Subjectively Interesting Component Analysis: Data Projections that Contrast with Prior Expectations

Bo Kang[1] Jefrey Lijffijt[1] Raúl Santos-Rodríguez[2] Tijl De Bie[12]
[1] Data Science Lab, Ghent University, Belgium
[2] Data Science Lab, University of Bristol, UK
{bo.kang;jefrey.lijffijt;tijl.debie}@ugent.be, enrsr@bristol.ac.uk

ABSTRACT

Methods that find insightful low-dimensional projections are essential to effectively explore high-dimensional data. Principal Component Analysis is used pervasively to find low-dimensional projections, not only because it is straightforward to use, but it is also often effective, because the variance in data is often dominated by relevant structure. However, even if the projections highlight real structure in the data, not all structure is interesting to every user. If a user is already aware of, or not interested in the dominant structure, Principal Component Analysis is less effective for finding interesting components. We introduce a new method called Subjectively Interesting Component Analysis (SICA), designed to find data projections that are *subjectively interesting*, i.e, projections that truly surprise the end-user. It is rooted in information theory and employs an explicit model of a user's prior expectations about the data. The corresponding optimization problem is a simple eigenvalue problem, and the result is a trade-off between explained variance and novelty. We present five case studies on synthetic data, images, time-series, and spatial data, to illustrate how SICA enables users to find (subjectively) interesting projections.

Keywords

Exploratory Data Mining; Dimensionality Reduction; Information Theory; Subjective Interestingness

1. INTRODUCTION

Dimensionality-reduction methods differ in two main aspects: whether (1) the aim is to predict or to explore data, e.g., random projections are linear projections used in classification, and whether (2) it yields linear or non-linear projections, e.g., Self-Organizing Maps find non-linear projections that are used mostly in exploratory analysis. We study an aspect of dimensionality reduction orthogonal to these two aspects, namely that it may be helpful to incorporate prior expectations to identify *subjectively interesting* projections.

KDD '16 August 13-17, 2016, San Francisco, CA, USA

ACM ISBN 978-1-4503-4232-2/16/08.

DOI: http://dx.doi.org/10.1145/2939672.2939840

In exploratory data analysis, users are typically interested in visualizations that highlight surprising information and patterns [8]. That is, users are interested in data projections that complement or contradict their prior expectations, rather than projections that confirm them. When the goal is predictive modelling, incorporating prior expectations may be useful as well, e.g., if the data has known structure that is unrelated to the prediction task. In that case, the variation corresponding to the irrelevant structure could be taken into account in the computation of the projection.

We propose a novel method, called Subjectively Interesting Component Analysis (SICA), which allows one to identify data projections that reveal sources of variation in the data *other than those expected a priori*. The method is based on quantification of the amount of information a visualization conveys to a particular user. This quantification is based on information theory and follows the principles of FORSIED (Formalising Subjective Interestingness in Exploratory Data Mining) [3, 4]. We briefly discuss this framework here, more details will follow in Section 2.

The central idea of FORSIED is to model a probability distribution, called the *background distribution*, over the space of possible data sets that reflects the knowledge a user has about the data. This probability distribution is chosen as the maximum entropy distribution subject to the user's prior beliefs about the data. The primary reason to choose the maximum entropy distribution is that it is the only choice that, from an information-theoretic perspective, is neutral. That is, it injects no new information.

Under FORSIED, *patterns*—in casu, *projection* patterns—are constraints on the possible values of the data under the background distribution, i.e., patterns specify the values of some statistics of the data. One can then quantify the probability of any pattern under the current background distribution and compute the self-information of each pattern to determine how surprising it is. Also, patterns shown can be integrated into the background distribution, after which the surprisal of other patterns can be updated. Hence, the method can continuously present surprising patterns.

We develop these ideas for a specific type of prior knowledge that a user may have: similarities (or distances) between data points. For example, users analyzing demographic data might have an understanding of the differences between cities and rural areas and think that, roughly, cities are like each other and rural areas are also like each other, but cities are not like rural areas. Another simpler example is that a user could expect adjacent geographic regions, e.g., neighboring villages, to have similar demographics.

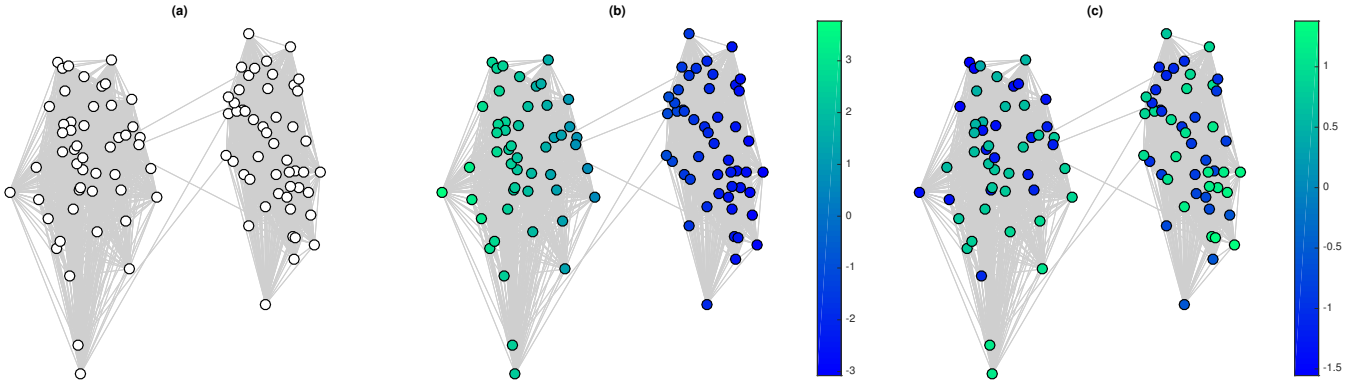

Figure 1: Communities data (§1, §4.2), (a) the actual network, (b) nodes colored according to their projected values using the first PCA component, (c) similar to (b), but for the first SICA component (our method). The x-axis corresponds to the first feature in the data, while the position of points on the y-axis is random.

We model these similarities that comprise the prior expectations in terms of a graph, where data points are nodes and nodes are connected by an edge iff they are expected to be similar. We argue that in many practical settings it is sufficiently easy to write out the graph representing the prior expectations and that it is also a powerful formalism. We illustrate the general principles in the following example.

Example. Given data comprising a social network of people, one would like to find groups that share certain properties, e.g., political views. Most trends in the data will follow the structure of the network, e.g., there is homophily (people are like their friends). Suppose that we, as the end-user, are no longer interested in the community structure, because we already know it. We synthesized data of 100 users over two communities, for details see Section 4.2. We encode the prior knowledge graph simply as the observed connections between users (Figure 1a). The result (Figure 1c) is that SICA finds a projection that is mostly orthogonal to the graph structure, actually highlighting new cluster structure unrelated to the structure of the social network.

Related work. Several unsupervised data mining and machine learning tasks, including manifold learning, dimensionality reduction, metric learning, and spectral clustering, share the common objective of finding low-dimensional manifolds that accurately preserve the relationships between the original data points. Different from PCA and ISOMAP [16], which intend to find subspaces that keep the global structure of the data intact, Locality Preserving Projections [9], Laplacian Embedding [1], and Locally Linear Embedding [15] focus on preserving the local properties of the data. Additionally, the optimization problems posed by both Locality Preserving Projections or Laplacian Embedding are very similar to spectral clustering, as they all explore the links among neighboring points, tying together those that are similar. In general, these algorithms are based on an eigendecomposition to determine an embedding of the data.

Closely related to our approach, some of the aforementioned and related methods (e.g., Laplacian-regularized models [19]) are also based on characterizing the pairwise similarity relationship among instances using graphs. Since these methods look for smooth solutions, they add a penalty in the objective function that grows for the eigenvectors corresponding to large eigenvalues of the Laplacian matrix of the graph in order to avoid abrupt changes on the graph. However, our framework follows an alternative approach: we identify mappings that, while maximizing the variance of the data in the resulting subspace, also target non-smoothness, to account for the user's interests. Interestingly, the resulting optimization problem is not simply the opposite of existing approaches. More details follow in Section 3.3.

Contributions. In this paper we introduce SICA, an efficient method to find subjectively interesting projections while accounting for known similarities between data points. To achieve this, several challenges had to be overcome. In short, we make the following contributions:

- We present a formalization of how to delineate prior knowledge in the form of expected similarities between data points. (Section 3.1)
- We derive a score for the interestingness of projection patterns given such prior knowledge. (Section 3.2)
- We show that this score can be optimized by solving a simple eigenvalue problem. (Section 3.3)
- We present five case studies, two on synthetic data and three on real data, and investigate the practical advantages and drawbacks of our method. (Section 4)

2. FORSIED AND PROJECTIONS

In this section we introduce necessary notation and review how to formalise projections as patterns within FORSIED.

Notation. Let the matrix $\hat{\boldsymbol{X}} \triangleq \begin{pmatrix} \hat{\boldsymbol{x}}_1' & \hat{\boldsymbol{x}}_2' & \cdots & \hat{\boldsymbol{x}}_n' \end{pmatrix}' \in \mathbb{R}^{n \times d}$ represent a dataset of n data points $\hat{\boldsymbol{x}} \in \mathbb{R}^d$. Methods for linear dimensionality reduction seek a set of k weight vectors $\boldsymbol{w}_i \in \mathbb{R}^d$, stored as columns of a matrix $\boldsymbol{W} \in \mathbb{R}^{d \times k}$, such that the projected data $\hat{\boldsymbol{\Pi}}_{\boldsymbol{W}} \triangleq \hat{\boldsymbol{X}}\boldsymbol{W} \in \mathbb{R}^{n \times k}$ is as informative about the data $\hat{\boldsymbol{X}}$ as possible. To fix the scale and avoid redundancies, we require (as is common) that the weight vectors have unit norm and are orthogonal to one another, i.e., that $\boldsymbol{W}'\boldsymbol{W} = \boldsymbol{I}$.

The background distribution. Our aim is to quantify the interestingness of data projections when considered against the prior belief state of the data analyst (the 'user'). This belief state is modeled by a probability density $p_{\boldsymbol{X}}$ over the set of possible values for the data $\boldsymbol{X}$ over data space $\mathbb{R}^{n \times d}$. Given this so-called *background distribution*, one can com-

pute the marginal probability density function of a projection $\mathbf{\Pi}_{\boldsymbol{W}} = \boldsymbol{X}\boldsymbol{W}$ defined by the projection matrix $\boldsymbol{W}$.

Formalizing projection patterns. We formalize a pattern as any information that restricts the set of possible values (the 'domain') of the data [3, 4]. This formalization applies to projection patterns in a natural way [5][1]: initially all the user knows is that the data belongs to $\mathbb{R}^{n\times d}$. After a specific projection $\hat{\mathbf{\Pi}}_{\boldsymbol{W}}$ is conveyed to a user through a scatter plot, the user knows that the data $\hat{\boldsymbol{X}}$ belongs to an affine subspace of $\mathbb{R}^{n\times d}$, namely: $\hat{\boldsymbol{X}} \in \left\{\boldsymbol{X} \in \mathbb{R}^{n\times d} | \boldsymbol{X}\boldsymbol{W} = \hat{\mathbf{\Pi}}_{\boldsymbol{W}}\right\}$. In practice, however, a scatter plot cannot be specified with an infinite accuracy. Instead, the projection of each data point is specified only up to a resolution Δ:

$$\hat{\boldsymbol{X}}\boldsymbol{W} \quad \in \quad [\hat{\mathbf{\Pi}}_{\boldsymbol{W}}, \hat{\mathbf{\Pi}}_{\boldsymbol{W}} + \Delta\mathbf{1}], \tag{1}$$

Note that Δ is typically very small, e.g. equal to the smallest distance that can be resolved by the human analyst on a scatter plot of the data projections. We refer to the form of expression (1) as the *projection pattern syntax*.

The subjective information content of a pattern. The Subjective Information Content (SIC) of a projection pattern is modeled adequately as minus the logarithm of the probability of the pattern. Making explicit only the dependency on the weight matrix $\boldsymbol{W}$, we have:

$$\mathrm{SIC}(\boldsymbol{W}) = -\log\Big(\Pr\big(\boldsymbol{X}\boldsymbol{W} \in [\hat{\mathbf{\Pi}}_{\boldsymbol{W}}, \hat{\mathbf{\Pi}}_{\boldsymbol{W}} + \Delta\mathbf{1}]\big)\Big), \tag{2}$$

where the probability is computed with respect to the background distribution. Indeed, this is the number of bits required to encode that the pattern is present (as opposed to absent), under a Shannon optimal code.[2] Thus a pattern is deemed subjectively more interesting if it is less plausible to the user. Note that for sufficiently small Δ, the probability of the projection pattern can be approximated accurately by $\Delta^{n\times k}$ times the probability density for $\mathbf{\Pi}_{\boldsymbol{W}}$, evaluated at the value $\hat{\mathbf{\Pi}}_{\boldsymbol{W}} = \hat{\boldsymbol{X}}\boldsymbol{W}$.

The effect of a pattern on the background distribution. Revealing a pattern to a user will affect her belief state. This effect can be modeled by specifying the user's newly learned aspects as constraints on her belief state about the data. That says, after seeing a projection, the user's background distribution should satisfy in expectation certain statistics of the projection. The distribution with maximum entropy, subject to these constraints, is an attractive choice, given its unbiasedness and robustness [4]. Further, as the resulting distribution belongs to the *exponential family*, its inference is well understood and often computationally tractable.

3. SICA

In order to apply the above framework, the following steps are required. First, we have to choose a syntax for the constraints that encode the prior expectations. Second, we need to compute the background distribution, i.e., the maximum entropy distribution subject to these constraints. Finally, we have to find the most subjectively-interesting projection given the background distribution.

3.1 The prior expectations

We consider the case where an analyst expects a priori that particular pairs of data points are similar to each other. For example, in a census data set, the user may expect that the data for adjacent regions are similar to each other. Alternatively, in a time series dataset, the user may expect that data belonging adjacent time points are similar. Such pairwise similarities can be conveniently encoded in an undirected graph $G([1..n], E)$ with n nodes labelled 1 through n and edge set E, where $(i, j) \in E$ if the user expects $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$ to be similar.

To ensure that these prior expectations hold under background distribution $p_{\boldsymbol{X}}$, we need to formalize them mathematically and enforce them as constraints on the maximum entropy optimization problem that finds $p_{\boldsymbol{X}}$. A user's expectation of the similarities between data points can be encoded by means of the following constraint:

$$\mathbb{E}_{\boldsymbol{X}}\left[\frac{1}{|E|}\sum_{(i,j)\in E} ||\boldsymbol{x}_i - \boldsymbol{x}_j||^2\right] \quad = \quad c. \tag{3}$$

However, this constraint on its own is meaningless as a small c could be the result of two things: (1) the data points paired in E are particularly close to each other, or (2) the scale of the data (as measured by the average squared norm of the data points) is expected to be small. To fix this, also the following constraint needs to be imposed, expressing the user's expectation about the overall scale of the data:

$$\mathbb{E}_{\boldsymbol{X}}\left[\frac{1}{n}\sum_{i}^{n} ||\boldsymbol{x}_i||^2\right] \quad = \quad b. \tag{4}$$

The values of b and c could be user-specified. However, a more practical implementation could assume that the user has *accurate* prior expectations, such that b and c can simply be computed based on the empirical data. We adopted this strategy in the experiments.

3.2 The Subjective Information Content

THEOREM 1. *With prior expectations defined, the Subjective Information Content (SIC) can be computed as follows.*

$$SIC(\boldsymbol{W}) = \mathrm{Tr}\left(\boldsymbol{W}'\boldsymbol{X}'\left[\lambda\boldsymbol{I} + \mu\boldsymbol{L}\right]\boldsymbol{X}\boldsymbol{W}\right) + C, \tag{5}$$

where $C = \log(Z) - nk\log(\Delta)$, a constant with respect to $\boldsymbol{W}$. $\boldsymbol{I}$ is the identity matrix and $\boldsymbol{L}$ is the Laplacian of the graph G defined as $\boldsymbol{L} = \boldsymbol{D} - \boldsymbol{A}$, with $\boldsymbol{A}$ the adjacency matrix of graph and $\boldsymbol{D}$ the diagonal matrix containing the row sums of $\boldsymbol{A}$ (i.e. the degrees of the nodes) on its diagonal.

The rest of this section is devoted to proving this theorem and can be skipped safely on a first read. To prove this, we must first derive the background distribution, and then compute the SIC of the projection pattern as in Eq. (2).

PROPOSITION 1. *The maximum entropy distribution subject to the constraints in Eqs. (3-4) (the background distribution) is given by the following probability density function:*

$$p_{\boldsymbol{X}}(\boldsymbol{X}) = \frac{1}{Z}\exp\left\{\mathrm{Tr}\left(-\boldsymbol{X}'[\lambda\boldsymbol{I} + \mu\boldsymbol{L}]\boldsymbol{X}\right)\right\}, \tag{6}$$

[1]We have presented this formalization already in a paper that is to appear. Hence, we did not include this formalization in the list of contributions of this paper.

[2]Conversely, note that by revealing a projection pattern, because of the conditioning operation the probability of the data under the user's belief state increases by a factor inversely proportional to the probability of the pattern itself. Thus, the subjective information content is also the number of bits of information the user gains about the precise value of the data by seeing the pattern.

where Z is the partition function in form:

$$Z = (2\pi)^{\frac{nd}{2}} |2[\lambda \boldsymbol{I} + \mu \boldsymbol{L}]|^{\frac{d}{2}}.$$

The proof, provided below, makes clear that the values of λ and μ depend on the values of b and c in the constraints, and can be found by solving a very simple convex optimization problem:

PROOF OF PROPOSITION 1. The optimization problem to maximize entropy subject to the prior belief constraints reads:

$$\begin{aligned} \max_{p_{\boldsymbol{X}}} \quad & -\int p_{\boldsymbol{X}}(\boldsymbol{X}) \log p_{\boldsymbol{X}}(\boldsymbol{X}) d(\boldsymbol{X}), \\ \text{s.t. } & \mathbb{E}_{\boldsymbol{X}}\left[\frac{1}{n}\sum_{i}^{n} ||\boldsymbol{x}_i||^2\right] = b, \\ & \mathbb{E}_{\boldsymbol{X}}\left[\frac{1}{|E|}\sum_{(i,j)\in E} ||\boldsymbol{x}_i - \boldsymbol{x}_j||^2\right] = c. \end{aligned}$$

This is a convex optimization problem with linear equality constraints, which can be solved using the method of Lagrange multipliers. Introducing the Lagrange multipliers λ and μ for the first and second constraint respectively, the first order optimality condition requires the partial derivative of the Lagrangian with respect to $p_{\boldsymbol{X}}(\cdot)$ to be equal to 0, i.e.:

$$-\log p(\boldsymbol{X}) - 1 - \lambda \sum_{i} ||\boldsymbol{x}_i||^2 - \mu \sum_{(i,j)\in E} ||\boldsymbol{x}_i - \boldsymbol{x}_j||^2 = 0.$$

This means that the optimal $p_{\boldsymbol{X}}$ is given by:

$$\begin{aligned} p_{\boldsymbol{X}}(\boldsymbol{X}) &= \frac{1}{Z}\exp\left\{-\lambda \sum_{i} ||\boldsymbol{x}_i||^2 - \mu \sum_{(i,j)\in E} ||\boldsymbol{x}_i - \boldsymbol{x}_j||^2\right\} \\ &= \frac{1}{Z}\exp\left\{\mathrm{Tr}\left(-\boldsymbol{X}'[\lambda \boldsymbol{I} + \mu \boldsymbol{L}]\boldsymbol{X}\right)\right\}. \end{aligned}$$

where $\boldsymbol{L}$ is as defined in the theorem statement. We observe that the distribution $p_{\boldsymbol{X}}$ is essentially a *matrix normal distribution*, namely, the matrix-valued random variable $\boldsymbol{X} \in \mathbb{R}^{n\times d}$ belongs to distribution $\mathcal{MN}_{n\times d}(\boldsymbol{M}, \boldsymbol{\Psi}, \boldsymbol{\Sigma})$. For distribution $p_{\boldsymbol{X}}$ in particular, we have $\boldsymbol{M} = \boldsymbol{0}$, $\boldsymbol{\Psi} = (2[\lambda \boldsymbol{I}_n + \mu \boldsymbol{L}])^{-1}$ and $\boldsymbol{\Sigma} = \boldsymbol{I}_d$, i.e.,

$$\boldsymbol{X} \sim \mathcal{MN}_{n\times d}(\boldsymbol{0}, (2[\lambda \boldsymbol{I}_n + \mu \boldsymbol{L}])^{-1}, \boldsymbol{I}_d),$$

where the partition function reads:

$$Z = (2\pi)^{\frac{nd}{2}} |2[\lambda \boldsymbol{I}_n + \lambda \boldsymbol{L}]|^{\frac{d}{2}}. \qquad \square$$

PROOF OF THEOREM 1. Given projection matrix $\boldsymbol{W} \in \mathbb{R}^{d\times k}$, the projected data matrix is denoted as $\boldsymbol{\Pi}_{\boldsymbol{W}} = \boldsymbol{X}\boldsymbol{W}$. Recall from *Proposition 1*, the background distribution $p_{\boldsymbol{X}} : \mathbb{R}^{n\times d} \to \mathbb{R}$ is as follows:

$$p_{\boldsymbol{X}}(\boldsymbol{X}) = \frac{1}{Z}\exp\left\{-\lambda \sum_{i} ||\boldsymbol{x}_i||^2 - \mu \sum_{(i,j)\in E} ||\boldsymbol{x}_i - \boldsymbol{x}_j||^2\right\}$$

where E is the edge set of graph $G([1\ldots n], E)$ that corresponds to the second constraint.

As the projection $\boldsymbol{\Pi}_{\boldsymbol{W}}$ is a linear transformation of random matrix $\boldsymbol{X}$, and $\boldsymbol{W}$ is of rank $k \leq n$ (full column rank) then $\boldsymbol{\Pi}_{\boldsymbol{W}} \sim \mathcal{MN}_{n\times k}\left(\boldsymbol{0}, (2[\lambda \boldsymbol{I}_n + \mu \boldsymbol{L}])^{-1}, \boldsymbol{I}_k\right)$, [7]. So the probability density function $p_{\boldsymbol{\Pi}_{\boldsymbol{W}}}$ of the projection $\boldsymbol{\Pi}_{\boldsymbol{W}}$ reads:

$$p_{\boldsymbol{\Pi}_{\boldsymbol{W}}}(\boldsymbol{\Pi}_{\boldsymbol{W}}) = \frac{1}{Z}\exp\left\{\mathrm{Tr}\left(-\boldsymbol{\Pi}_{\boldsymbol{W}}'\left[\lambda \boldsymbol{I} + \mu \boldsymbol{L}\right]\boldsymbol{\Pi}_{\boldsymbol{W}}\right)\right\}$$

By the definition of subjective information content (2), we then obtain the SIC of a projection:

$$\begin{aligned} \mathrm{SIC}(\boldsymbol{W}) &= -\log\left(\Pr\left(\boldsymbol{X}\boldsymbol{W} \in [\hat{\boldsymbol{\Pi}}_{\boldsymbol{W}}, \hat{\boldsymbol{\Pi}}_{\boldsymbol{W}} + \Delta\boldsymbol{1}]\right)\right) \\ &= -\log\left(p_{\boldsymbol{\Pi}_{\boldsymbol{W}}}(\boldsymbol{\Pi}_{\boldsymbol{W}})\right) - \log(\Delta\boldsymbol{1}) \\ &= \log(Z) + \mathrm{Tr}\left(\boldsymbol{W}'\boldsymbol{X}'\left[\lambda \boldsymbol{I} + \mu \boldsymbol{L}\right]\boldsymbol{X}\boldsymbol{W}\right) - nk\log(\Delta) \\ &= \mathrm{Tr}\left(\boldsymbol{W}'\boldsymbol{X}'\left[\lambda \boldsymbol{I} + \mu \boldsymbol{L}\right]\boldsymbol{X}\boldsymbol{W}\right) + C \end{aligned}$$

where $C = \log(Z) - nk\log(\Delta)$. $\square$

3.3 Finding the most informative projections

If we assume the resolution parameter is constant, it can be safely left out from the objective function. The most interesting projection can be obtained by solving:

$$\begin{aligned} \max_{\boldsymbol{W}} \ & \mathrm{Tr}\left(\boldsymbol{W}'\hat{\boldsymbol{X}}'\left[\lambda \boldsymbol{I} + \mu \boldsymbol{L}\right]\hat{\boldsymbol{X}}\boldsymbol{W}\right), \\ \text{s.t. } & \boldsymbol{W}'\boldsymbol{W} = \boldsymbol{I}. \end{aligned}$$

The solution to this problem consists of a matrix $\boldsymbol{W} \in \mathbb{R}^{d\times k}$ with columns equal to the eigenvectors corresponding to the top-k eigenvalues of the matrix $\hat{\boldsymbol{X}}'\left[\lambda \boldsymbol{I} + \mu \boldsymbol{L}\right]\hat{\boldsymbol{X}} \in \mathbb{R}^{d\times d}$ [10].

The computational complexity of finding an optimal projection $\boldsymbol{W}$ consists of two parts: (1) solving a convex optimization problem to obtain the background distribution. This can be achieved by applying, e.g., a steepest descent method, which uses at most $\mathcal{O}(\varepsilon^{-2})$ [13] steps (until the norm of the gradient is $\leq \varepsilon$). For each step, the complexity is dominated by matrix inversion, which has complexity $\mathcal{O}(n^3)$ with n the size of data. (2) Given the background distribution, we find an optimal projection, the complexity of which is dominated by eigenvalue decomposition ($\mathcal{O}(n^3)$). Hence, the overall complexity of SICA is $\mathcal{O}(\frac{n^3}{\varepsilon^2})$.

Note that both traditional spectral clustering [12, 14] and different manifold learning approaches [1, 9, 19] also try to solve a related eigenproblem in order to discover the intrinsic manifold structure of the data, using an eigendecomposition to preserve the local properties of the data. However, differently from our approach, these methods are interested in the eigenvectors corresponding to the smallest k eigenvalues of the Laplacian, as they provide insights into the local structure of the underlying graph, while we are maximizing our objective and therefore, in contrast to other methods, targeting non-smooth solutions.

4. CASE STUDIES

In this section, we demonstrate the use of SICA on both synthetic and real-world data, including images, time-series, and spatial data. We show how the proposed method can effectively encode the user's prior expectations and then discover interesting projections, thereby providing insight into the data that is not apparent when using alternative data exploration techniques.

We compare the behavior of SICA with Principal Component Analysis (PCA), because the latter is a very popular dimensionality reduction method, and because our method also explains the observed variance in the data, although

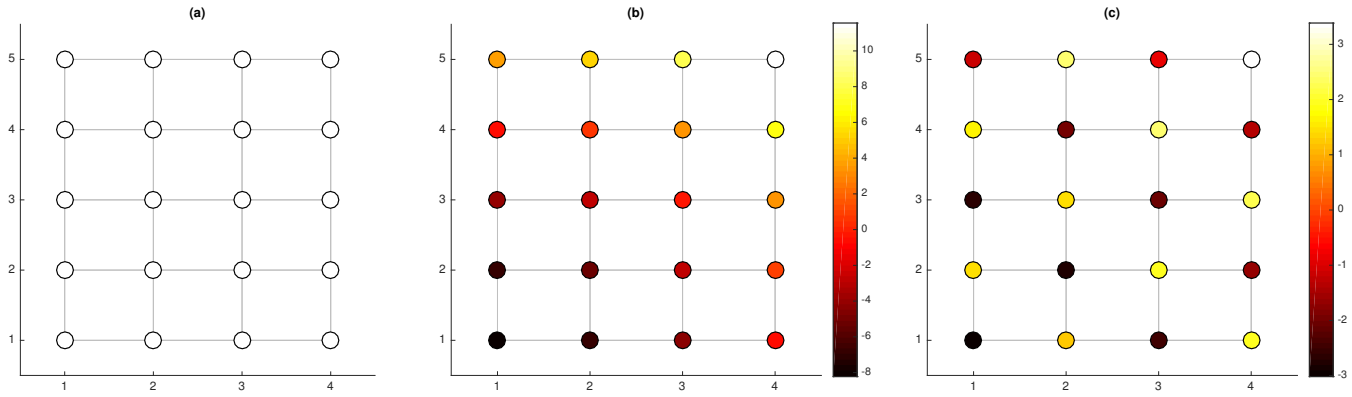

Figure 2: Synthetic Grid data (§4.1), (a) graph representing the user's prior belief, (b) projection onto PCA's first component, and (c) projection onto SICA's first component.

	Feature 1	Feature 2	···
PCA 1st component	-0.995	-0.021	···
SICA 1st component	-0.369	-0.916	···

Table 1: Grid data (§4.1), weights of first component for PCA and SICA.

	Feature 1	Feature 2	···
PCA 1st component	-0.998	0.015	···
SICA 1st component	0.186	0.957	···

Table 2: Communities data (§4.2), weights of first component for PCA and SICA.

traded off with novelty. Hence, PCA is the most related dimensionality reduction method and, as we will see, results of PCA and SICA may also coincide. That happens, for example, if the specified prior expectations are irrelevant.

4.1 Synthetic Grid

Task. As an example, we consider a simple scenario where a user is aware of the main structure in the data and is interested in finding whether any additional structure exists.

Dataset. We generated a dataset of 20 data points with 10 real-valued attributes, i.e., $\hat{\boldsymbol{X}} \in \mathbb{R}^{20\times 10}$. The data points are linked to the vertices of a rectangular grid, where each vertex v is indexed as (l, m), $l \in \{1, \ldots, 4\}$ and $m \in \{1, \ldots, 5\}$. We assume the first attribute of the data varies strongly along one diagonal of the grid, i.e., $\boldsymbol{x}_1(v(l, m)) = 0.5l^2 + 0.5m^2$. As for the second attribute, the values between neighboring vertices alternate between -1 and $+1$ plus Gaussian noise. The remaining features are standard Gaussian noise.

Prior expectation. We assume that the user already knows that there is a smooth variance along the grid. Such knowledge can, for example, be encoded in a graph by connecting adjacent data points, as shown in Figure 2a.

Results. Figures 2b and 2c present the resulting top components from PCA and SICA; the nodes on the grid are colored according to the values after projection. The projection onto the first PCA component varies along the diagonal of the grid, from $(1, 1)$ to $(4, 5)$. This confirms the user's expectations, and hence is not informative. On the other hand, SICA gives a projection that assigns high vs. low scores to every other vertex. This reveals another underlying property of the data, complementing the user's prior beliefs.

Another view on the difference between the PCA and SICA projections is given in Table 1. We observe that PCA assigns a large weight to the first feature, which is the one that varies globally. In contrast, SICA emphasizes the second feature, which is to the feature that changes between the neighboring vertices.

4.2 Synthetic Communities

Task. A user analyzing social network data is typically interested in identification of group structure, i.e., finding groups of people that share certain properties. Suppose that we have studied the network already for a while and become bored with patterns that relate the network structure with attributes that characterize the communities. One may ask *is there anything more?* We show that with SICA, one can encode community structures as prior expectations, and hence find structure orthogonal to the network layout.

Dataset. We synthesized a dataset of information about 100 users, where each is described by 10 features, i.e., $\hat{\boldsymbol{X}} \in \mathbb{R}^{100\times 10}$. The first feature is generated from two Gaussian distributions, where each distribution corresponds to half of the users. The means and variances are chosen such that, according to the first feature, the data can be clearly separated into two communities. To have a more realistic simulation, we also assume that few connections exist between two communities. The second feature is generated by uniformly assigning a value -1 or $+1$ to the users. For instance, this could represent the users' sentiment towards a certain topic. The remaining features are standard Gaussian noise.

Prior expectation. We take as prior expectation the observed network. That means we expect people to be alike if they are connected. The resulting prior knowledge graph consists of two cliques and few edges in-between, see Figure 1a.

Results. To compare the projections given by PCA and SICA, we visualized the projections in Figures 1b and 1c. For both PCA and SICA projections, we colored the data points according to their projected values, i.e., $\boldsymbol{X}\boldsymbol{w}$, where $\boldsymbol{w}$ is the first component of either PCA or SICA. In Figure 1b, we see that PCA's projection gives one cluster a higher score (green vertices) than the other (blue vertices). Clearly,

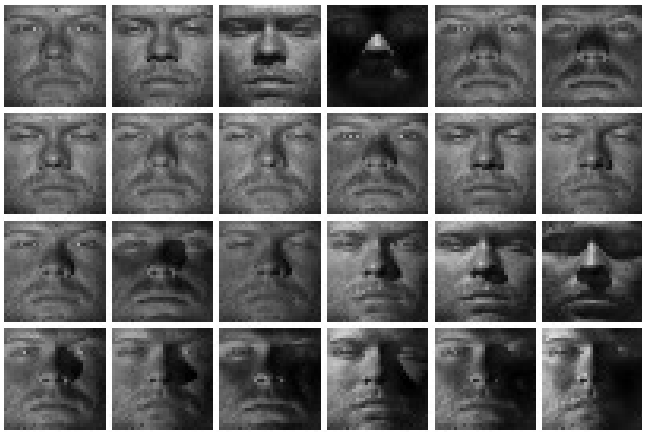

Figure 3: Faces data (§4.3), subject one, first 24 lighting conditions.

PCA reveals the structure of the two communities. In contrast, SICA assigns both high and low scores within each cluster (Figure 1c). That is, it highlights variance *within* the clusters. Table 2 lists the weight vectors of the projections. As expected, PCA gives a large weight to the first feature, which has higher variance. However, SICA's first component is dominated by the second feature. Hence, by incorporating the community structure as prior expectation, SICA finds the alternative community structure corresponding to the second feature.

4.3 Images and Lighting

Task. Consider the problem of learning to recognize faces. As input data, we may have images shot under different conditions (e.g., variable lighting). PCA can be used to find eigenfaces [17]. However, PCA preserves both global features (e.g., global illumination) and local features (e.g., facial structure). If the images are shot under similar conditions, this approach may work well. However, if the conditions vary, for example if the direction of lighting varies, then PCA will mix the variation between faces with the variation between conditions. This may be undesirable and we could prefer to ignore the variation associated with the lighting condition. As illustrated in this section, this may be helpful not only to find more intuitively meaningful eigenfaces, but also to improve the accuracy of face recognition.

Dataset. We studied the Extended Yale Face Database B[3]. The dataset contains frontal images of 38 human subjects under 64 illumination conditions. We ignored the images of seven subjects whose illumination conditions are not specified. We centered the data by subtracting, per pixel, the overall mean. Hence, the input dataset contains 1684 data points, each of which is described by 1024 features.

Prior expectation. We assume that the user already knows the lighting conditions and is no longer interested in them. This knowledge can be encoded into SICA constraints by connecting the graph nodes (images) with the same lighting condition with edges. Hence, the graph of SICA constraints consists of 64 cliques, one for every lighting condition.

[3]A Matlab data file (already preprocessed) is available at http://www.cad.zju.edu.cn/home/dengcai/Data/FaceData.html. The original dataset is described in [6, 11].

Figure 4: Faces data (§4.3), top five eigen faces for PCA (top) and SICA (bottom).

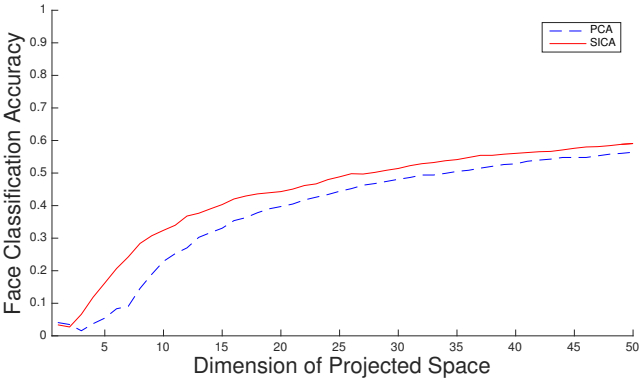

Figure 5: Faces data (§4.3), accuracy of 3-NN subject classification in projected feature spaces.

Results. We expect PCA to find a mixture of illumination and facial features, while SICA should find mainly facial structure. Note that illumination conditions vary across the same human subjects, and facial structures are more *subject specific* features. Intuitively, if we project the image onto the top components of both PCA and SICA, the projection given by SICA would separate the subjects better than the projection produced by PCA. To test this intuition, we computed the accuracy (w.r.t. the subjects as labels) of a k-Nearest Neighbors (k-NN) classifier on the projected features with respect to the top N components of PCA and SICA. A projection that separates the subjects well will have high classification accuracy. We applied k-NN on feature spaces with N ranging from 1 to 50 and $k = 3$. Figure 5 shows that SICA (solid red line) gives a better separation than PCA (dashed blue line) for any number of components.

The top eigenfaces from PCA and SICA are presented in Figure 4. We observe that the eigenfaces given by PCA are influenced substantially by the variation in lighting conditions. These conditions vary from back-to-front, right-to-left, top-to-down, down-to-top and left-top-to-right-bottom. Because the images of each subject contain every lighting condition, it is indeed more difficult to separate the subjects based only on the top PCA components. On the other hand, the eigenfaces from SICA highlight local facial structures like the eye area (first, third and fifth faces), mouth and nose (first, third and fifth faces), female face structure (fourth face), and pupils (third, fourth and fifth faces). Intuitively, these areas indeed correspond to facial structure that could discriminate between individuals. Note though that the first and second SICA faces also pick up some lighting

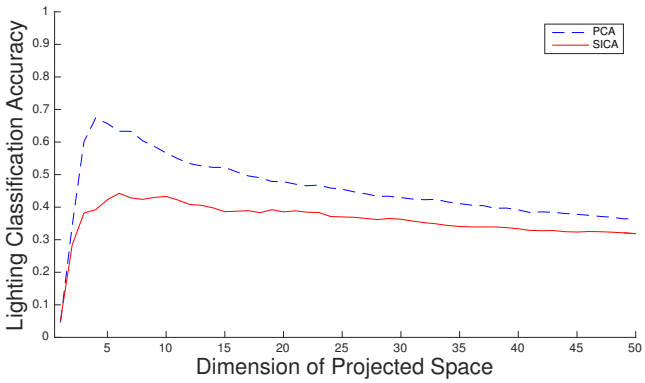

Figure 6: Faces data (§4.3), accuracy of 3-NN lighting condition classification.

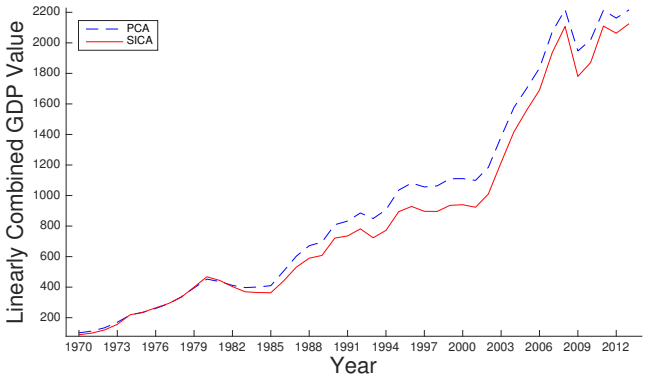

Figure 7: GDP data (§4.4), projections against first PCA and SICA component.

variation, which possibly explains the low accuracy of SICA if we use only the top two components (Figure 5).

Finally, as PCA mainly preserves image structures that correspond to lighting conditions, we suspect that PCA will actually give a better separation in terms of different lighting conditions. To evaluate this, instead of classifying subjects, we try to use k-NN to classify different illumination conditions. Figure 6 shows that PCA indeed gives better separation than SICA. This also strengthens our previous observation that PCA preserves both global variance (illumination conditions) and local structures (facial features), while SICA reveals more local structures.

4.4 World Economy

Task. A fundamental task in time series analysis is to extract global and local characteristics, e.g., trends and events. Again, PCA projections probably reveal both types of features, but potentially mixed so that it is hard to separate the global vs. the local features. However, if a user has prior expectations about one (for example, trends), other features may become more visible. PCA cannot adapt to changes in the user's knowledge about data. However, we expect that SICA can be used to find more surprising features.

Dataset. We compiled GDP per capita (in US Dollars) time series from the World Bank website.[4] By filtering

[4]http://data.worldbank.org/indicator/NY.GDP.PCAP.CD

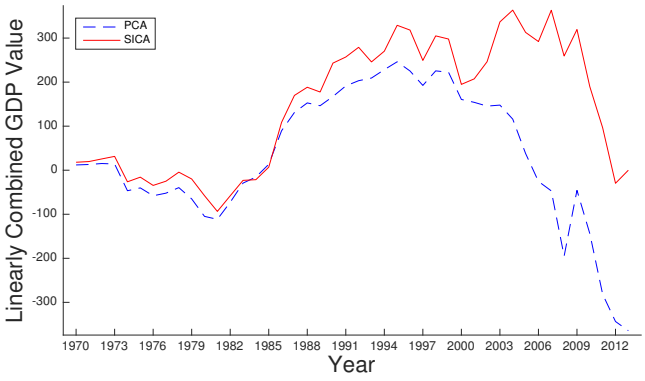

Figure 8: GDP data (§4.4), projections against second PCA and SICA component.

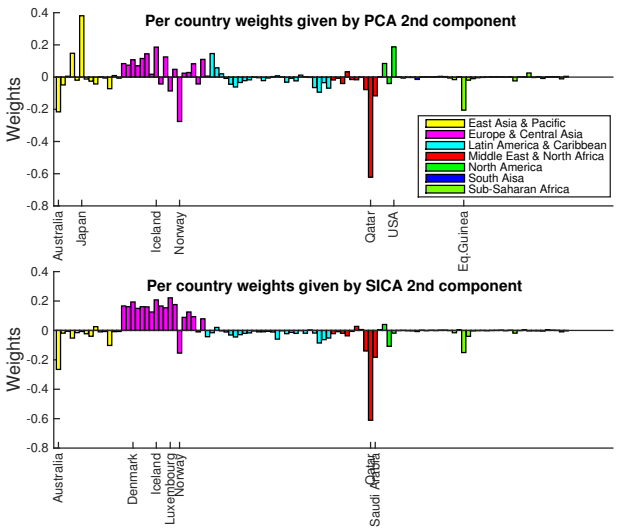

Figure 9: GDP data (§4.4), per country weights given by PCA and SICA second component. The 7 countries with largest absolute weights are marked.

out the countries with incomplete GDP records, the compiled dataset consists of the GDP per capita records of 110 countries from 1970 to 2013. The World Bank categorises countries into seven regions: East Asia & Pacific, Europe & Central Asia, Latin America & Caribbean, Middle East & North Africa, North America, South Asia, Sub-Saharan Africa. So the input data for both methods consists of 44 data points, where each data point (year) is described by 110 features (the GDP per capita for the 110 countries of that specific year). The data is centered, but not standardized.

Prior expectation. GDP values of adjacent years are unlikely to have drastic changes. We can translate this expected local similarity into prior expectations: by treating each time point as a graph node, local similarity can be represented by edges between temporally adjacent time points. The resulting graph is a chain with 44 nodes. By incorporating these prior expectations, we expect SICA to find fluctuations over short time periods, while PCA remains agnostic of time.

Results. A projection of the time series data onto one PCA or SICA component is again a time series. It is essentially

	PCA	SICA
Variance terms	**1.131e+12**	1.023e+12
Non-Smoothness terms	0.967e+10	**1.106e+10**

Table 3: GDP data (§4.4), sum of values of SIC terms w.r.t top four PCA and SICA components.

	CDU/CSU	SPD	FDP	GREEN	Left
PCA 1st	0.53	-0.13	0.22	0.13	-0.80
SICA 1st	0.72	-0.65	0.10	-0.09	-0.19

Table 4: German socio-economics data vote attributes (§4.5), weights given by first PCA and SICA component.

a linear combination of all country's GDP series, where the weights correspond to countries. Since most countries' GDPs are correlated and rising over time (see Figure 7), both top projections given by PCA (dashed blue line) and SICA (solid red line) show simply an overall increase of the GDP (essentially the average GDP series) over the years.

More interesting are the second projections: in Figure 8 it is shown that the projection onto the second SICA component has more local fluctuation, and the projection given by the second PCA component is smoother. Arguably, the difference is not very large. To check whether there is indeed a significant difference between the solutions, we computed the values of the variance ($\boldsymbol{W'X'\lambda IXW}$) and non-smoothness terms ($\boldsymbol{W'X'\mu LXW}$) (in SIC definition 2) over the top four PCA and SICA components. As shown in Table 3, PCA's components give projections with greater variance while the projections of the SICA's components have smaller global variance but more local variances (non-smoothness). The differences are not very large, probably because the growth of most countries is very smooth and there have been few events with large impact.

To investigate the time series in more detail, we considered the time series against a list of financial crises[5]. Note that in Figure 8 there are two sudden drops in 1974 and around 1979, that might well be due to the 1970 energy crisis[6]. In the 1973 crisis, the western countries were heavily affected. The 1979 crisis is caused by the interruption of export from the Middle East and the Iranian Revolution.

According to the bar charts depicting the weight vectors (Figure 9), PCA assigned positive and negative weights to different regions while SICA gives positive weights majorly to countries in Europe & Central Asia and negative weights to countries in the Middle East & North Africa. It is quite remarkable that among all positively-weighted European & Central Asian countries, Norway, which is the major fossil fuel producer in Europe, is also assigned a large negative weight by SICA, similar to the Middle Eastern countries. The same holds for Australia, which is the world's second largest coal exporter. So, rather than experiencing the fuel crisis as the other Western and East Asian countries, these two countries benefited from it, which we found by studying the projections and weight vectors from SICA.

4.5 Spatial Socio-Economics

Data. The German Socio-economic dataset [2] was compiled from the database of the German Federal Office of Statistic. The dataset consists of socio-economic records of 412 administrative districts in Germany. The features can be divided into three groups: election vote counts, workforce distribution, and age demographics. We additionally coded for each district the geographic coordinates of the district center and which districts share a border with each other.

[5]https://en.wikipedia.org/wiki/Financial_crisis

[6]https://en.wikipedia.org/wiki/1970s_energy_crisis

	Elderly	Old	Mid-Age	Young	Child
PCA	-0.61	-0.42	0.43	0.09	0.51
SICA	-0.62	-0.32	0.69	0.19	0.06

Table 5: German socio-economics data age demographics (§4.5), weights given by first PCA and SICA component.

4.5.1 Vote Attribute Group

Task. We are interested in exploration of the voting behavior of different districts in Germany. We already know that the traditional East-West divide is still a large influence, and would like to find patterns orthogonal to that division.

Dataset. The attributes in this group come from the 2009 German elections. It covers the five largest political parties: CDU/CSU, SPD, FDP, GREEN, and LEFT. We centered the data by subtracting the mean from each data point. Since the values of the five features add up to a constant, the data is not standardized.

Prior expectation. We assume the voting behavior of the districts in east Germany are similar to each other, and so do the remaining districts. By treating each district as a graph node, we can translate our knowledge into prior expectation by connecting similar districts by edges. This results in a graph with two cliques: one clique consists of all districts in east Germany, the other clique contains the rest.

Results. The projection onto the first PCA component (Figure 10a) shows a large global variance across the map. Districts in west Germany and Bavaria (south) receive high scores (red circles) and districts in east Germany (Brandenburg and Thuringa) have low scores (dark blue circles). Table 4 additionally shows the weight vectors of the first components of PCA and SICA. The first PCA component is dominated by the difference between CDU/CSU and Left. This is expected, because this indeed is the primary division in the elections; eastern Germany votes more Left, while in Bavaria, CSU is very popular.

However, SICA picks up a different pattern; the fight between CDU/CSU and SDP is more local. Although there is still considerable global variation (in this case between the south and the north), we also observe that the Ruhr area (Dortmund and around) is similar to eastern Germany in that the social-democrats are preferred over the Christian parties. Arguably, the districts where this happens are those with a large fraction of working class, like the Ruhr area. It is perhaps understandable that they vote more leftist, e.g., they vote for parties that put more emphasis on interests of the less-wealthy part of the population.

4.5.2 Demographic Attribute Group

Task. We are interested in exploration of the age demographics of different districts in Germany. Again, we already know that the traditional East-West divide is still a large

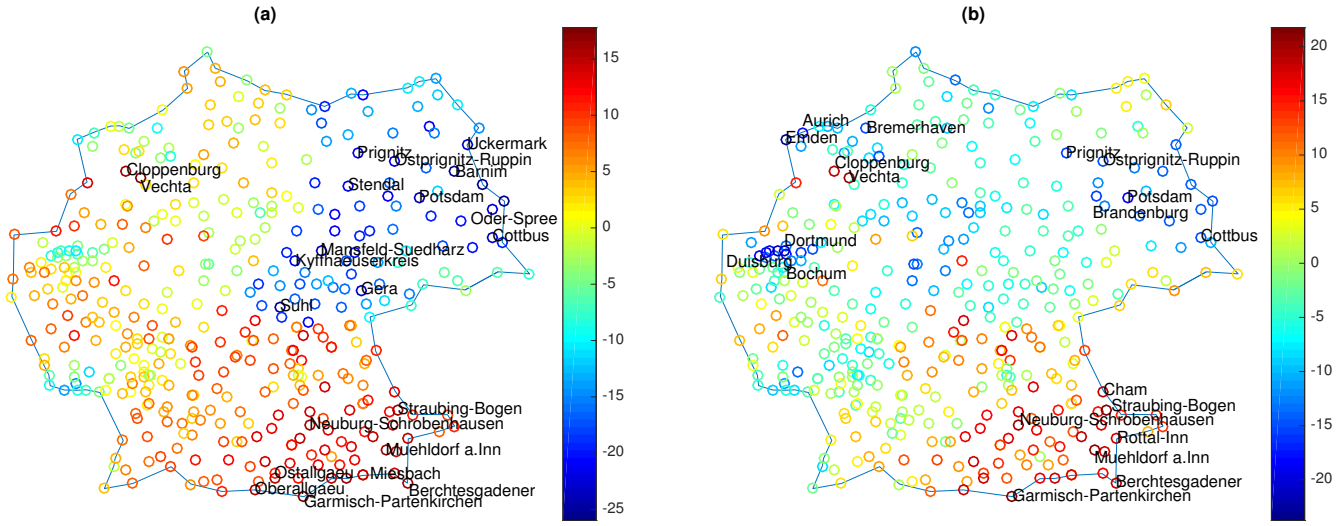

Figure 10: German socio-economics data (§4.5), (a) vote attributes, scatter plot onto first PCA component, (b) vote attributes, scatter plot onto first SICA component. The top 10 districts with most positive and most negative weights are labeled.

influence, although for somewhat different reasons. We are interested in finding patterns orthogonal to that division.

Dataset. The demographic attribute group describes the age structure of the population (in fractions) for every district. There are five attributes: *Elderly People* (age > 64), *Old People* (between 45 and 64), *Middle Aged People* (between 25 and 44), *Young People* (between 18 and 24), and *Children* (age < 18). Since the values of the five features add up to a constant, the data is not standardized.

Prior expectation. Due to historical reasons, the population density is lower in east Germany than the rest of country. According to Wikipedia[7]: "1.7m (12%) people left new federal states since fall of the Berlin Wall, [...], high number of them were women under 35". Also Berlin-Institute for Population and Development[8] reports: "the birth rate in east Germany dropped down to 0.77 after unification, and raised to 1.30 nowadays compare to 1.37 in the West". Given this (in Germany common sense) knowledge, we would like to find out something surprising. Hence, we assume again that the demographics of the districts in east Germany are similar, and the remaining districts are also similar. This results in a graph with two cliques: one consists of all districts in the east Germany, another one contains the rest.

Results. Projection of the first PCA component confirms our prior expectations. Figure 11a shows that there is a substantial difference between the east and west of Germany. In the projections, high values (red color) are assigned to the cities in east Germany, while low values (blue color) are given to the rest of Germany. If we look at the weights for the first PCA component (Table 5), we find the projection is based on large negative weights to people above 44 (old and elder), and large positive weights to the younger population (age $<$ 45). This confirms that indeed the demographic status of east Germany deviates from the rest of the country.

[7]https://en.wikipedia.org/wiki/New_states_of_Germany#Demographic_development

[8]http://www.berlin-institut.org/fileadmin/user_upload/Studien/Kurzfassung_demografische_lage_englisch.pdf

As opposed to PCA, SICA gives an alternative projection, see Figure 11b. The difference is more subtle as in the analysis of the voting behavior. Although SICA also assigns large negative scores to east Germany, because there are relatively many elderly there, SICA also highlights the large cities, e.g., Munich, Cologne, Frankfurt, Hamburg, Kiel, Trier. Rather than just showing the global trend, the result from SICA picks out districts whose demographic status deviates from this surrounding districts. Indeed, from the weight vector (Table 5) we see that these districts are found by consedering the number of middle aged people against the number of elderly. We know that many middle-aged $(24 - 44)$ working people live in large cities, and, according to the report from Berlin-Institute for population and Development, "large cities generally have fewer children, since they offer families too little room for development". Indeed, we find that families live in the neighboring districts, highlighting a perhaps less-expected local contrast.

4.6 Summary

We have found that SICA enables us to find alternative projections that are more interesting given the specified prior expectations. Often, the difference with PCA is large, but sometimes, e.g., in the GDP time-series case, one may find that the data contains little variation besides the main trend, in which case the methods produce similar results.

5. CONCLUSIONS

PCA is one of the most popular dimensionality reduction techniques, comparing favourably with non-linear approaches in many real-world tasks [18]. However, if we are aware of a user's prior expectations, PCA is suboptimal for finding interesting projections. To address this, we presented SICA, a new linear dimensionality reduction approach that explicitly embraces the subjective nature of interestingness. In this paper, we showed how to encode the prior expectations as constraints in an optimization problem that can be solved in a similar manner to PCA. Results from

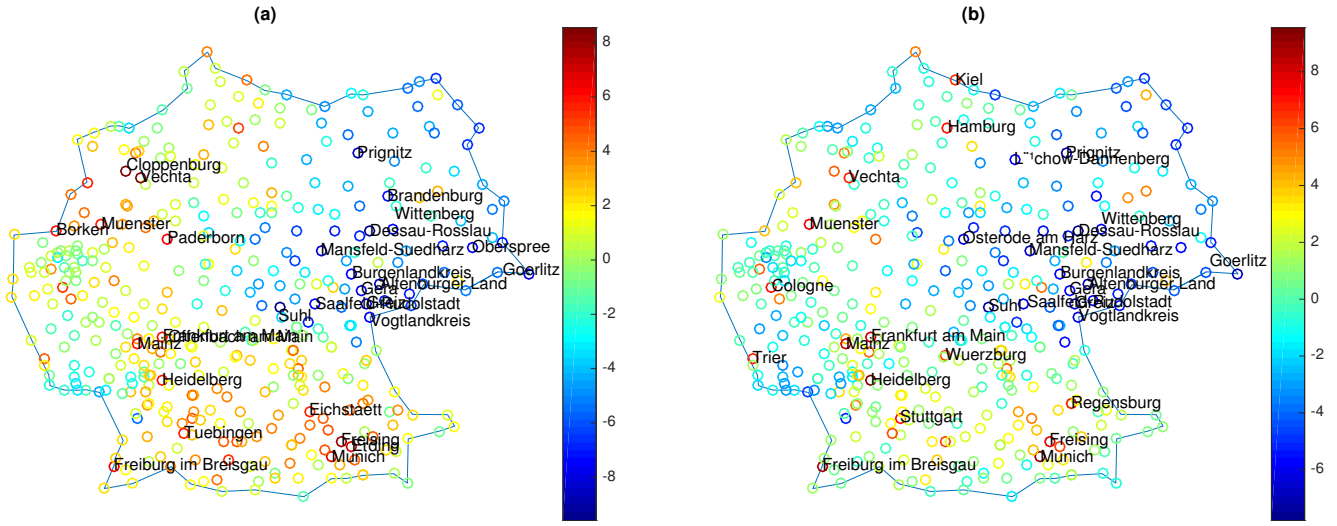

Figure 11: German socio-economics data (§4.5), (a) demographic attributes, scatter plot against first PCA component, (b) demographic attributes, scatter plot against first SICA component. The top 10 districts with most positive and most negative weights are labeled.

several case studies show clearly that it can be meaningful to account for available prior knowledge about the data.

A potentially interesting avenue for further work is to incorporate multiple prior expectations simultaneously, to enable more flexible iterative analysis. This involves solving a MaxEnt optimization problem subject to multiple graph constraints. We also plan to study graded similarities between data points. Such prior beliefs result in a graph with weighted edges. Although this is technically already possible, the question is how a user can conveniently input these expectations into the system. Finally, alternative types of prior expectations are also worth examining.

Acknowledgements

This work was supported by the European Union through the ERC Consolidator Grant FORSIED (project ref. 615517) and by EPSRC grant DS4DEMS (EP/M000060/1).

6. REFERENCES

[1] M. Belkin and P. Niyogi. Laplacian eigenmaps for dimensionality reduction and data representation. *Neural Comput.*, 15(6):1373–1396, 2003.

[2] M. Boley, M. Mampaey, B. Kang, P. Tokmakov, and S. Wrobel. One click mining: Interactive local pattern discovery through implicit preference and performance learning. In *Proc. of KDD*, pages 27–35, 2013.

[3] T. De Bie. An information theoretic framework for data mining. In *Proc. of KDD*, pages 564–572, 2011.

[4] T. De Bie. Subjective interestingness in exploratory data mining. In *Proc. of IDA*, pages 19–31, 2013.

[5] T. De Bie, J. Lijffijt, R. Santos-Rodríguez, and B. Kang. Informative data projections: A framework and two examples. In *Proc. of ESANN*, to appear.

[6] A. Georghiades, P. Belhumeur, and D. Kriegman. From few to many: Illumination cone models for face recognition under variable lighting and pose. *IEEE PAMI*, 23(6):643–660, 2001.

[7] A. K. Gupta and D. K. Nagar. *Matrix Variate Distributions*. CRC Press, 1999.

[8] D. J. Hand, H. Mannila, and P. Smyth. *Principles of Data Mining*. MIT Press, 2001.

[9] X. He and P. Niyogi. Locality preserving projections. In *Proc. of NIPS*, 2003.

[10] E. Kokiopoulou, J. Chen, and Y. Saad. Trace optimization and eigenproblems in dimension reduction methods. *Num. Lin. Alg. Applic.*, 18(3):565–602, 2011.

[11] K.-C. Lee, J. Ho, and D. J. Kriegman. Acquiring linear subspaces for face recognition under variable lighting. *IEEE PAMI*, 27(5):684–698, 2005.

[12] U. V. Luxburg. A tutorial on spectral clustering. *Stat. Comp.*, 17(4):395–416, 2007.

[13] Y. Nesterov. *Introductory lectures on convex optimization: A basic course*. Springer, 2004.

[14] A. Y. Ng, M. I. Jordan, and Y. Weiss. On spectral clustering: Analysis and an algorithm. In *Proc. of NIPS*, pages 849–856, 2001.

[15] L. K. Saul and S. T. Roweis. Think globally, fit locally: Unsupervised learning of low dimensional manifolds. *J. Mach. Learn. Res.*, 4:119–155, Dec. 2003.

[16] J. B. Tenenbaum, V. de Silva, and J. C. Langford. A global geometric framework for nonlinear dimensionality reduction. *Science*, 290(5500):2319, 2000.

[17] M. Turk, A. P. Pentland, et al. Face recognition using eigenfaces. In *Proc. of CVPR*, pages 586–591, 1991.

[18] L. J. P. van der Maaten, E. O. Postma, and H. J. van den Herik. Dimensionality reduction: A comparative review. Technical Report TiCC TR 2009–005, Tilburg University, 2009.

[19] K. Q. Weinberger, F. Sha, Q. Zhu, and L. K. Saul. Graph Laplacian regularization for large-scale semidefinite programming. In *Proc. of NIPS*, pages 1489–1496, 2006.