ISSN 1213-6972 (print)
ISSN 1213-6964 (online)
Journal of WSCG
http://www.wscg.eu
Vol.33, No-1-2, 2025

Efficient Regularization-based Normalization for Interactive Multidimensional Data Analysis Without Scaling Artifacts

Vladimir Molchanov
University of Münster
Einsteinstraße 62
48149 Münster, Germany
molchano@uni-muenster.de

Hennes Rave
University of Münster
Einsteinstraße 62
48149 Münster, Germany
hennes.rave@uni-muenster.de

Lars Linsen
University of Münster
Einsteinstraße 62
48149 Münster, Germany
linsen@uni-muenster.de

ABSTRACT

Attribute values in multidimensional datasets often have different measurement units, making data normalization an essential preprocessing step for visualization algorithms such as multidimensional data projections. However, existing normalization techniques are often sensitive to noise, rely on specific data models, are computationally expensive, or have other limitations. The state-of-the-art method for computing optimal scalings of multidimensional data attributes is based on Lloyd relaxation in a linearly projected space. However, its high computational complexity hinders its applicability to datasets of moderate or large sizes. We overcome this limitation by efficiently regularizing the distribution of projected samples using integral images. Our method reduces scaling-induced artifacts, leading to more reliable multidimensional data analysis. In numerical experiments, we demonstrate that our approach, generally, outperforms state-of-the-art methods in computation time, scalability, accuracy, and stability.

Keywords

Multidimensional data visualization, linear projection, data normalization, attribute scaling

1 INTRODUCTION

Visualizations can represent complex, abstract, and large datasets in a form that can be efficiently perceived, interpreted, and understood by humans. Visual representations serve both as a means of displaying and transmitting information and as an essential step in data mining and analysis pipelines. Consequently, visualization algorithms are commonly integrated into many data processing tools.

Data visualization pipelines commonly incorporate data transformations. Improper transformations may lead to information loss, distortion, or misleading interpretations. Therefore, extensive research has been conducted over decades to analyze the limitations and applicability of existing visualization methods, develop advanced data-, domain-, and task-specific approaches, and propose quality control mechanisms.

In the visualization field, multidimensional data projections are dimensionality reduction methods that map the multidimensional data to a, commonly, two-dimensional visual domain. Since projections are generally lossy transformations, a large variety of projection methods exist, each aiming to optimally preserve different characteristics of the original multidimensional data. Understanding the properties of a given projection method allows for identifying patterns in the projected layout and relating them to features of the original multidimensional data.

Basic low-dimensional patterns include clusters and outliers. The detection of such patterns in a projection can be hindered by various visualization effects, such as overplotting [ED07, MG13, RGE19] and poor visual contrast. However, even when a pattern is visually detected, it remains uncertain whether this low-dimensional feature represents an intrinsic structure of the multidimensional data or is an artifact introduced by the applied projection. One potential source of erroneous patterns in the projection domain is the scaling of data attributes.

Attributes of multidimensional data are often different in nature, meaning attribute values are provided in different units and, consequently, have different value ranges. Since units of measurement can be arbitrarily changed, large absolute values of any attribute can be rescaled to small values and vice versa. Converting units, such as from Celsius to Fahrenheit or Kelvin, constitutes a linear transformation of data attributes. Thus, the set of values for each attribute is defined up to an unknown linear mapping.

There are standard approaches for scaling multidimensional data before applying dimensionality reduction

techniques such as *range scaling* or *whitening*. Range scaling linearly maps the range of attribute values for each attribute to a common interval such as $[0,1]$ or $[-1,1]$, thus making values of all attributes comparable in magnitude. This allows for better control over each dimension's contribution to the resulting projection. However, this type of scaling is highly sensitive to noise and outliers, potentially leading to unstable calculations, especially when projecting temporally varying data. Whitening normalizes data such that the standard deviation of each attribute is equal to unity. This method is less affected by noise and outliers but assumes that all attributes have similar statistical properties – a restrictive condition. Consequently, applying whitening to general multidimensional data may introduce projection artifacts.

Recently, Lehmann and Theisel [LT18] proposed a technique for computing optimal linear scalings of multidimensional data. Their approach searches for scalings that result in a projected point layout that is as regular as possible, ensuring that remaining detectable structures are inherent to the data rather than artifacts of attribute scaling. This method employs the Lloyd relaxation algorithm to construct a target regularized layout and adapts the scaling coefficients to approximate the target distribution optimally. While the method makes no assumptions about data statistics, it is computationally intensive, and interactive rates are only achievable through data sampling. Additionally, it is limited to projections into a two-dimensional visualization domain.

In this paper, we propose an efficient method for computing linear scalings of multidimensional data with minimal scaling artifacts. Given a linear projection, we construct a regularized projected data distribution using the integral image-based approach developed by Rave et al. [RML25]. Following the idea of Lehmann and Theisel [LT18], we compute scaling coefficients that ensure that the projection layout closely approximates a regularized distribution. Utilizing *Integral Images* (InIms) for computing regularized distributions overcomes limitations and significantly improves the state-of-the-art *LloydRelaxer* approach by Lehmann and Theisel [LT18]. Our contributions can be summarized as follows:

(1) We adapt the InIm-based approach by Rave et al. [RML25] for computing optimal scalings of multidimensional data.

(2) We adapt a regularity measure for 2D point distributions to assess the quality of the resulting projections.

(3) We perform numerical tests comparing our proposed method with the state-of-the-art in terms of computational efficiency, regularization quality, stability within interactive usage scenarios, and scalability with respect to data size.

2 RELATED WORK

Multidimensional data projections. Projections are dimensionality reduction methods aimed at visualizing data in a low-dimensional domain. However, projections typically introduce distortions, limiting the reliability of analyzing original multidimensional data through projection layouts. Prominent projection methods mitigate distortion of specific data characteristics, making the choice of a suitable projection method dependent on the analysis task.

Linear projection methods are preferred for interactive analysis of large datasets due to their low computational cost and algorithmic simplicity. Principal Component Analysis [Jol86] maximizes variance along two orthogonal directions in multidimensional space. For labeled data, Linear Discriminant Analysis [Fis36, McL04] optimizes class separation in the projection domain. *Star Coordinates* (SC) [Kan00, Kan01] enable interactive exploration of the entire space of linear projections, providing insight into the influence of each data attribute on the projected sample distribution. However, arbitrary linear projections may severely distort data structure in the projected view. Lehmann and Theisel [LT13] proposed Orthographic Star Coordinates, which restrict arbitrary linear projections to orthographic ones. Molchanov and Linsen [ML19] relaxed the Orthographic Star Coordinates constraints to improve computational efficiency. The resulting *Shape-Preserving Star Coordinates* project multidimensional spheres as discs. A subsequent study [MHL20] introduced an efficient morphing technique for shape-preserving SC.

Non-linear projection methods provide greater accuracy in reflecting multidimensional data properties but at the cost of higher computational complexity and limited scalability. Multidimensional Scaling [BG10] optimally preserves sample similarities based on Euclidean distances in the multidimensional domain. *Isomap* [TSL00] employs geodesic distances derived from neighborhood graphs to evaluate sample similarities. T-Distributed Stochastic Neighbor Embedding (t-SNE) [vdMH08] focuses on preserving local neighborhood relationships, though distances in the projection domain do not necessarily correspond to those in the original space. Uniform Manifold Approximation and Projection (UMAP) [MHSG18] also aims at preserving local neighborhoods but manifold structures, as well.

Pagliosa et al. [PPM+15] introduced the *Projection Inspector*, an approach that enables users to explore multiple projection techniques and visually compare the resulting projected layouts. Sacha et al. [SZS+17] developed a classification of interactive visualization methods and systems integrating dimensionality reduction algorithms.

Scaling. The selection of a scatterplot's aspect ratio can be viewed as an a posteriori scaling of projected 2D data. Fink et al. [FHSW13] proposed selecting the aspect ratio based on Delaunay triangulation properties. More recently, Wang et al. [WWF+19] constructed a scatterplot density field using anisotropic kernels and selected an optimal aspect ratio to optimize the density field's properties.

Scaling multidimensional data is a crucial preprocessing step, as an improper choice can significantly impact analysis and is difficult to compensate for, even with an appropriate projection method. Common approaches include mapping each attribute's range to a fixed interval or whitening the data. Lehmann and Theisel [LT18] recently proposed a general, model-free approach for computing scaling coefficients. They employed the Lloyd relaxation algorithm to construct a regularized version of the projected layout, selecting the optimal scaling to approximate a uniform distribution, see Section 3.3 for details.

Uniformity of sampling and distributions is a key characteristic of many mathematical algorithms. For example, quasi-Monte Carlo methods for numerical integration and optimization rely on uniformity in pseudo-random sequences. Machicao et al. [MNM+21] recently evaluated the quality of such generators using a Markov-chain approach. Ong et al. [OKO12] studied uniformity in 2D point distributions, identifying three types of uniformity measures: discrepancy, point-to-point, and volumetric uniformity. They reviewed seven existing methods and proposed a novel approach based on potential energy analogies. In our study, we apply this measure to evaluate the quality of computed scaling by assessing the regularity of resulting projection distributions, see Section 3.2 for details.

Various approaches address *density regularization* in data visualization applications. Scatterplot occlusion reduction can be achieved using image-based approaches, such as pixel mapping [RGE19], or smooth interactive transformations of the scatterplot domain using InIms, as proposed by Rave et al. [RML25]. We adapt the latter approach for efficient computation of regularized layouts, discussed in detail in Section 3.3.

InIms represent precomputed sums of rasterized 2D density functions over rectangular domains. Originally introduced by Crow [Cro84] and later by Viola and Jones [VJ02] for object detection in image analysis, InIms were applied by Rave et al. [RML25] to assess imbalance in discrete density distributions over scatterplot domains. At each discrete location, a displacement vector is computed, mapping the spatial domain accordingly. Iterative application of this transformation converges to a nearly uniform density distribution of projected samples.

3 FOUNDATIONS

In this section, we introduce the necessary notations and provide essential background information on the methods and concepts that form the foundation of our proposed approach, i.e., scaling in linear multidimensional data projections (Section 3.1), uniformity measures of 2D distributions (Section 3.2), and regularization of projection outcome (Section 3.3). For easy reference, we adopt the notations from Lehmann and Theisel [LT18] whenever possible.

3.1 Scaling in Linear Projections

Let n be the dimensionality of the data space and m the number of data points. The j-th point in the data space is represented as a vector $\mathbf{d}_j = (d_{j1}, \ldots, d_{jn})^T$, and the entire multidimensional dataset can be expressed as an $n \times m$ matrix $\mathbf{D} = (\mathbf{d}_1, \ldots, \mathbf{d}_m)$.

Any linear projection operator mapping $\mathbb{R}^n$ to $\mathbb{R}^2$ can be represented by a $2 \times n$ matrix $\mathbf{A} = (\mathbf{a}_1, \ldots, \mathbf{a}_n)$. The projection of a data point $\mathbf{d}_j$ is given by $\mathbf{p}_j = \mathbf{A} \cdot \mathbf{d}_j$, where the set of projected samples forms the $2 \times m$ matrix $\mathbf{P}$.

Linear normalization of data attributes involves a combination of scaling and translation operations. As noted by Lehmann and Theisel [LT18], translating multidimensional data attributes does not alter the relative positions of the projected samples $\mathbf{P}$. Consequently, translation has no effect on data patterns in the projection domain. Therefore, we restrict our discussion to pure scaling operations and use the terms *normalization* and *scaling* interchangeably.

Scaling coefficients are represented by an n-dimensional vector $\mathbf{k} = (k_1, \ldots, k_n)$. It is often convenient to express scaling as a diagonal matrix $\mathbf{K} = diag(\mathbf{k})$, where the scaling coefficients appear along the diagonal. The projection of the multidimensional dataset $\mathbf{D}$ using the linear operator $\mathbf{A}$ after applying scaling coefficients $\mathbf{k}$ is given by:

$$\mathbf{P} = \mathbf{A} \cdot \mathbf{K} \cdot \mathbf{D}. \tag{1}$$

SC represent the projection matrix $\mathbf{A}$ as a system of n vectors, each corresponding to a column of $\mathbf{A}$ and, thus, to a specific data attribute. SC allow users to interactively manipulate these vectors, steering elements of the projection matrix in real-time. The updated matrix is then used to re-project the dataset $\mathbf{D}$, enabling users to explore the space of linear projection operators in relation to the configuration space of projections $\mathbf{P}$.

3.2 Uniformity Measure

Visual analysis of projected data aids in understanding the characteristics and structure of multidimensional data. Identifying patterns in the projection domain, such as clusters and outliers, is a key aspect of data analysis. Some projections provide more insight than

others. To compare the quality of different projections in this regard, it is necessary to numerically evaluate how structured or structure-free a given projection layout is.

Ong et al. [OKO12] proposed a physics-based approach for characterizing the uniformity of 2D sample distributions. In this method, samples in the 2D projection domain are treated as particles in a repulsive force field. The potential energy of the system reaches its minimum when particles are as evenly distributed as possible, resulting in a uniform distribution. The potential energy between a pair of particles, separated by a distance r_{ij}, is modeled as $p_{ij} = \Theta/(\Theta - r_{ij}^2)$, where Θ is a constant. The global uniformity measure is then defined as

$$Q = \frac{1}{m} \cdot \frac{1}{q} \sum_{i=1}^{m} \sum_{j=1}^{q} \frac{r_{ij}^2}{\Theta + r_{ij}^2}. \tag{2}$$

Here, q nearest neighbors of each sample are considered, and Θ is selected based on the typical distance between uniformly distributed points. Ong et al. [OKO12] suggested $\Theta = 3m^{-1}$ for m points sampled in a unit square. However, our numerical tests confirm that Q is not invariant to the choice of q under this Θ. We, therefore, correct this by setting $\Theta = 3qm^{-2}$, ensuring consistent scaling of Q. This corrected Θ is used in our numerical experiments in Section 5.

Structures detectable in the projection domain can be imprints of intrinsic structures of the original multidimensional data or artifacts of improper scaling and projection. To enhance data analysis reliability, artificial patterns caused by improper scaling should be eliminated. Lehmann and Theisel [LT18] proposed that such artifacts vanish when adjusting vector $\mathbf{k}$. The most uniform projected distribution, achieved by optimizing scaling coefficients for a given projection operator $\mathbf{A}$, provides the most reliable structures for analysis. Thus, we seek $\mathbf{k}$ that maximizes the uniformity measure:

$$\mathbf{k}^* = \arg\max_{\mathbf{k}} Q(\mathbf{A} \cdot \mathbf{K} \cdot \mathbf{D}). \tag{3}$$

We denote diagonal matrix of optimal scaling coefficients $K^* = diag(\mathbf{k}^*)$ and the most uniform samples' distribution with respect to all possible scalings.

3.3 Regularization

While directly solving problem (3) numerically is possible, e.g., via iterative maximization, evaluating the measure Q requires computing nearest neighbors. The complexity of this approach is $\mathcal{O}(m \log m)$, making interactive rates unattainable for large m.

Lehmann and Theisel [LT18] avoided explicit evaluation of Q. Since Q is maximized for uniform distributions in the 2D projection domain, they proposed regularizing P using the Lloyd relaxation algorithm [Llo06] and determining the scaling $\mathbf{k}^*$ that makes P as close as possible to the regularized distribution. In other words, maximizing the quality measure is achieved by approximating a point distribution for which Q is maximal. Their iterative algorithm follows these steps: First, construct the Voronoi diagram [Aur91] of the 2D point set P using the Fortune algorithm [For86, For04]. Second, compute the centroids of the Voronoi cells, denoted as $\mathbf{C}$. Then, update the current scaling $\mathbf{k}$ such that P approximates $\mathbf{C}$. These steps are repeated until convergence.

The approach by Lehmann and Theisel has several limitations. First, its complexity is $\mathcal{O}(n \cdot m^2)$, which is higher than direct optimization of Q in Equation (2). This issue was mitigated by uniform data sampling [LT18]. Second, Lloyd relaxation is only guaranteed to converge in 1D or 2D domains [SG86], restricting its generalizability. Additionally, numerical instabilities may arise due to duplicate points, requiring special handling. Finally, the regularity of the resulting projected distribution was not evaluated, leaving it unclear how close Q can be brought to its optimal value, Q^*.

Recently, Rave et al. [RML25] proposed an efficient regularization technique using InIms. For a given 2D sample distribution, a rasterized density function is computed by summing isotropic kernel functions centered at sample positions. Then, eight InIms are computed by summing density values over rectangular areas, rotating around each pixel in 45° increments. This set of InIms provides integral density distribution information, which is then used to compute a deformation field that moves the sample distribution toward a more regular state. The mapping is applied iteratively, with density field and InIm computations repeated until a stable configuration is reached.

The algorithm by Rave et al. [RML25] enables efficient parallel implementation on the GPU with complexity $\mathcal{O}(m)$, as demonstrated in Section 5.3. In our work, we apply InIms-based regularization to the projected sample distribution, replacing the Voronoi centroids $\mathbf{C}$ with the resulting distribution. The optimal scaling $\mathbf{k}^*$ is then computed using the steepest descent method. Details are presented in Section 4.

4 EFFICIENT REGULARIZATION-BASED NORMALIZATION

Projections map multidimensional data into a low-dimensional visual domain. When used for data analysis, projections should be both informative and reliable. Informative projections provide insights into the multidimensional data structure. A comprehensive exploration of multidimensional data typically requires examining a large set of projections, such as those presented in data tours [Asi85, FT74, CBCH95].

Identifying a minimal subset of the most informative projections remains an active research area [LT16].

Projection reliability pertains to the interpretability of patterns in the projected data, ensuring they accurately reflect structures in the original multidimensional space. Misleading patterns can emerge due to improper data scaling, which can distort relationships and create artifacts. The goal of the proposed method is to eliminate such scaling artifacts, thereby enhancing the reliability of the resulting projection.

In the context of Equation (1), a projection $\mathbf{P}$ of data $\mathbf{D}$ is reliable when the projection operator $\mathbf{A}$ is chosen to maximize the preservation of meaningful structures, while the scaling matrix $\mathbf{K}$ (or equivalently, the scaling vector $\mathbf{k}$) is selected to minimize artificial patterns. Since choosing an appropriate projection operator $\mathbf{A}$ is beyond the scope of this paper, we assume it is fixed. The optimal scaling $\mathbf{K}^*$ is the one that suppresses artificial patterns, ensuring that $\mathbf{P}$ is as uniform as possible. Various measures can assess the uniformity of $\mathbf{P}$, including measure Q defined in Equation (2).

We optimize the scaling vector $\mathbf{k}$ using the steepest descent method, assuming a fixed projection operator $\mathbf{A}$. The optimization objective is to bring $\mathbf{P}$ as close as possible to a regularized counterpart $\mathbf{R}$. That is, we seek scaling coefficients $\mathbf{k}^*$ such that $\mathbf{P}$ closely approximates $\mathbf{R}$. Following the approach in [LT18], we use our regularization $\mathbf{R}$ instead of Voronoi centroids.

Construction of $\mathbf{R}$ follows the method developed by Rave et al. [RML25]. Several practical optimizations were introduced by integrating the regularization procedure with the adjustment of scaling coefficients using the steepest descent method. The full algorithm is described below:

1. Given a linear projection operator $\mathbf{A}$, initialize the scaling coefficients $\mathbf{k}_0$: If $\mathbf{A}$ has been modified interactively (e.g., using SC), the previously optimized scaling coefficients can be reused. Otherwise, standard normalization techniques (such as range scaling or whitening, see Section 1) provide the initial values.
2. Compute the projection $\mathbf{P}$ of the multidimensional dataset $\mathbf{D}$ according to Formula (1) using the current scaling coefficients.
3. Regularize $\mathbf{P}$ using the InIm-based approach by Rave et al. [RML25]. The steps for computing $\mathbf{R}$ are as follows:
   (a) Construct a rasterized density field by convolving a Gaussian kernel with the sample distribution $\mathbf{P}$ and store as a texture.
   (b) Compute eight InIms for each texel of the density texture.
   (c) Calculate the deformation field by evaluating the eight InIms for each texel.
   (d) Map all samples to their new positions by moving them in the direction of the deformation vectors.
   (e) Repeat the process until a stable configuration $\mathbf{R}$ is achieved.

   Our experiments indicate that recomputing $\mathbf{R}$ every s steps suffices, which further reduces computation time (see Section 5.1). Thus, the computed regularization from one iteration can be reused s times. Moreover, our numerical tests also show that using a relatively low-resolution texture has a negligible effect on the final result while significantly reducing computation time (see Section 5.3).
4. Optimize the scaling coefficients $\mathbf{k}$ using the steepest descent method. The updated coefficients can be immediately applied to recompute the projection layout $\mathbf{P}$, effectively restarting the optimization process from step 2. The iterations of the steepest descent method terminate when changes in $\mathbf{k}$ become negligible or when the maximum number of iterations is reached.

5 RESULTS

We conducted numerical experiments using the *Wine* and *Communities* datasets from the UCI Machine Learning Repository [DG19], as well as the *Swiss Roll* dataset [vdMPvdH09]. The numerical experiments presented in this section were performed on a PC equipped with an Intel Core i7-13620H CPU and an NVIDIA GeForce RTX 4070 Laptop GPU. To construct the discretized density field, we used textures of varying sizes with a truncated Gaussian kernel function of size 13×13 pixels. Effects of varying kernel size were presented by Rave et al. [RML25]. We implemented the Lloyd algorithm as described in [RLW+11] and used the Eigen library [GJ+10] for linear algebra operations. Parallel computations were performed using OpenGL and CUDA [NVF20].

5.1 Computation Times and Uniformity

We compared the proposed approach with the state-of-the-art method by Lehmann and Theisel [LT18] by conducting a series of tests with varying parameters. The texture size was set to values from $\{128, 256, 512, 1024\}$ for the proposed method and from $\{128, 256, 512, 1024, 2048, 4096\}$ for Lloyd-Relaxer. The number of steepest descent iterations was chosen from the set $\{20, 40, 60, 80, 100\}$, and the step size was set to multiples of 0.1 within the interval $[0.1, 1.0]$. We recomputed $\mathbf{R}$ every $s = 1, \ldots, 5$ iterations. For each combination of parameters, we recorded the execution time of each algorithm as well as the uniformity measure Q.

Results for the *Communities* and *Wine* datasets are presented in Figure 1. For the former, the proposed method shows a noticeable improvement in achieving higher

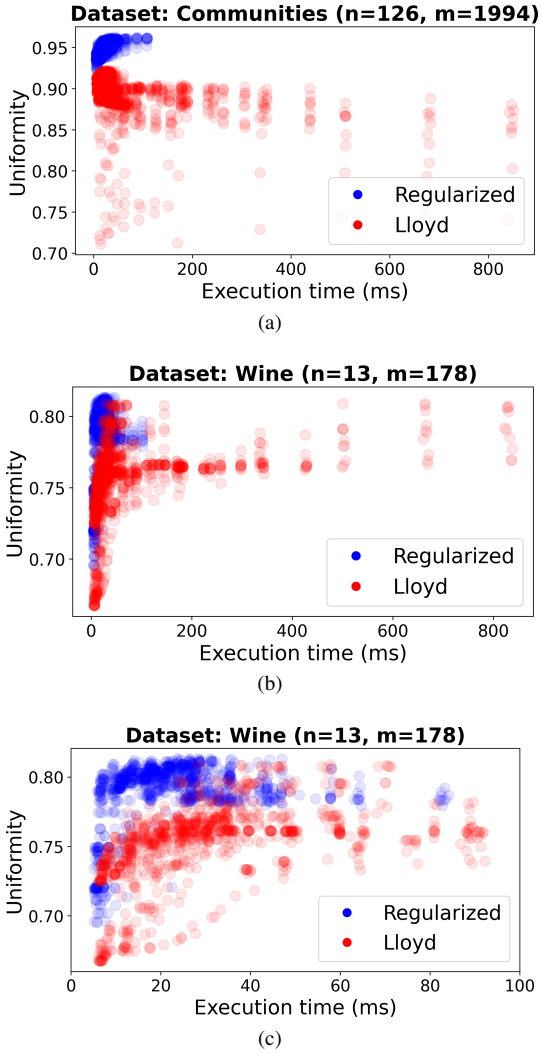


Figure 1: Computation times and uniformity Q of our approach for a set of parameter configurations (blue) in comparison to the state-of-the-art (red): (a) *Communities* dataset, (b) *Wine* dataset, (c) zoomed-in version of (b).

uniformity values Q for most parameter combinations while maintaining comparable computational timings (see Figure 1a). For the latter, the choice of parameters has a greater impact on the measured characteristics. However, the proposed method achieves interactive rates for any parameter combination (Figure 1b) and significantly higher Q values in most cases (Figure 1c). When computing the uniformity measure Q, we considered all pairs of samples, i.e., $q = m$.

5.2 Stability

The primary application scenario for both LloydRelaxer and the proposed method involves linear projections, potentially encoded in SC, for interactive exploration of multidimensional data. When the user interacts with the SC axes, i.e., modifies the linear projection operator, scaling coefficients are recalculated to eliminate artificial patterns in the projection domain. However, if the scaling coefficients change too dramatically, the projection layout undergoes significant distortions, hindering consistent data exploration.

In our next set of experiments, we compare the stability of the evolution of scaling coefficients computed by the two methods. Starting from a radial SC layout for the *Wine* dataset, we moved one axis along a circular trajectory (without changing its length) or, alternatively, along a lemniscate trajectory (where both length and orientation change):

$$x(t) = \cos t/(1+\sin^2 t),$$
$$y(t) = 0.2+\sin t\cdot\cos t/(1+\sin^2 t),$$

where $t \in [0, 2\pi]$. The positions of the other SC axes remained unchanged. Along the moving axis path, we recorded changes in the projected sample distances. Figure 2 presents detailed results for the movement of the SC axis corresponding to attribute 13, as well as aggregated data for all moving axes. In all tests, the changes caused by the proposed optimization procedure were (often substantially) smaller than or equal to those produced by LloydRelaxer.

Figure 3 summarizes the change of the scaling coefficients and the uniformity measure in the same tests. The uniformity measure Q reached higher values with the proposed algorithm while requiring smaller adjustments to the scaling coefficients. Thus, our approach ensures smoother behavior of the projected samples and higher uniformity when the user interactively explores the space of linear projection operators.

5.3 Scalability

The key difference between the proposed method and the LloydRelaxer approach lies in the regularization algorithm. Experiments presented in Figure 1 show that the proposed approach results in a higher uniformity measure while maintaining comparable execution times for small datasets. Next, we examine the scalability of both approaches with respect to data size.

Lehmann and Theisel stated that the complexity of the regularization algorithm used in [LT18] was $\mathcal{O}(m \log m)$. However, the implementation by Rong et al. [RLW+11] allows for $\mathcal{O}(m)$ complexity, which is numerically confirmed by our experiment presented in Figure 4a. Results for the InIm-based regularization are also shown in Figure 4b. Although both methods exhibit linear complexity in m, the latter algorithm achieves lower execution times.

Reducing the texture size for the InIm-based approach to 256×256 significantly improves computation speed. Further reductions in texture size have little effect on execution time but lead to a noticeable drop in accuracy. Additionally, texture sizes below 256×256 are

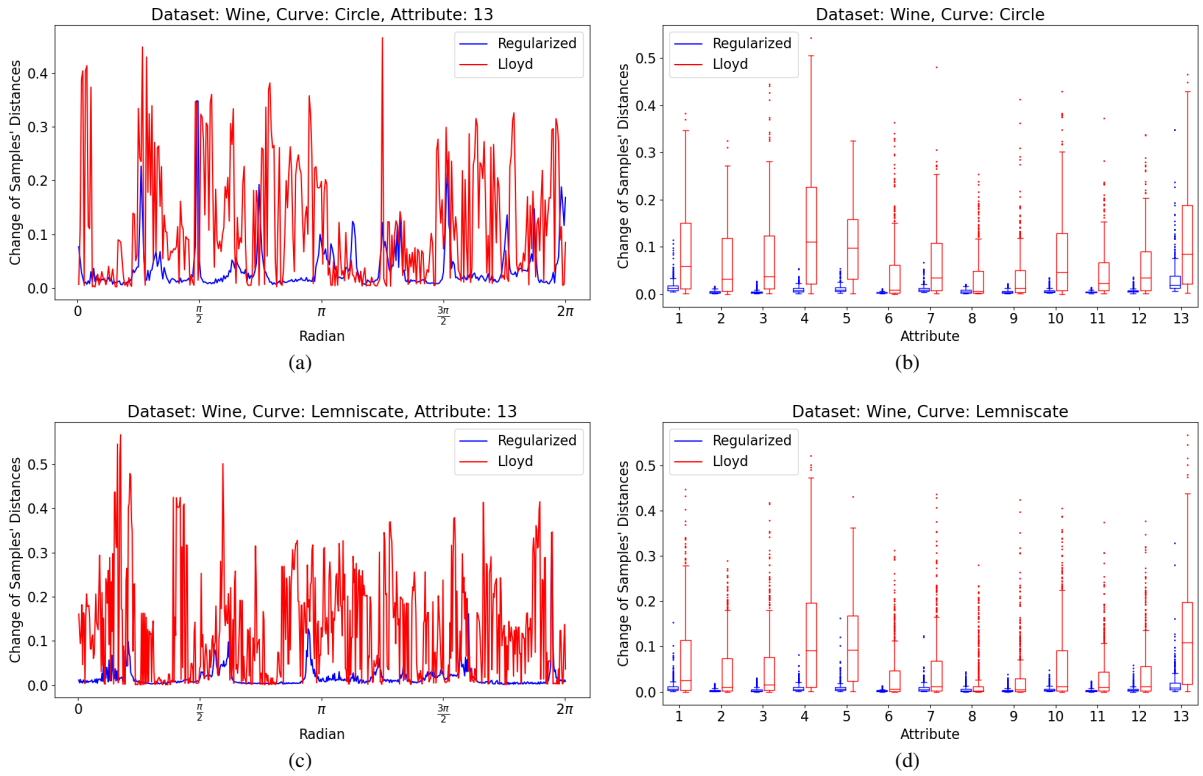


Figure 2: Stability (average sample distances) of our approach (blue) compared to the state-of-the-art (red). The *Wine* dataset is projected with a linear operator corresponding to a standard radial layout of SC. When the SC axis corresponding to attribute 13 is moved along a circular path (Figures 2a) or lemniscate (Figures 2c), changes in the average samples' distances are recorded. Summary of similar tests when moving other SC axes are presented in Figures 2b and 2d.

not supported by the implementation of the algorithm in Rong et al. [RLW+11].

5.4 Case Study

We applied the proposed normalization of data attributes to the *Swiss Roll* dataset [vdMPvdH09]. Figure 5 compares the classical linear projections with the results obtained using LloydRelaxer and the proposed approach for three different SC configurations.

The projection layout shown in Figure 5i may seem counterintuitive. Our method searches for a scaling that produces the most uniform sample distribution in the projection domain. However, the distribution in Figure 5i exhibits a highly structured pattern. Despite this, the uniformity measure is very high, with $Q = 0.95$, differing only slightly from $Q = 0.949$ computed for the projection in Figure 5h. Although the local sample density along the spiral curve is high, the samples remain well spread on the global scale. Moreover, this structured pattern is intrinsic to the artificial dataset, making its presence in the projection not only expected, but also desirable.

6 CONCLUSION

We presented a method for computing scalings of multidimensional data where scaling artifacts are minimized. The proposed technique enhances the state-of-the-art approach by Lehmann and Theisel [LT18] in multiple aspects. First, interactive rates are achieved for significantly larger datasets. No down-sampling is required, and the scalability of the proposed algorithm is superior to that of the existing method. Second, the novel approach achieves higher values of the uniformity measure Q. In this work, we corrected the uniformity measure proposed by Ong et al. [OKO12], making it invariant to the number of nearest neighbors considered. Third, numerical instabilities related to duplicate points in the dataset are avoided. As a result, scaling coefficients – and consequently, the projection layout – evolve more smoothly during user interactions with SC. Several application cases are shown in the accompanying video. The proposed approach is currently applied to linear projection operators. In future work, we will generalize the algorithm to non-linear dimensionality reduction methods.

7 ACKNOWLEDGMENTS

This work was funded by the Deutsche Forschungsgemeinschaft (DFG) grants MO 3050/2-3 – 360330772 and CRC 1450 – 431460824.

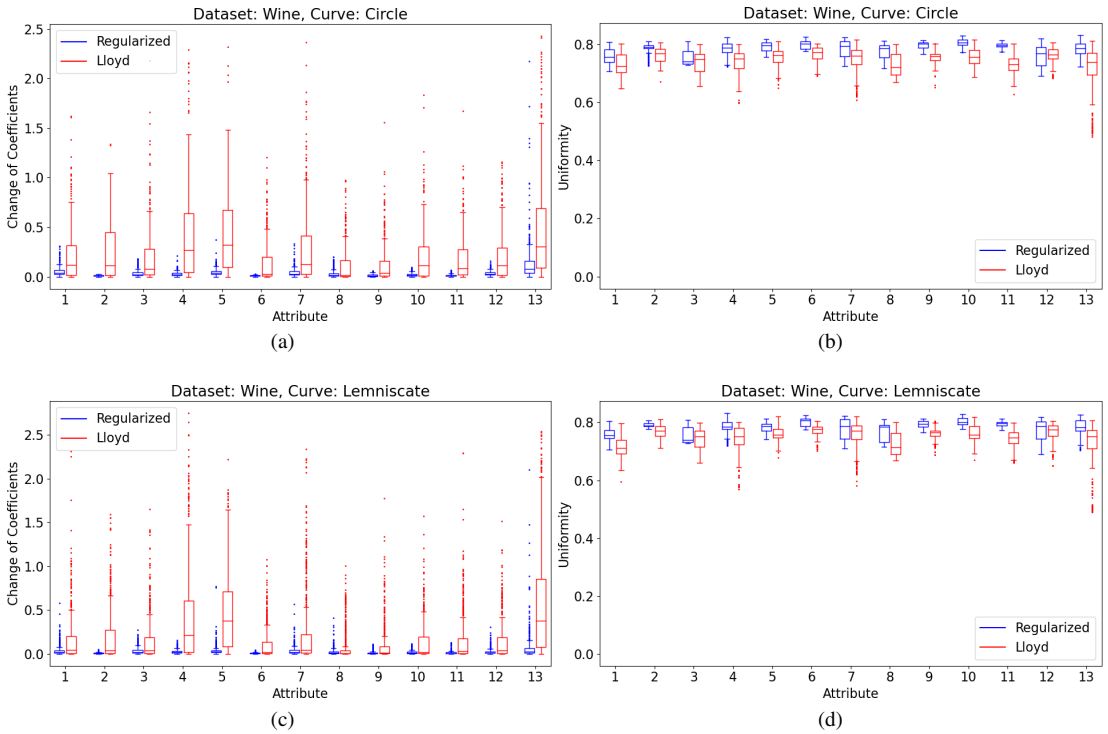
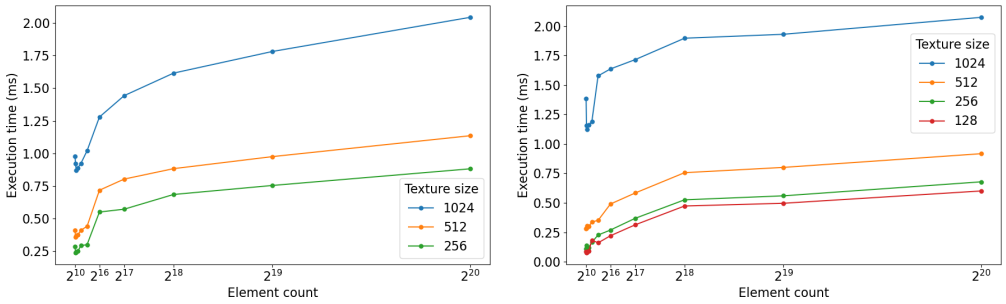


Figure 3: Stability (uniformity and change in scaling coefficients) of our approach (blue) compared to the state-of-the-art (red). *Wine* dataset is projected with a linear operator corresponding to a standard radial layout of SC. When one axis of SC is moved along a circular path (Figures 2a and 2b) or lemniscate (Figures 2c and 2d), changes in the uniformity measure and scaling coefficients are recorded.



(a) Algorithm from Rong et al. [RLW+11] (b) InIm-based approach

Figure 4: Scalability of the regularization methods with respect to data and texture size for artificial datasets with normally distributed samples. Comparison of our approach (b) to the state-of-the-art (a).

REFERENCES

[Asi85] Daniel Asimov. The grand tour: A tool for viewing multidimensional data. *SIAM Journal on Scientific and Statistical Computing*, 6(1):128–143, January 1985.

[Aur91] Franz Aurenhammer. Voronoi diagrams – a survey of a fundamental geometric data structure. *ACM Computing Surveys*, 23(3):345–405, September 1991.

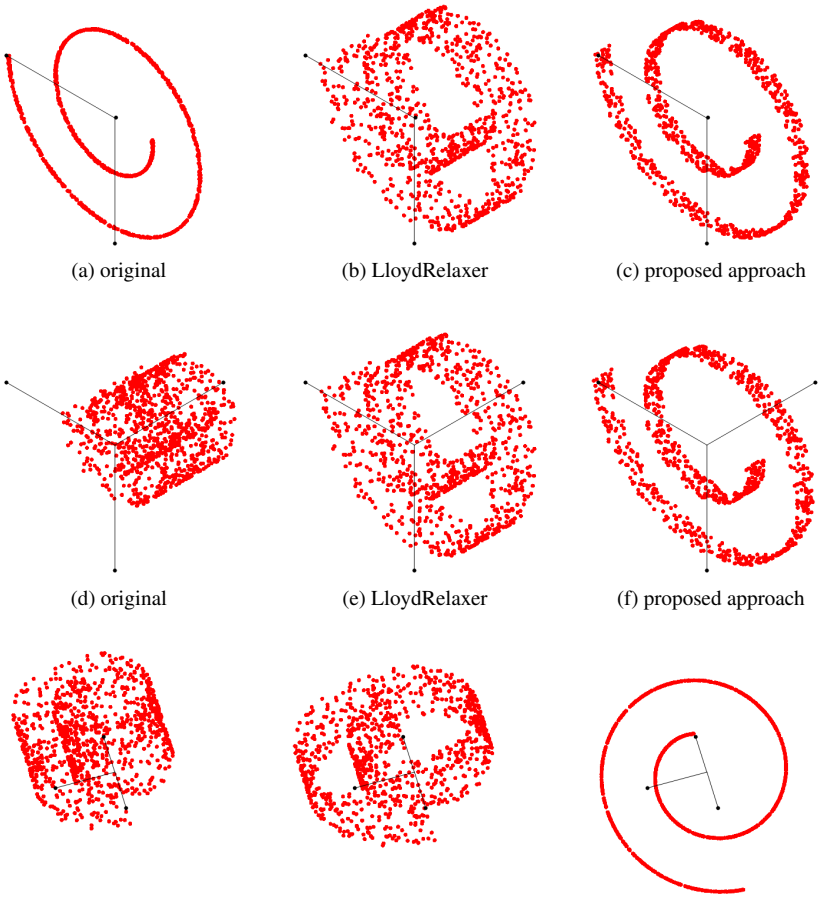
[BG10] Ingwer Borg and Patrick J. F. Groenen. *Modern Multidimensional Scaling Theory and Applications*. Springer Series in Statistics. Springer, 2nd edition edition, 2010.

[CBCH95] Dianne Cook, Andreas Buja, Javier Cabrera, and Catherine Hurley. Grand tour and projection pursuit. *Journal of Computational and Graphical Statistics*, 4:155–172, 1995.

[Cro84] Franklin C. Crow. Summed-area tables for texture mapping. In *Proceedings of the 11th Annual Conference on Computer Graphics and Interactive Techniques*, pages 207–212, New York, NY, USA, 1984. Association for Computing Machinery.

[DG19] Dheeru Dua and Casey Graff. UCI machine learning repository, 2019.

[ED07] Geoffrey Ellis and Alan Dix. A taxonomy of clutter reduction for information visualisation. *IEEE*



(a) original (b) LloydRelaxer (c) proposed approach

(d) original (e) LloydRelaxer (f) proposed approach

(g) original (h) LloydRelaxer (i) proposed approach

Figure 5: Optimal scaling of dimensions for *Swiss Roll* dataset [vdMPvdH09] for three SC configurations (rows): (a) without scaling, (b) state-of-the-art, and (c) our approach.

Transactions on Visualization and Computer Graphics, 13(6):1216–1223, November 2007.

[FHSW13] Martin Fink, Jan-Henrik Haunert, Joachim Spoerhase, and Alexander Wolff. Selecting the aspect ratio of a scatter plot based on its delaunay triangulation. *IEEE Transactions on Visualization and Computer Graphics*, 19(12):2326–2335, December 2013.

[Fis36] R. A. Fisher. The use of multiple measurements in taxonomic problems. *Annals of Eugenics*, 7(7):179–188, 1936.

[For86] Steven Fortune. A sweepline algorithm for Voronoi diagrams. In *SCG '86: Proceedings of the second annual symposium on Computational geometry*, pages 313–322, New York, NY, USA, 1986. ACM Press.

[For04] Steven Fortune. Voronoi diagrams and delaunay triangulations. In Jacob E. Goodman and Joseph O'Rourke, editors, *Handbook of Discrete and Computational Geometry, Second Edition*, pages 513–528. Chapman and Hall/CRC, 2004.

[FT74] Jerome H. Friedman and John W. Tukey. A projection pursuit algorithm for exploratory data analysis. *IEEE Transactions on Computers*, C-23(9):881–890, Sept 1974.

[GJ+10] Gaël Guennebaud, Benoît Jacob, et al. Eigen v3. http://eigen.tuxfamily.org, 2010.

[Jol86] I. T. Jolliffe. *Pincipal Component Analysis*. Springer-Verlag, 1986.

[Kan00] Eser Kandogan. Star coordinates: A multi-dimensional visualization technique with uniform treatment of dimensions. In *Proceedings of IEEE Information Visualization Symposium*, pages 4–8, 2000.

[Kan01] Eser Kandogan. Visualizing multi-dimensional clusters, trends, and outliers using star coordinates. In *Proceedings of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '01, pages 107–116, New York, NY, USA, 2001. ACM.

[Llo06] Stuart Lloyd. Least squares quantization in PCM. *IEEE Transactions on Information Theory*, 28(2):129–137, September 2006.

[LT13] Dirk J. Lehmann and Holger Theisel. Orthographic star coordinates. *IEEE Transactions on Visualization and Computer Graphics*, 19(12):2615–2624, 2013.

[LT16] Dirk J. Lehmann and Holger Theisel. Optimal sets of projections of high-dimensional data. *IEEE Transactions on Visualization and Computer Graphics*, 22(1):609–618, Jan 2016.

[LT18] Dirk J. Lehmann and Holger Theisel. The LloydRelaxer: An approach to minimize scaling effects for multivariate projections. *IEEE Transactions on Visualization and Computer Graphics*, 24(8):2424–2439, Aug 2018.

[McL04] Geoffrey J. McLachlan. *Discriminant analysis and statistical pattern recognition*. Wiley-Interscience Hoboken, N.J., 2004.

[MG13] Adrian Mayorga and Michael Gleicher. Splatterplots: Overcoming overdraw in scatter plots. *IEEE Transactions on Visualization and Computer Graphics*, 19(9):1526–1538, sep 2013.

[MHL20] Vladimir Molchanov, Sagad Hamid, and Lars Linsen. Efficient morphing of shape-preserving star coordinates. In *2020 IEEE Pacific Visualization Symposium (PacificVis)*, pages 136–145. IEEE, June 2020.

[MHSG18] Leland McInnes, John Healy, Nathaniel Saul, and Lukas Großberger. UMAP: Uniform Manifold Approximation and Projection. *Journal of Open Source Software*, 3(29):861, 2018.

[ML19] Vladimir Molchanov and Lars Linsen. Shape-preserving star coordinates. *IEEE Transactions on Visualization and Computer Graphics*, 25(1):449–458, 2019.

[MNM+21] Jeaneth Machicao, Quynh Quang Ngo, Vladimir Molchanov, Lars Linsen, and Odemir Bruno. A visual analysis method of randomness for classifying and ranking pseudo-random number generators. *Information Sciences*, 558:1–20, 2021.

[NVF20] NVIDIA, Peter Vingelmann, and Frank H.P. Fitzek. CUDA, release: 10.2.89, 2020.

[OKO12] Meng Sang Ong, Ye Chow Kuang, and Melanie Po-Leen Ooi. Statistical measures of two dimensional point set uniformity. *Computational Statistics & Data Analysis*, 56(6):2159–2181, 2012.

[PPM+15] Paulo A. Pagliosa, Fernando Vieira Paulovich, Rosane Minghim, Haim Levkowitz, and Luis Gustavo Nonato. Projection inspector: Assessment and synthesis of multidimensional projections. *Neurocomputing*, 150:599–610, 2015.

[RGE19] Renata G. Raidou, M. Eduard Gröller, and Martin Eisemann. Relaxing dense scatter plots with pixel-based mappings. *IEEE Transactions on Visualization and Computer Graphics*, 25(6):2205–2216, June 2019.

[RLW+11] Guodong Rong, Yang Liu, Wenping Wang, Xiaotian Yin, Xianfeng David Gu, and Xiaohu Guo. GPU-assisted computation of centroidal Voronoi tessellation. *IEEE Transactions on Visualization & Computer Graphics*, 17(3):345–356, March 2011.

[RML25] Hennes Rave, Vladimir Molchanov, and Lars Linsen. De-cluttering scatterplots with integral images. *IEEE Transactions on Visualization and Computer Graphics*, 31(4):2114–2126, 2025.

[SG86] Michael J. Sabin and Robert M. Gray. Global convergence and empirical consistency of the generalized Lloyd algorithm. *IEEE Transactions on Information Theory*, 32(2):148–155, 1986.

[SZS+17] Dominik Sacha, Leishi Zhang, Michael Sedlmair, John A. Lee, Jaakko Peltonen, Daniel Weiskopf, Stephen C. North, and Daniel A. Keim. Visual interaction with dimensionality reduction: A structured literature analysis. *IEEE Transactions on Visualization and Computer Graphics*, 23(1):241–250, Jan 2017.

[TSL00] Joshua B. Tenenbaum, Vin de Silva, and John C. Langford. A global geometric framework for nonlinear dimensionality reduction. *Science*, 290(5500):2319–2323, 2000.

[vdMH08] Laurens van der Maaten and Geoffrey E. Hinton. Visualizing high-dimensional data using t-SNE. *Journal of Machine Learning Research*, 9:2579–2605, 2008.

[vdMPvdH09] Laurens van der Maaten, Eric O. Postma, and Jaap van den Herik. Dimensionality reduction: A comparative review. *Journal of Machine Learning Research*, 10:66–71, 2009.

[VJ02] Paul Viola and Michael Jones. Robust real-time object detection. *International Journal of Computer Vision*, 57(2):137–154, 2002.

[WWF+19] Yunhai Wang, Zeyu Wang, Chi-Wing Fu, Hansjörq Schmauder, Oliver Deussen, and Daniel Weiskopf. Image-based aspect ratio selection. *IEEE Transactions on Visualization and Computer Graphics*, 25(1):840–849, January 2019.