Generating Low-Dimensional Denoised Representations of Nonlinear Data with Superparamagnetic Agents

Thomas Ott, Zurich University of Applied Sciences

Thomas Eggel, Zurich University of Applied Sciences

Markus Christen, University of Zurich

Overview

The Problem: Visualizing high-dimensional data spaces

The conceptual idea of superparamagnetic agents

Level 1: Spin System

Level 2: Agent System

Toy Examples

Real World Example

Open Questions

The Problem – Visualizing high-dimensional data

Visualization of high-dimensional data by means of a low-dimensional embedding plays a key role in explorative data analysis. Classical approaches for dimensionality reduction aim to represent the data structure on a linear subspace of the original data space:

- **PCA** performs a projection onto the axes with maximal data variance
- **MDS** finds a low-dimensional embedding that preserves interpoint distances

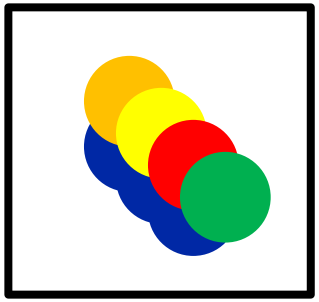
Problems:

- These methods perform poorly when applied to nonlinear data structures.
- For many real-world applications, researchers are faced with similarity or proximity data with correct ordering, but potentially unreliable data values.

In this contribution, we present a novel approach that is able to deal with nonlinear structures in data space and that includes a mechanism to reduce the distraction by noise.

The Conceptual Idea of Superparamagnetic Agents

- The core idea of our approach is to translate the data into a set of agents.
- These agents ’construct’ a low-dimensional representation of the data in a self-organized way by moving according to laws of local spin interactions.



«Approach those who you like»

«Avoid those who you dislike»

Level 1 – Spin System (1)

We assume a given data set with *n* data points and its corresponding dissimilarity matrix with values $g_{ij} = g_{ji}$. Our method can be divided into two levels.

1. In the first level, each data item is represented by a Potts spin variable and the dissimilarity matrix is encoded in the spin couplings J_{ij} (k: k-nearest neighbors; a: average distance between neighbors):

$$J_{ij} = J_{ji} = \frac{1}{k} \exp\left(\frac{-g_{ij}^2}{2a^2} \right)$$

2. The spin system is treated in the formalism of the canonical ensemble, giving the probability for a certain spin configuration s_i as follows:

$$p(s) = \frac{1}{Z} \exp(-H(s)/T) \qquad H(s) = \sum_{(i,j)} J_{ij}(1 - \delta_{s_i s_j})$$

Level 1 – Spin System (2)

3. By introducing a temperature-like parameter *T*, a cluster hierarchy can be generated. For smaller *T*, all spins tend to be in the same state. Upon an increase in *T*, large clusters break up into smaller clusters in a cascade of (pseudo-)phase transitions. For small *T*, spins that belong to data items of a noisy background can be filtered out as singletons that do not cluster (*M*: number of Monte Carlo Steps).

$$G_{ij} = \frac{1}{M} \sum_{t=1}^{M} \underbrace{\delta_{s_i^t s_j^t}}_{G_{ij}(t)}$$

The result the calculations of level 1 is a specific spin configuration. It serves as input for the calculation on levels 2 that moves the points representing data on the plain. After the calculations on level 2, a new spin configuration is calculated.

Level 2 – Agent System (1)

In the second level, each data item is represented by an agent in a 2-dimensional coordinate system and the agents move according to laws that are governed by the local interactions of the spin system. The algorithmic procedure is summarized as follow:

1) Choose a random distribution of agents in $\mathbb{R}^2$, a random spin configuration s^0 and set the temperatures $T=T_{min}$, T_{max} (both in the superparamagnetic phase) as well as ΔT (in dependence of M).

2) For T, calculate a new spin configuration s^{t+1} (Swendsen-Wang algorithm) and then the actual pair correlation $G_{ij}(t+1)$

3) Calculate the pairwise attraction / repulsion of agents and relocate them (see next page)

4) Repeat the procedure starting from step 2 until $T=T_{max}$.

Level 2 – Agent System (2)

- If $G_{ij}(t+1) = 1$ and $J_{ij} > 0$ then

$$\begin{aligned}\vec{x_i^{t+1}} &= \vec{x_i^t} + \alpha \cdot (\vec{x_j^t} - \vec{x_i^t}) \\ \vec{x_j^{t+1}} &= \vec{x_j^t} + \alpha \cdot (\vec{x_i^t} - \vec{x_j^t})\end{aligned}$$

- else

$$\begin{aligned}\vec{x_i^{t+1}} &= \vec{x_i^t} + \beta \cdot e^{-d_{ij}^t} \cdot (\vec{x_i^t} - \vec{x_j^t}) \\ \vec{x_j^{t+1}} &= \vec{x_j^t} + \beta \cdot e^{-d_{ij}^t} \cdot (\vec{x_j^t} - \vec{x_i^t})\end{aligned}$$

where $d_{ij}^t = |\vec{x_i^t} - \vec{x_j^t}|$.

Methodological remarks

(Optional) noise cleaning is performed by removing agents whose spins are in no clusters even for T_{min} (calculated in step 1).

Usually, the procedure is repeated for several temperatures *T* and then the mean location of the points is calculated.

The method does not offer unique solutions, which highlights the importance of the parameters involved: $0 < \alpha < 0{,}5$ controls attraction, $0 < \beta$ controls repulsion. Simulations show that α and β strongly determine the scaling of the final agent configuration:

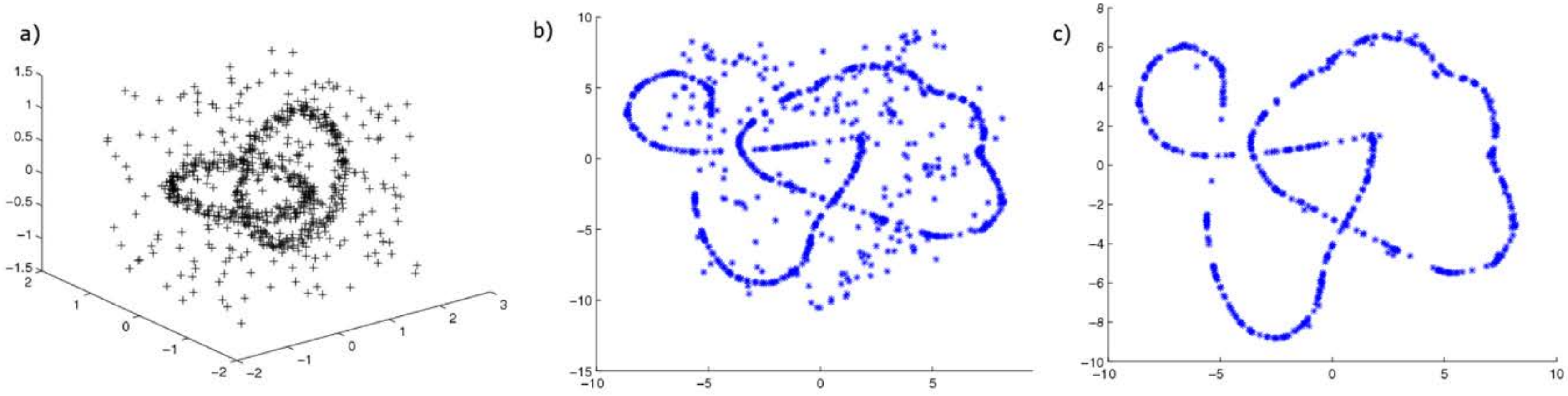
- α mainly affects the intra-cluster distances.
- β mainly affects the inter-cluster distances.

For the examples in this paper we used the values α = 0.1 and β = 0.01 that have proven useful to balance inter- and intra-cluster distances.

Toy Example 1

The example consists of two interlocked rings on a noisy background.

a) Original data set in 3D
b) Superparamagnetic agent map (SAM) without noise cleaning
c) SAM with noise cleaning



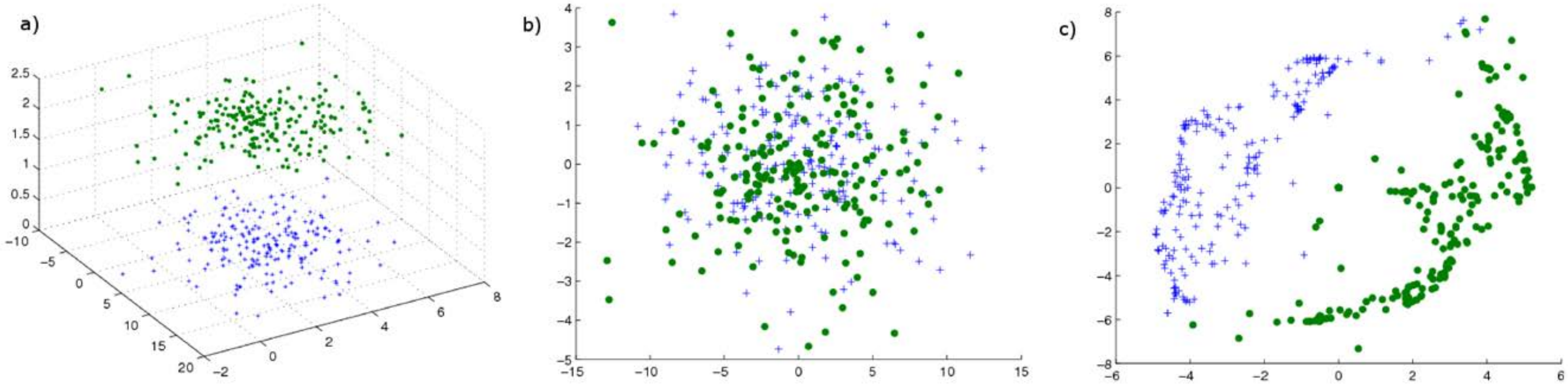
Toy Example 2

The data set of this example consists of two Gaussian clusters with 200 points each and means $\mu_1 = (0, 0, 0)$ and $\mu_2 = (0, 0, 2)$

a) Original data set in 3D

b) PCA solution: the clusters are invisible in 2D because the extension in the x- and y-direction is larger than in z-direction.

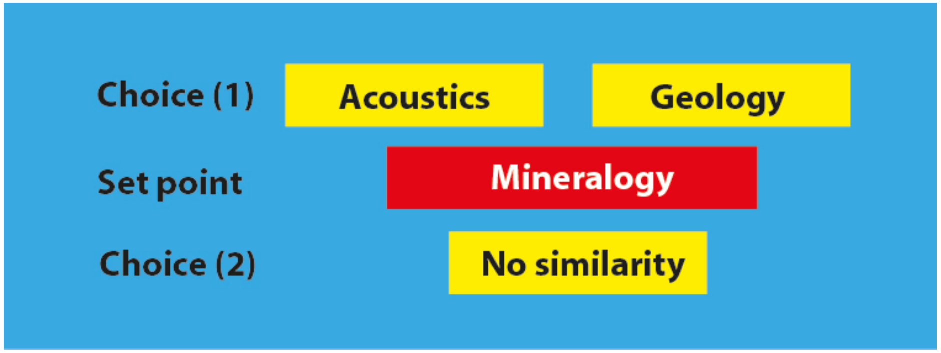
c) SAM solution: The clusters are clearly distinguable.



Real World Example – The Experiment

We use data from a survey on the similarity of 249 scientific disciplines represented as subject categories that classify journals contained in the citation indexing and search service Web of Science provided by Thomson Reuters.

In the internet survey, the participants were presented with subject category X (including short text) as well as two other categories Y and Z and they had to choose to which discipline X is more similar. The task is robust for sequence effects and allows that subjects can stop the survey whenever they like.



Choice (1) Acoustics Geology

Set point Mineralogy

Choice (2) No similarity

Real World Example – The Data

876 researchers from all disciplines have been approached in multiple ways (e.g., via scientific associations) and they provided 33’558 assessments of the similarity of such subject category triplets, leading to a similarity matrix.

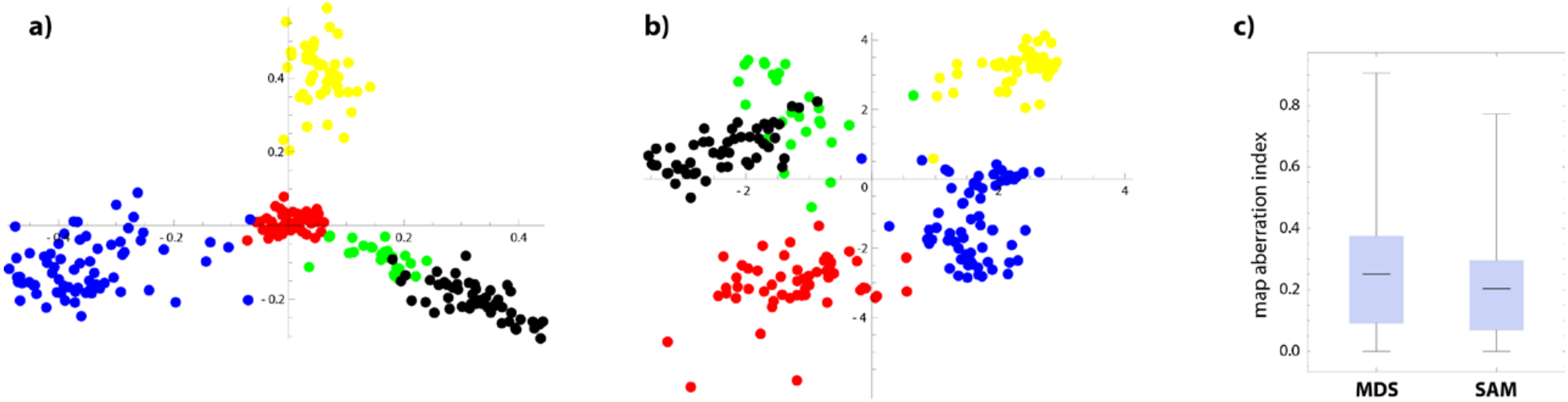
To manage combinatorial explosion, we presuppose that disciplines from the same main fields (engineering, humanities, medicine, (social) science) are considered to be more similar when compared to a discipline from another field; i.e. participants that relate themselves to a specific field obtain random triplets where 90% emerge from ”their” field.

As similarity measure we use the ratio of positive attributions of two disciplines X and Y compared to the total number of possibilities to attribute X with Y.

Real World Example – Comparing MDS-SAM (1)

Both approaches display a similar cluster discernibility, but the topology of the original space is less well preserved in MDS compared to SAM.

We calculated for each item the sum of the absolutes of the normalized distance differences for each pair (original space vs. map space). The smaller the mean of this distribution (map aberration index), the better does the map preserve the topology of the original space.

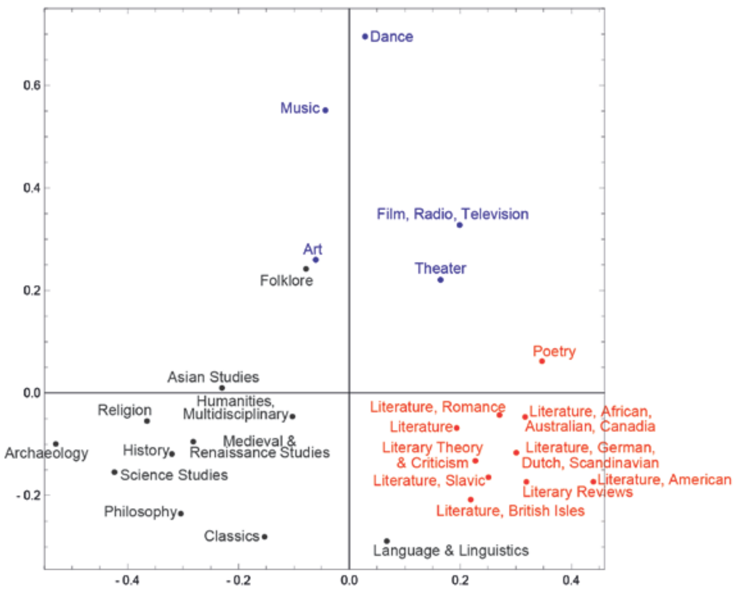


Blue: science; red: engineering; yellow: medicine; green: humanities; black: social science

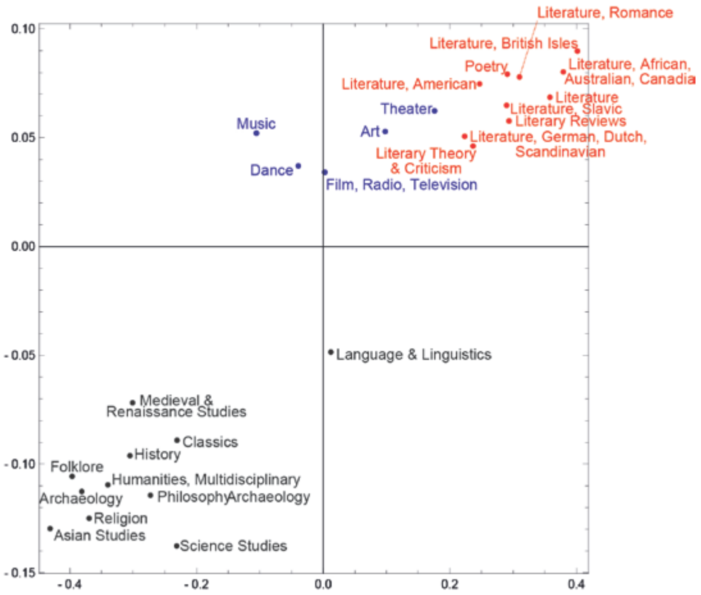
Real World Example – Comparing MDS-SAM (2)

When comparing the visualizations of the data for the group ”humanities”, for which most data was achieved in the survey, we find that SAM provides a more plausible representation:

MDS



SAM



Open Questions

Although the heuristic superparamagnetic agents algorithm was successful in several applications, questions remain regarding the theoretical understanding:

- How can we quantify the role of the parameters *α* and *β*?

- How can the theoretical connections to other methods such as nonmetric multidimensional scaling be elaborated?

- Can we also use the technique to determine the true dimensionality of higher-dimensional data structures?

- What other rules or clustering methods could be used instead of our heuristics to generate a low-dimensional representation?

Thank you!