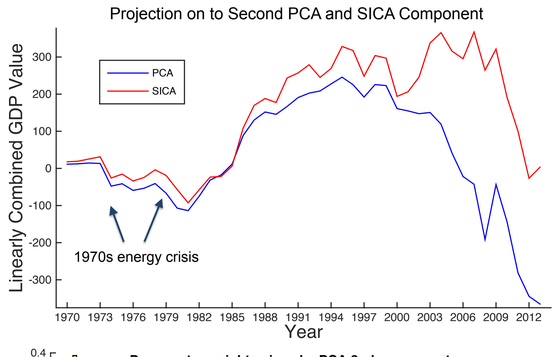
Top Five Eigenfaces of SICA

[3] A.S. Georghiades, P.N. Belhumeur, and D.J. Kriegman. From few to many: Illumination cone models for face recognition under variable lighting and pose. *IEEE Trans. Pattern Anal. Mach. Intelligence*, 23 (6): 643–660, 2001.
[4] Kuang-Chih Lee, Jeffrey Ho, and David J Kriegman. Acquiring linear subspaces for face recognition un- der variable lighting. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 27(5):684–698, 2005.

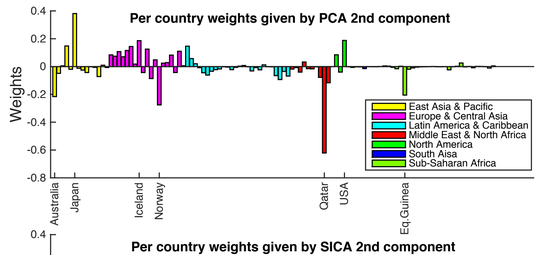
## 3.2 Case Study: Time Series

- **Dataset:** GDP per Capita[5] of 110 countries between year 1970 and 2013, i.e., $\hat{\boldsymbol{X}} \in \mathbb{R}^{44 \times 110}$. Countries are categorized into seven regions.
- **Prior expectation:** GDP value between adjacent years is unlikely to have drastic change. The resulted graph is a chain with 44 nodes.
- **Results:** projections on to first PCA and SICA components show a rather smooth increase over years.
- The projection on to second SICA component shows more local fluctuations.

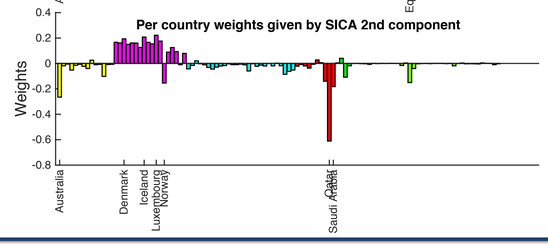



- PCA second component distributes weights equally to different regions.



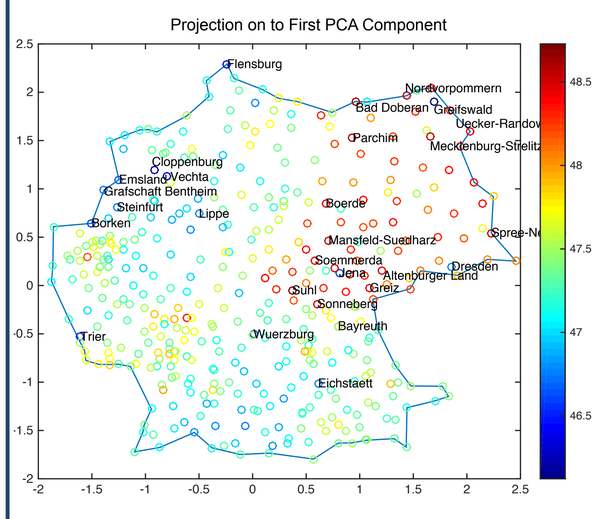

- SICA second component mainly gives positive weights to European countries, and negative weights to Middle-Eastern countries.

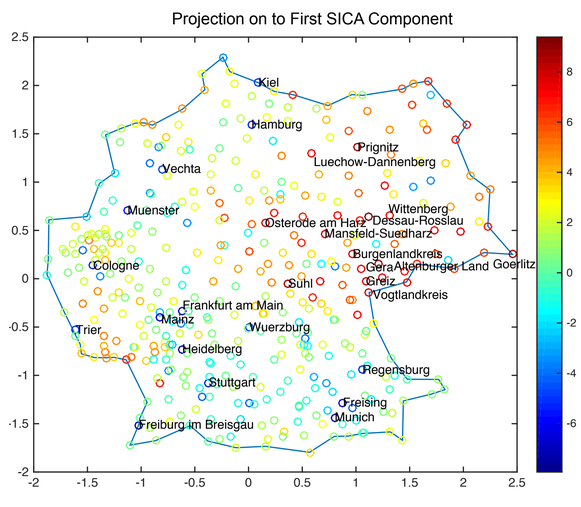



[5] The dataset is compiled from The World Bank website http://data.worldbank.org/indicator/NY.GDP.PCAP.CD

## 3.3 Case Study: Spatial Data

- **Dataset:** age structure of 412 districts (Landkreise) in Germany[6]. It contains five categories: Elder (age > 64), Old (between 45 and 64), Middle Aged (between 25 and 44), Young (between 18 and 24), Children (age < 18), i.e., $\hat{\boldsymbol{X}} \in \mathbb{R}^{412 \times 5}$.
- **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:** data points are colored by their weighted scores w.r.t the first PCA and SICA component.




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

| | Elder | Old | Mid-Age | Young | Children |
|---|---|---|---|---|---|
| PCA First Component | 0.44 | 0.60 | 0.54 | 0.17 | 0.35 |
| SICA First Component | 0.62 | 0.31 | -0.69 | -0.19 | -0.05 |

[6] The dataset is constructed based on the database provided by the German Federal Office of Statistic https://www.destatis.de