Structure-aware Data Consolidation

Shihao Wu, Peter Bertholet, Hui Huang, *Member, IEEE,*
Daniel Cohen-Or, Minglun Gong, *Member, IEEE,* Matthias Zwicker, *Member, IEEE*

Abstract—We present a structure-aware technique to consolidate noisy data, which we use as a pre-process for standard clustering and dimensionality reduction. Our technique is related to mean shift, but instead of seeking density modes, it reveals and consolidates continuous high density structures such as curves and surface sheets in the underlying data while ignoring noise and outliers. We provide a theoretical analysis under a Gaussian noise model, and show that our approach significantly improves the performance of many non-linear dimensionality reduction and clustering algorithms in challenging scenarios.

Index Terms—Data consolidation, filtering, clustering, dimensionality reduction, manifold denoising.

1 Introduction

We present a structure-aware filtering (SAF) method that consolidates noisy data by projecting it onto underlying, lower dimensional structures. To reveal structures in noisy inputs, SAF concentrates sample points toward latent and lower dimensional data manifolds, while maintaining an even distribution of samples across these manifolds. We achieve this by adding a regularization to the weighted data averaging in conventional mean shift [1]. A theoretical analysis under a Gaussian noise model is provided, which reveals the parameter settings needed to balance between data concentration on the manifolds and even distribution across them. Empirical experiments show that SAF can significantly boost the performance of state-of-the-art clustering and dimensionality reduction approaches.

In clustering applications, data may form arbitrary, lower dimensional structures embedded in a feature space. A general strategy to address this problem is to project the data into lower dimensional subspaces where the clusters are more apparent. Often numerous subspaces are required, for example if each cluster manifests itself in a different subspace. The problem is more challenging, however, when the clusters form non-linear structures as demonstrated in Figure 1. Here each cluster has a curvy non-convex structure and the two clusters are intertwined such that they are not separated in any linear subspace. It becomes even more difficult in the presence of irrelevant features or data measurement uncertainties, which appear as noise. As shown in Figure 1, standard techniques such as spectral clustering or DBSCAN may fail to cluster such data.

Our structure-aware filtering technique (SAF) excels when the data forms low-dimensional structures that are contaminated by higher-dimensional noise. It is most effective when the low-dimensional manifolds are highly non-linear, like the curvy clusters in Figure 1, which SAF recovers succinctly (red points). After processing the data with SAF, standard clustering techniques are successful as demonstrated in Figure 1. A key advantage is that SAF does not require any local parametric representations of the underlying manifolds. Testing and evaluating our method on various benchmarks shows that SAF improves performance of many standard clustering techniques.

Clustering is also related to (nonlinear) dimensionality reduction, as many clustering techniques strive to construct a lower dimensional embedding of the data before clustering. In comparison, our approach does not project the data to lower dimensional spaces directly. Instead, it projects noisy data onto lower dimensional manifolds, but each data point still maintains its high dimensional features. It can therefore be viewed as a "dimension consolidation" technique. We demonstrate that our strategy can be used as a pre-process to standard dimensionality reduction, and our consolidation indeed leads to more robust results of a number of standard clustering approaches as well.

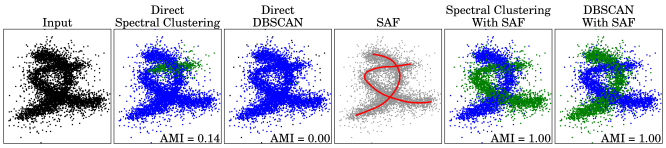


Fig. 1. A challenging example with two intertwined clusters, corrupted with noise. After projecting the input data (left) onto the underlying structure using our structure-aware filtering (SAF) approach (red points in the middle), spectral clustering or DBSCAN (right) detect the proper clusters.

In summary, the main contribution of this paper is to demonstrate how to boost the performance of common clustering and dimensionality reduction techniques by applying a novel structure-aware data consolidation approach as a pre-process. Our theoretical analysis proofs that, under simplifying assumptions (isotropic Gaussian noise with known variance, planar manifolds in arbitrary dimensions), the proposed structure-aware filtering converges to the underlying data manifolds. Empirical evidence shows that our analysis provides valid guidance to the selection of algorithmic parameters in practical applications.

2 Related Work

For an overview of research on clustering we refer the reader to recent surveys [2], [3] and discuss only selected works here. Many clustering techniques are derived from k-means clustering, including for example k-medians [4] and many others. Gaussian mixture models (GMMs) [5] and the expectation-maximization (EM) algorithm are also closely

related to k-means clustering. One of the main limitations of most k-means related algorithms is their assumption that clusters exhibit simple shapes, such as isotropic (k-means clustering) or ellipsoid (GMMs) distributions. In contrast, hierarchical clustering does not require the user to specify the number of clusters and can naturally produce arbitrary cluster shapes. We refer readers to standard textbooks for an overview [6]. These techniques either proceed top-down (divisive) or bottom-up (agglomerative). DBSCAN [7] is one of the most popular agglomerative algorithms. It greedily aggregates points in high density neighborhoods to clusters, which may form arbitrary shapes.

Clustering by seeking modes of an underlying density distribution is another popular approach. The best known example is the mean shift algorithm [1] and its variations. These techniques do not require the specification of the number of clusters and can also find clusters with arbitrary shapes. The underlying kernel density estimation, however, also suffers from the curse of dimensionality, which restricts them to lower dimensions.

Subspace clustering is a general strategy to work around the curse of dimensionality, and we refer the reader to Kriegel et al.'s recent survey of these techniques [2]. Dimensionality reduction maps the data to a lower dimensional space, often using non-linear techniques [8], before further processing such as visualization or clustering. Spectral clustering [9] relies on an embedding given by the spectral analysis of the similarity matrix of the data. It is highly related to dimensionality reduction techniques using Laplacian eigenmaps [10] and diffusion maps [11]. Typically, spectral clustering techniques produce their final output by applying k-means clustering after dimensionality reduction. Kannan et al. [12] provide a thorough analysis of spectral clustering using a novel quality criterion. Dhillon et al. [13] made an interesting connection between kernel k-means [14] and multiclass spectral clustering [15].

The method proposed in this paper is not a clustering or dimensionality reduction technique on its own, but it can significantly improve the performance of many approaches mentioned above by consolidating the data before clustering or embedding. Technically, our consolidation algorithm falls into the locally optimal projection (LOP) framework [16], [17], [18] in its discretized form. Huang et al. further introduce a *L_1-medial skeleton* [19] as a curve skeleton representation for 3D point clouds using such locally optimal contraction. Wu et al. augment each surface point to a *deep point* [20] by associating it with an inner point that resides on a structural mixture of skeletal curves and sheets. The common objective of these works is to seek a proper interpretation of the noisy input using a data fitting term complemented with a repulsion term. In a similar spirit, we consider a large sample scenario and have made the first attempt on analyzing continuous densities, resulting in a convergence proof for a special case, and a structure-aware data consolidation method that greatly assists many data mining applications.

Our technique also shares similarities with manifold denoising algorithms [21], [22], [23], [24], [25], [26]. Manifold blurring mean shift [22] restricts mean shift directions to be parallel to manifold normals estimated using local PCA. Sparse subspace denoising [25] builds on sparse subspace clustering and includes a subspace reconstruction error by estimating locally linear subspaces using PCA to achieve denoising. Robust PCA [27] suppresses outliers by decomposing a highly corrupted measurement matrix into a low-rank and a sparse matrix. Hein and Maier [21] propose Manifold Denoising (MD) using a neighborhood graph Laplacian of the data. Laplacian smoothing, however, shrinks the manifold and ultimately collapses it to a single point. Hence, manual tuning of the desired amount of smoothing is in general required. Most recently, Deutsch et al. [26] propose a Manifold Frequency Denoising (MFD) algorithm by removing the high frequency bands in the spectral graph wavelet domain. It is a global method and can produce clean output when there exists only one underlying manifold. However, when the noise is severe and the underlying manifolds are nearby, the results of MFD degenerate.

3 METHOD

Our goal is to improve existing clustering and dimensionality reduction algorithms by developing a data consolidation technique as a pre-process, which we call structure-aware filtering (SAF). Here we first introduce the SAF approach (Section 3.1) by starting with an intuitive continuous formulation, where input and output data are modeled as continuous density functions. This facilitates a theoretical analysis that allows us to explicitly derive the behavior of SAF (Section 3.2). Finally, we discuss a discrete implementation (Section 3.3).

3.1 Structure-aware Filtering

SAF consolidates noisy data densities by contracting them locally to remove noise and reveal high density structures. We first formulate SAF by modeling the noisy input data densities as continuous functions, and expressing SAF as a continuous flow given by a time dependent velocity field $v(z,t) : (\mathbb{R}^n, R) \rightarrow \mathbb{R}^n$. Denoting the input data density $f_p(z)$, we initialize a time-dependent output density $f_x(z,t)$ as $f_x(z,0) = f_p(z)$. Then, the goal of the SAF flow is to advect $f_x(z,t)$ to gradually remove noise while revealing the underlying structures in the input density $f_p(z)$.

We model the noisy input data density $f_p(z)$ by adding noise to an underlying m-dimensional data manifold M. Let us assume the data is mapped to $\mathbb{R}^n$ via an embedding $i : M \rightarrow \mathbb{R}^n$, and we have a probability density p_M on M. Then we express the data-generating process in $\mathbb{R}^n$ as $X = i(\theta) + \epsilon$, where $\theta \sim p_M$ and we assume isotropic Gaussian noise $\epsilon \sim N(0, \sigma)$. Hence, the noisy input density is represented as

$$f_p(z) = (2\pi\sigma^2)^{-\frac{n}{2}} \int_M e^{-\frac{||z-i(\theta)||^2}{2\sigma^2}} p_M(\theta) d\theta. \qquad (1)$$

While noise distributions other than Gaussian could be used, we will focus on Gaussian noise in our theoretical analysis. Note that Equation (1) can be considered a generalization of the Gaussian latent variable model used in Probabilistic PCA [28], where θ is Gaussian and $i(\cdot)$ is linear.

The SAF velocity field consists of two components: the first one "pulls" along the gradients of the noisy input data density. This term tries to accumulate output density in local extrema of the noisy input density, and we call it the data term. The second term "pushes" output density along its negative gradients, hence we call it a repulsion

term. This term makes sure that the output density does not "clump" around weak density extrema in the noisy input data density. The repulsion term allows us to consolidate and enhance latent continuous structures in the input data, such as one-dimensional (curve) or higher-dimensional (surface) manifolds. More precisely, we define the SAF flow with the velocity field $v(z,t)$ as

$$v(z,t) = \nabla(f_p * K)(z) - \lambda(z,t)\nabla(f_x * L)(z,t).$$

Here the smoothing kernel K serves to remove noise from the input density, and L smooths the output density itself with a balancing weight function $\lambda(z,t)$. The output density $f_x(z,t)$ is time dependent, and related to the velocity field via the continuity equation

$$\frac{\partial f_x(z,t)}{\partial t} = -\nabla \cdot \left(f_x(z,t)v(z,t)\right).$$

Let us assume the smoothing kernels K and L are radially symmetric, so we can write them as $K(\xi) = k(\frac{1}{2}\|\xi\|^2)$ and $L(\xi) = l(\frac{1}{2}\|\xi\|^2)$, $\xi \in \mathbb{R}^n$. Further assuming k and l are differentiable, we have

$$\nabla K(\xi) = \xi k'(\frac{1}{2}\|\xi\|^2), \quad \text{and} \quad \nabla L(\xi) = \xi l'(\frac{1}{2}\|\xi\|^2),$$

where k' and l' are derivatives with respect to the argument $\frac{1}{2}\|\xi\|^2$. In addition, we define the weighting as

$$\lambda_x(z,t) = \mu \frac{\left(f_p(\xi) * k'(\frac{1}{2}\|\xi\|^2)\right)(z)}{\left(f_x(\xi) * l'(\frac{1}{2}\|\xi\|^2)\right)(z,t)},$$

with a global user parameter $\mu > 0$. Here ξ is the integration variable in the convolution, which we may omit in the following for clarity. After rearranging and scaling the velocity field we obtain the final SAF formulation

$$v(z,t) = \frac{f_p * \left(\xi k'(\frac{1}{2}\|\xi\|^2)\right)}{f_p * k'(\frac{1}{2}\|\xi\|^2)}(z) - \mu \frac{f_x * \left(\xi l'(\frac{1}{2}\|\xi\|^2)\right)}{f_x * l'(\frac{1}{2}\|\xi\|^2)}(z,t). \tag{2}$$

3.2 Theoretical Analysis

Given a noisy input density representing an underlying manifold, the data term of SAF attracts output density towards local maxima of the noisy input density, while the repulsion tries to maintain a smooth output density that is evenly distributed over the underlying manifold. These two terms need to be properly balanced: if the data term is too strong, data may be concentrated at isolated density modes; if the repulsion force is too strong, data may diffuse away. Here we provide a theoretical analysis of this process to understand under which circumstances SAF manages to attract density to the underlying manifold.

To make analysis tractable, we consider the special case where the smoothing kernels k and l are Gaussian, and the underlying data manifold is a hyperplane in $\mathbb{R}^n$, i.e., the noisy input data density f_p and the initial output density are degenerate Gaussians. We first analyze the one-dimensional case, and then generalize to arbitrary dimensions.

3.2.1 One-dimensional Case

Let $G_{0,\sigma^2}(\xi)$ denote a zero-mean, univariate Gaussian distribution with variance σ^2. We introduce the notation $g_{0,\sigma^2}(\frac{1}{2}\|\xi\|^2) = G_{0,\sigma^2}(\xi)$, and note that $g'_{0,\sigma^2} = -g_{0,\sigma^2}$. In the 1D case, the data manifold is represented by a Dirac impulse, the noisy input density is given by the 1D Gaussian $f_p = G_{0,\sigma^2}$, and the smoothing kernels are $k = l = g_{0,h^2}$. The following theorem describes how SAF converges to the noise free input (the Dirac impulse) in this scenario.

Theorem 1. *Let $f_p(z) = f_x(z,0) = G_{0,\sigma^2}(z)$, then for any $t \geq 0$ the output density is a Gaussian with some standard deviation $\omega(t)$, i.e., $f_x(z,t) = G_{0,\omega(t)}(z)$. In addition, if*

$$\sigma^2 < \frac{1-\mu}{\mu}h^2, \tag{3}$$

then $\omega(t) \to 0$ as $t \to \infty$. In other words, the output density converges to a Dirac impulse, i.e., the true 1D data manifold.

Proof. The initialization $f_p(z) = f_x(z,0) = G_{0,\sigma^2}(z)$ ensures that the first part of the theorem holds at $t = 0$. Assuming it is also true at time $t > 0$, we note that

$$\begin{aligned}
\frac{f_p * (\xi \cdot k')}{f_p * k'}(z) &= \frac{G_{0,\sigma^2} * (\xi \cdot g'_{0,h^2})}{G_{0,\sigma^2} * g'_{0,h^2}}(z) \\
&= -\frac{h^2}{\sigma^2 + h^2}z, \\
\frac{f_x * (\xi \cdot l')}{f_x * l'}(z,t) &= \frac{G_{0,\omega(t)^2} * (\xi \cdot g'_{0,h^2})}{G_{0,\omega(t)^2} * g'_{0,h^2}}(z,t) \\
&= -\frac{h^2}{\omega(t)^2 + h^2}z,
\end{aligned}$$

and hence the velocity from Equation (2) becomes

$$v(z,t) = \left(-\frac{h^2}{\sigma^2 + h^2} + \mu\frac{h^2}{\omega(t)^2 + h^2}\right)z. \tag{4}$$

This suggests that $v(z,t)$ corresponds to a uniform scaling of space, which means that the output density stays Gaussian at a time $> t$ and proofs the first part of the theorem by induction. To show the second part, let us denote the instantaneous scaling factor at time t as

$$\tau(t) = 1 - \frac{h^2}{\sigma^2 + h^2} + \mu\frac{h^2}{\omega(t)^2 + h^2}. \tag{5}$$

Space is monotonically contracted (scaled down) if $0 \leq \tau(t) < 1$ for all t, which is guaranteed by Equation 3. □

As a key result, the theorem above provides a simple equation that characterizes the user parameters μ (repulsion strength) and h^2 (size of smoothing kernel) that lead to guaranteed convergence.

3.2.2 Hyperplanes in Arbitrary Dimensions

We generalize this analysis to the n-dimensional case by considering axis-aligned hyperplanes with uniform density $p_M \equiv 1$ as the data manifolds[1]. Let us define an axis aligned hyperplane by the set H of coordinate axes that lie in the hyperplane. That is, H is a subset of the indices $\{1,\ldots,n\}$, and its cardinality $|H| = m$ corresponds to the

1. p_M cannot be considered a proper probability density in this case, but this is not an issue for our analysis.

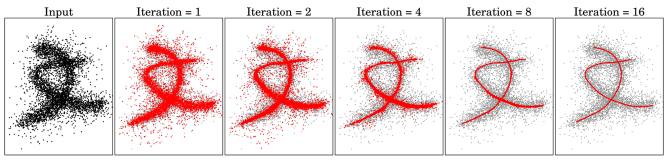


Fig. 2. Consolidation process using anisotropic SAF.

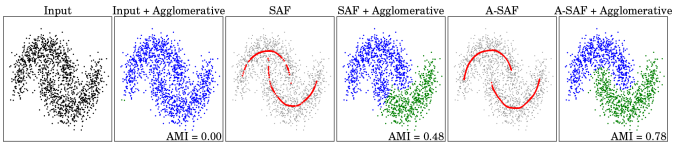


Fig. 3. Clustering without (SAF) and with anisotropic repulsion (A-SAF).

dimensionality of the hyperplane. Using Equation (1), this leads to noisy input densities represented by degenerate, axis aligned multivariate Gaussians with zero-means, which are equivalent to products of 1D Gaussians,

$$f_p(z) = \prod_{i \in \bar{H}} G_{0,\sigma^2}(z_i),$$

where $\bar{H} = \{\{1, \dots, n\} \setminus H\}$, z denotes an n-dimensional vector, and z_i is the i-th element in the vector.

Similar as in 1D, we assume the intermediate distribution is initialized as $f_x(z,0) = f_p(z)$, and the smoothing kernels are $k = l = g_{0,h^2}$. In this special setting, the n-dimensional case directly reduces to the 1D case as all involved functions are separable into products of 1D functions, which means that also the convolutions are separable. As a consequence, the velocity $v_i(z,t)$ in each dimension $i \in \bar{H}$ is analogous to the 1D case in Equation (4),

$$v_i(z,t) = \left(-\frac{h^2}{\sigma^2 + h^2} + \mu \frac{h^2}{\omega(t)^2 + h^2} \right) z_i, \quad \text{for } i \in \bar{H}, \quad (6)$$

and Theorem 1 applies to each dimension $i \in \bar{H}$ separately. On the other hand, the velocities parallel to the hyperplane are zero, $v_i(z,t) = 0$ for $i \in H$. Note that our analysis includes arbitrarily oriented hyperplanes, since we can simply rotate the coordinate system to align with the hyperplane, and then define the hyperplane as above.

3.2.3 *Curved Manifolds*

We can rewrite Equation (2) as

$$\begin{aligned} v(z,t) = & \underbrace{(1-\mu)(i(\theta_{\min}) - z)}_{\text{(I)}} \\ & - \underbrace{\left(i(\theta_{\min}) - \frac{f_p(\xi)\xi * k'(\frac{1}{2}\|\xi\|^2)}{f_p(\xi) * k'(\frac{1}{2}\|\xi\|^2)}(z) \right)}_{\text{(II)}} \\ & + \mu \underbrace{\left(i(\theta_{\min}) - \frac{f_x(\xi)\xi * l'(\frac{1}{2}\|\xi\|^2)}{f_x(\xi) * l'(\frac{1}{2}\|\xi\|^2)}(z,t) \right)}_{\text{(III)}}. \end{aligned} \quad (7)$$

where $i(\theta_{\min}) = \arg\min_{i(\theta)} ||z - i(\theta)||$ denotes the closest point to z on the data manifold M. The first term (I) represents motion towards the manifold M, as desired. Hein et al. [21] show that the second term (II) approximates $-mH - \frac{2}{p_M}\langle \nabla p_M, \nabla i \rangle$, where H is the mean curvature normal of M. The mean curvature term here smooths and shrinks the manifold, and the gradient of the data density ∇p_M leads to "clumping". Both these undesirable effects can be observed in practice.

In contrast, the repulsion in SAF is represented by a third, similar term (III), but with opposite sign compared to (II), and for an evolving manifold f_x. Hence repulsion in SAF counteracts the mean curvature smoothing and shrinkage, and it leads to more uniform densities on the manifold due to tangential diffusion. In practice (Section 3.3.2), our approach converges to stable structures without collapsing. In addition, if we choose the parameters according to Theorem 1, we obtain thin manifolds without noise. Nonetheless, the analysis provided here can only give an intuition; a thorough proof for non-linear cases is left for future work.

3.3 Discrete SAF with Anisotropic Repulsion

We implement a discretized version of SAF following a Lagrangian approach, i.e., we represent densities by sets of sample points. The input density f_p is given by points $\{p_j\}$, and f_x by points $\{x_i(t)\}$. Then the data term from Equation (2) (without normalization) is

$$f_p * \left(\xi k'(\frac{1}{2}\|\xi\|^2) \right)(z) = \sum_j (p_j - z) k'(\frac{1}{2}\|p_j - z\|^2).$$

In addition, let us generalize the continuous formulation from Equation (2) to anisotropic kernels for repulsion, which will allow for more effective repulsion in practice as discussed below. We implement an anisotropy by including a matrix A to linearly deform the repulsion kernel, that is $L(A\xi) = l(\frac{1}{2}\|A\xi\|^2)$. This leads to the generalized repulsion term and its discretized form

$$\begin{aligned} & f_x * \left(A^T A \xi k'(\frac{1}{2}\|A\xi\|^2) \right)(z) \\ = & \sum_j A^T A (p_j - z) k'(\frac{1}{2}\|A(p_j - z)\|^2). \end{aligned}$$

Now we evaluate the velocity field only at the sample points $\{x_i(t)\}$, and advect the samples with unit time steps. Their updated positions $\{x_i(t+1)\}$ in the next time step immediately define f_x for the next iteration. For simplicity of notation, denote $x_i = x_i(t)$ and $x_i' = x_i(t+1)$. Then the discretized version of SAF defined at each sample point is

$$\begin{aligned} x_i' = & x_i + \frac{\sum_j (p_j - x_i) k'(\frac{1}{2}\|p_j - x_i\|^2)}{\sum_j k'(\frac{1}{2}\|p_j - x_i\|^2)} \\ & - \mu \frac{\sum_{i'} A^T A (x_{i'} - x_i) l'(\frac{1}{2}\|A(x_{i'} - x_i)\|^2)}{\sum_{i'} l'(\frac{1}{2}\|A(x_{i'} - x_i)\|^2)} \\ = & \frac{\sum_j p_j k'(\frac{1}{2}\|p_j - x_i\|^2)}{\sum_j k'(\frac{1}{2}\|p_j - x_i\|^2)} \\ & - \mu \frac{\sum_{i'} A^T A (x_{i'} - x_i) l'(\frac{1}{2}\|A(x_{i'} - x_i)\|^2)}{\sum_{i'} l'(\frac{1}{2}\|A(x_{i'} - x_i)\|^2)}. \end{aligned}$$

The motivation behind the anisotropic repulsion is that, if the data forms a lower dimensional structure embedded in a higher dimensional space, we would like to direct repulsion

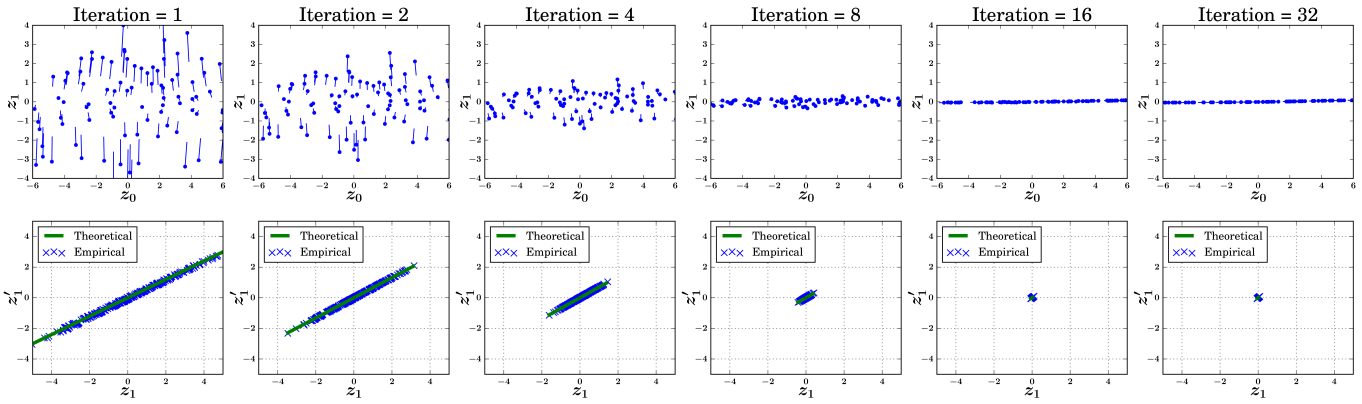


Fig. 4. We illustrate the iterative consolidation process in a 2D example with a data distribution $f_p = G_{0,(\infty,4)}$. The user parameters are $h = 4$ and $\mu = 0.5$, which leads to convergence according to Equation (3). The first row shows the actual point movements where the dots are current positions and the vectors are pointing to the next locations. The second row shows both empirical and theoretical (Equation (6)) update ratios of z_1'/z_1.

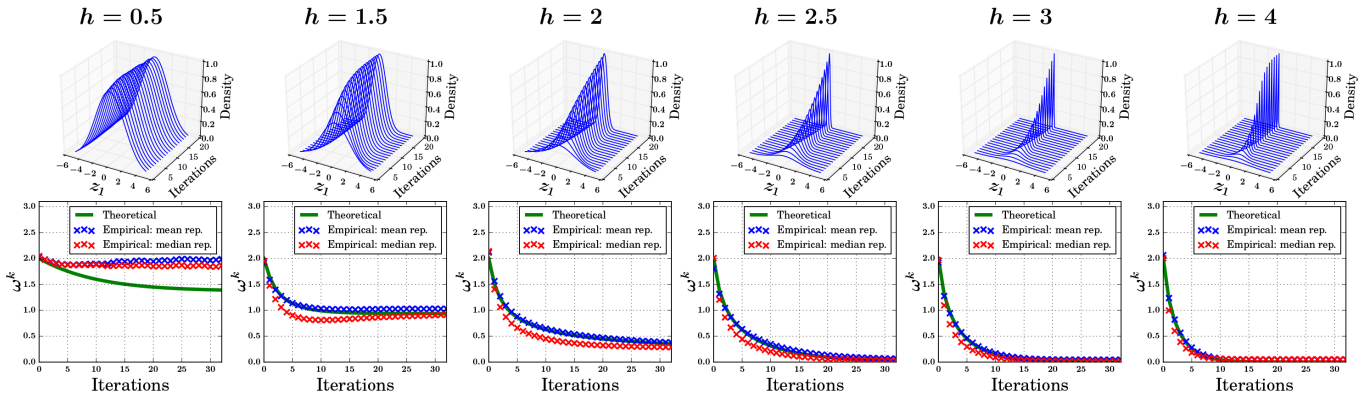


Fig. 5. Top row: we illustrate changes of point densities during the iterative consolidation process with different h values. Bottom row: empirical estimate and theoretical prediction of the variances ω of the intermediate point distributions. We show results using mean and median repulsion with blue and red crosses, respectively. While theoretical analysis with median repulsion is difficult, it empirically follows our prediction.

to move points around on the structure itself [19], [20]. In practice, for each output point x_i we perform a PCA analysis on its k nearest neighbors and get the corresponding eigenvectors $\{v_i^1, v_i^2, \cdots, v_i^n\}$ and eigenvalues $\{\lambda_i^1, \lambda_i^2, \cdots, \lambda_i^n\}$, where n is the dimension of the input data. We denote the $n \times n$ column matrix $A_i = [\lambda_i^1 v_i^1; \lambda_i^2 v_i^2; \cdots; \lambda_i^D v_i^D]$, and use it to adjust the shape of the repulsion kernel. Note that our analysis from Section 3.2 also applies to anisotropic repulsion. Anisotropic repulsion simply means that the variances h_i in the repulsion term in Equation (6) are scaled with the PCA eigenvalue along the corresponding coordinate axis.

Figure 2 illustrates the consolidation process using anisotropic SAF. The anisotropic repulsion force mainly pushes points along the local major PCA directions, which improves the regularity of data distribution and eventually can lead to better clustering, as demonstrated in Figure 3.

3.3.1 Kernel Selection

The kernel k for the data term should be smooth to eliminate noise in the input density, hence we use a (multidimensional) Gaussian $k(\frac{1}{2}\|\xi\|^2) = g_{0,h^2}(\frac{1}{2}\|\xi\|^2)$. The kernel l for the repulsion term should have large derivatives around the origin, such that close-by points are effectively pushed away from each other. Therefore, in practice we also use a modified repulsion kernel defined by its derivative

$$l'(\frac{1}{2}\|\xi\|^2) = \begin{cases} -g_{0,h^2}(\frac{1}{2}\|\xi\|^2)/\|\xi\| & \xi \neq 0, \\ 0 & \xi = 0. \end{cases} \quad (8)$$

In the discrete setting, using Gaussian repulsion kernels means that the repulsion term vanishes if each output point x_i minimizes a locally weighted sum of square distances, i.e., if each point is the mean of its locally weighted neighbors. We refer to it as "*mean repulsion*". With modified repulsion kernels defined in Equation (8), the repulsion term vanishes if each point minimizes a locally weighted sum of absolute distances, i.e., if each point is the median of its locally weighted neighbors. We refer to this as "*median repulsion*".

3.3.2 Empirical Validation

We empirically validate the theoretical results, that is, the velocity from Equation (6) and the convergence criterion from Equation (3), in a 2D setup with axes (z_0, z_1), following the notation in Equation (6). The input points are uniformly distributed along z_0, and normally distributed along z_1, that is, $\sigma_0 = \infty, \sigma_1 = 2$ (subscript indices are dimensions as in Equation (6)) and $f_p = G_{0,(\infty,4)}$. This is a degenerate 2D Gaussian modeling a 1D line in 2D. We set the repulsion strength to $\mu = 0.5$. According to Equation (3) we need $h^2 > 4$ (omitting the subscript index 1 for clarity) for convergence.

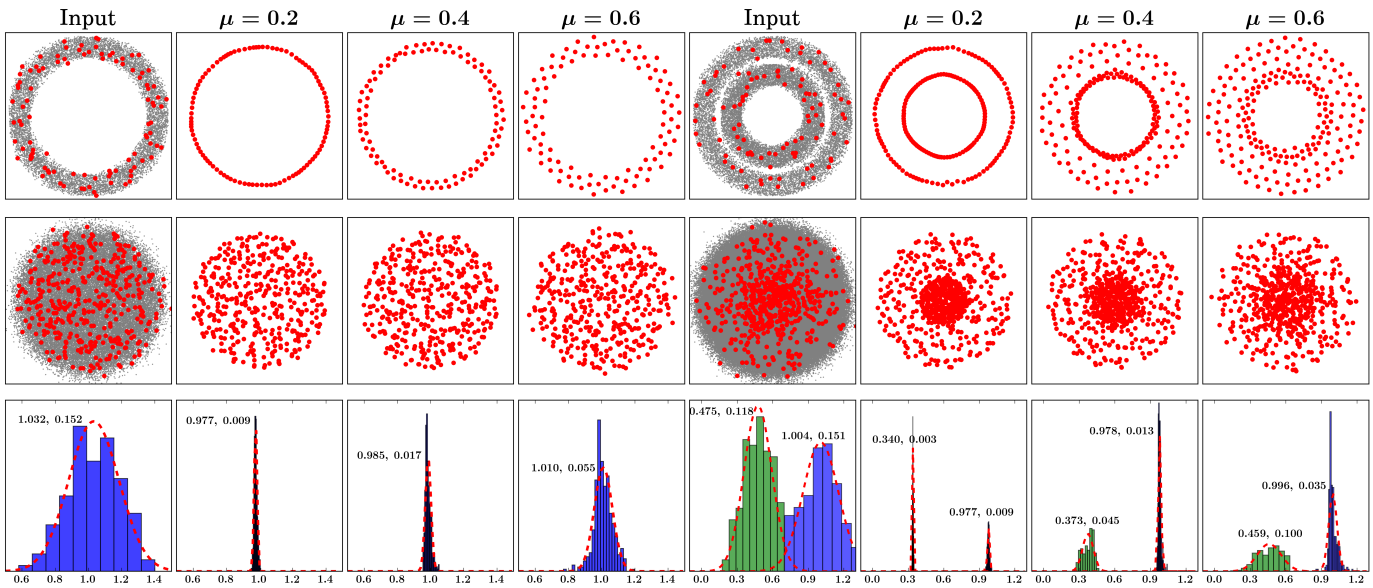


Fig. 6. Convergence in different dimensions and with median repulsion. We add Gaussian noise ($\sigma = 0.15$) to 2D circles (first row, single circle and two concentric ones) and 4D spheres (second row, single sphere and two concentric ones). We set the kernel size $h = 0.1$, and show results with different μ values. Equation (3) predicts convergence for $\mu < 0.31$. The third row shows histograms of distances to the center of the 4D spheres. The values next to the red bell curves are their empirical means and variances, demonstrating convergence to thin manifolds as predicted for $\mu < 0.31$.

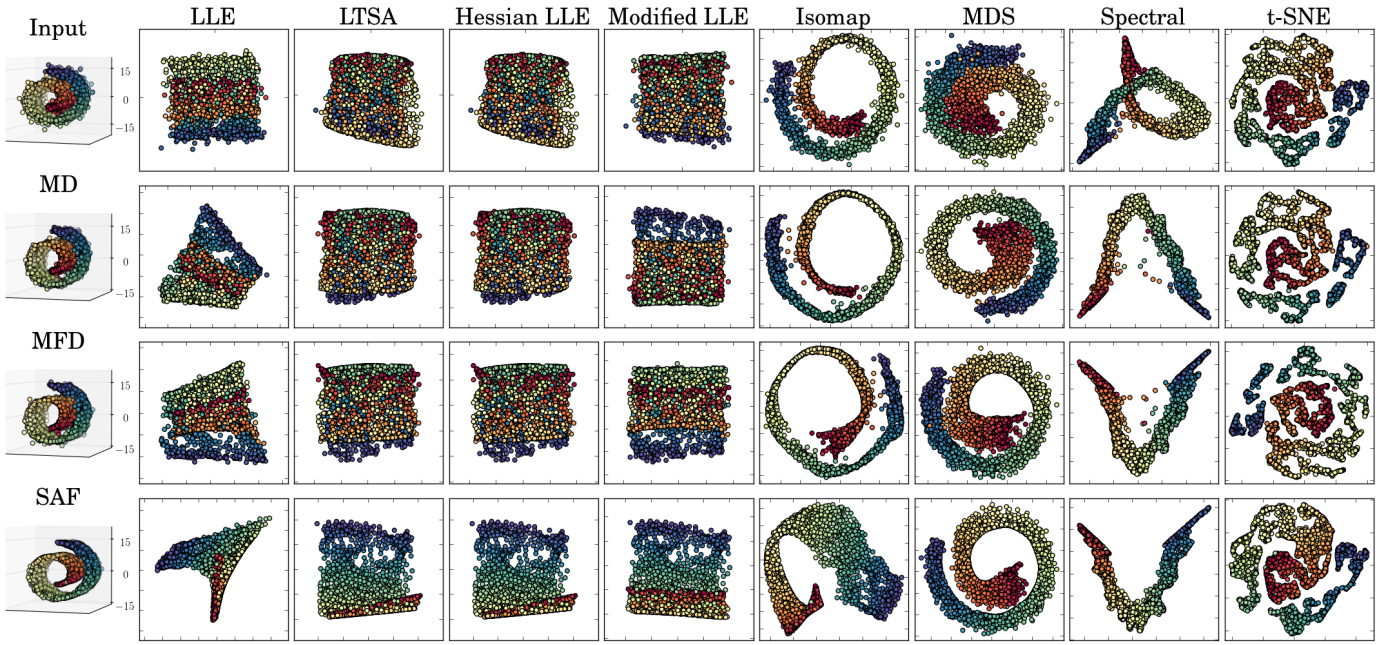


Fig. 7. Performance of dimensionality reduction without (top) and with MD [21] (second row), MFD [26] (third row), and SAF (bottom) consolidation.

We visualize the convergence process for $h = 4$ in Figure 4, where mean repulsion is used.

In Figure 5 we show evolution of the point density over the iterations for different values of h^2. This shows that for $h^2 < 4$ the distribution fails (or stops) to contract because of the repulsion term. For $h^2 > 4$ the distribution continuously sharpens at a rate that is well predicted by the theory. Deviations of empirical behavior from the theory can be explained by the fact that the discrete point sets do not exactly correspond to continuous Gaussian distributions.

Figure 6 illustrates that our theory well predicts the convergence behavior with median repulsion for data in different dimensions. While there is some shrinkage due to the curved manifolds, SAF converges to stable structures because of repulsion. In the supplementary we also compare empirical results for mean and median repulsion, which generates more uniform point distributions in practice.

3.3.3 *Comparison with Mean Shift and LOP*

The SAF data term is equivalent to mean shift, which accumulates output density at local extrema of the input. We are not interested in finding modes of the input, however. Instead, we want to produce an output density that removes noise

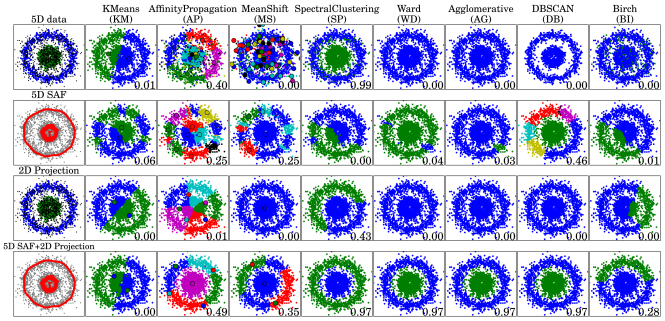


Fig. 8. Dimensionality reduction and clustering: the input consists of two concentric 2D circles corrupted with 5D noise (the black dots). We elevate 2D rings to 5D by appending zeros, and then add standard Gaussian noise. The leftmost column shows ground truth labeling (green and blue, outliers black), the consolidated points (red) and the input points (gray). We use PCA to project the 5D data to 2D. We compare four strategies (from the top row to the bottom), "direct clustering", "consolidation + clustering", "projection + clustering" and "consolidation + projection + clustering", with different clustering algorithms (from left to right). The second row reveals the sensitivity of most clustering techniques to higher dimensional data, which is not clustered well even though the consolidation exposes the structure of the data. The third row shows that reducing the dimensionality of the data does not solve the problem due to the noisy data. In general, the "consolidation + projection + clustering" strategy in the bottom row gives the best performance (AMI scores in bottom right of subfigures). Some techniques (k-means, affinity propagation, mean shift) are not suitable to cluster this type of data, and they do not benefit from consolidation.

from the input and consolidates and reveals its continuous structures. For this, the repulsion term is crucial.

The discrete formulation reveals that SAF is also a generalization of LOP operators [16], [17], [18], where the LOP weights correspond to the derivatives of certain radial kernels k and l. LOP with isotropic repulsion based on Euclidean distances [17] corresponds to median repulsion (Equation 8), while anisotropic SAF puts more effort on revealing and consolidating continuous high density structures in the underlying noisy data; see e.g., Figures 2 and 3.

4 Results

Here we discuss the application of our approach to dimensionality reduction and clustering. We provide more extensive experimental results in the supplemental material.

4.1 Dimensionality Reduction

Dimensionality reduction, or manifold learning, is an indispensable tool for data analysis, such as visualization or clustering. Existing techniques, however, often suffer from noise present in the high dimensional data. Our SAF can serve as a dimension consolidation tool that removes noise in high dimensional space, which greatly improves the performance of subsequent dimensionality reduction. In Figure 7 we test some of the most common dimensionality reduction methods with and without SAF consolidation.

We also compare SAF with two recent manifold denoising methods, Manifold Denoising (MD) [21], and Manifold Frequency Denoising (MFD) [26]. The resulting 2D embeddings show that the intrinsic shape of the data can be best preserved when the data is consolidated with SAF before dimensionality reduction.

4.2 Clustering

We evaluate our method by comparing the performance of selected clustering techniques with and without our data consolidation approach. As baseline techniques we selected clustering algorithms that are commonly used, widely available with source code, and representative for various clustering strategies: KMeans clustering (KM) [29], Affinity propagation (AP) [30], Mean Shift clustering (MS) [1], Spectral clustering (SP) [31], Ward clustering (WD) [32], Agglomerative clustering (AG) [33], DBSCAN clustering (DB) [7], and Birch clustering (BI) [34], all implemented in the scikit-learn library [35]. In each experiment, we tune the parameters for all the selected algorithms to achieve optimal consolidation results, and we compare results with and without our data consolidation approach. For each clustering algorithm, however, we use the same parameters regardless of using consolidation or not. SAF parameters for all the experiments can be found in supplemental material. When the ground truth labeling is given, we compute the Adjusted Mutual Information (AMI) to evaluate the clustering results. Note that we exclude the extremely noisy points (depicted as small black dots in the figures) from the calculation of AMI scores, because assigning ground truth labels for those points could be very ambiguous.

4.2.1 *Dimensionality Reduction and Clustering*

Many clustering algorithms cannot cope with high dimensional data well. In Figure 8, although our method can successfully clean up high dimensional noise and expose the low dimensional structure, the clustering algorithms do not benefit from the consolidation because the consolidated points remain in high dimension. Projecting the input data directly to a lower dimensional space often does not solve the problem if the data is noisy. Once we project the consolidated data into a lower dimensional space, however, the improvement of clustering is significant.

This suggests that our consolidation method neither stops working in high dimensional space, nor does it solve the high dimensionality problem on its own. Consolidation using SAF followed by dimensionality reduction and finally clustering, however, is an effective scheme.

4.2.2 *Different Dimensionalities*

In Figure 9 we investigate how our consolidation performs with increasing dimensionality. We test on data consisting of two concentric hyperspheres with different radii, corrupted by Gaussian noise. We keep the radii, the number of sample points, and the noise level constant independent of the data dimensionality. Note that we do not apply any dimensionality reduction in this experiment.

As shown in Figure 9, the clusterings degrade as the dimensionality increases up to 6D. One reason is that many clustering algorithms themselves cannot handle high dimensional data well. And, in higher dimensions, the data points become sparser, which inevitably affects the robustness of our consolidation. In other words, our method requires denser data points or a better neighborhood definition (distance metric) as the dimensionality increases. We also observe similar effects with MD [21] and MFD [26]. In general, however, SAF performs better in high dimensional cases.

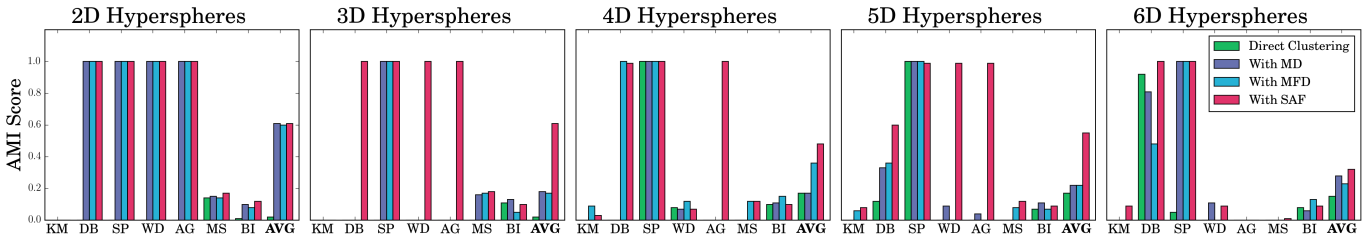


Fig. 9. Performance of SAF with increasing dimensionality, compared with MD [21] and MFD [26]. The data consists of two concentric hyperspheres with different radii, corrupted with Gaussian noise. The rightmost bars "AVG" show the average over all clustering methods.

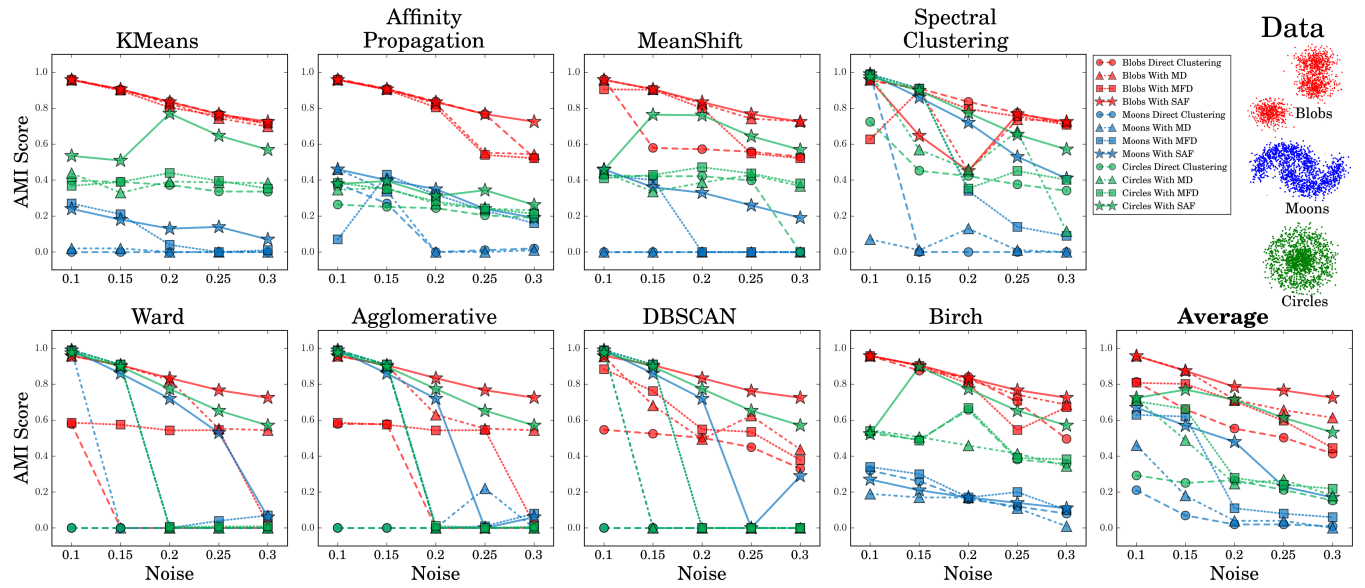


Fig. 10. Performance under different noise levels and for different datasets. We compare the clustering scores of our SAF consolidation to those of direct clustering, MD [21] and MFD [26]. SAF performs the best, especially when the noise level is high where the structures are better preserved by SAF. See also the supplemental material for visual comparisons.

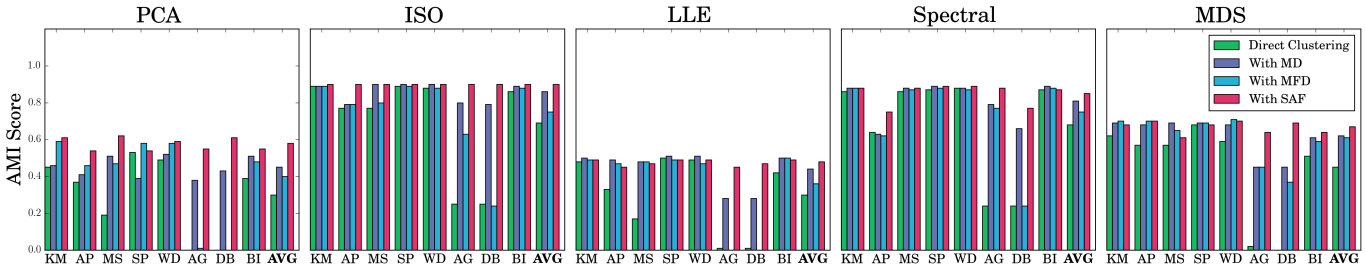


Fig. 11. Clustering the MINST data with different 3D embedding spaces using PCA, isomap [36], LLE [8], spectral [10], and MDS [37]. The AMI scores suggest that here agglomerative clustering (AG) and DBSCAN (DB) benefit most from consolidation, whereas spectral clustering does not.

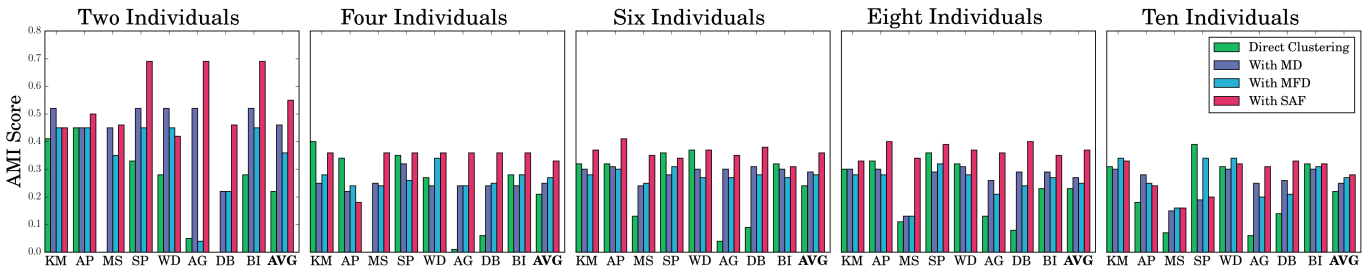


Fig. 12. Clustering the Yale face data. The AMI scores show the benefit of SAF in almost all cases.

4.2.3 Different Noise Levels

In Figure 10 we evaluate the performance of consolidation under different input noise levels. We use the data generator in the scikit-learn library to generate test data with desired Gaussian noise. As shown in the figure, our consolidation approach can substantially improve the clustering performance under a wide range of input noise levels. MD [21] and MFD [26] also improve the clustering performance under low noise levels, but SAF better preserves underlying structures when noise levels are high. Please refer to the supplementary for visual comparisons.

4.2.4 Different Embedding Spaces

As discussed above, we found that dimensionality reduction is important when dealing with high dimensional data. In the experiment in Figure 11 we investigate the influence of using different embedding spaces before clustering. We test on the MINST data set and project the 96 dimensional input data into 3D. Results suggest that using different dimensionality reduction techniques will not make a big impact on our consolidation method, as long as the underlying structure can be preserved in the embedding space. While MD [21] and MFD [26] also improve clustering performance, SAF shows an overall performance advantage.

4.2.5 Different Target Cluster Numbers

We test the clustering performance for different target cluster numbers using the extended Yale Face Dataset B [38]. This dataset contains 38 individuals and around 64 frontal images. We randomly selected 2, 4, 6 ,8 and 10 individuals from the dataset and report the AMI scores in Figure 12. The original face image is 1024 dimensions, which is projected onto a 9D affine subspace via PCA. The subspace is constructed by randomly selecting 1900 training images from the dataset. This pre-processing step was adapted and justified by Wang et al. [39]. In average, our method can best improve the clustering performance for different numbers of clusters compared to MD [21] and MFD [26]. All three methods work best for clustering only two groups of faces, where the underlying structure can be more easily found. Please refer to the supplementary for further details.

5 Conclusions

We present a novel structure-aware filtering (SAF) algorithm with applications in dimensionality reduction and clustering. We also provide a theoretical analysis of SAF that shows under what circumstances SAF converges to a point distribution on the underlying structures.

Through consolidating high-dimensional data, SAF can greatly facilitate existing dimensionality reduction and clustering techniques. Experiments on both synthetic and real data demonstrate that the performances of a variety of clustering algorithms are significantly boosted after SAF is applied, and SAF outperforms other state of the art techniques for manifold denoising.

Acknowledgments

We thank the anonymous reviewers for their constructive comments. This work was supported in part by Swiss National Science Foundation (169151), National Science Foundation of China (61522213, 61379090), NSFC-ISF Joint Research Program (6171101005, 2472/17), Natural Sciences and Engineering Research Council of Canada (2017-06086), Guangdong Science and Technology Program (2015A030312015), and Natural Science Foundation of Shenzhen University (827-000196). Hui Huang (hhzhiyan@gmail.com) is the corresponding author of this paper.

References

[1] D. Comaniciu and P. Meer, "Mean shift: a robust approach toward feature space analysis," *IEEE Trans. Pattern Analysis & Machine Intelligence*, vol. 24, no. 5, pp. 603–619, 2002.
[2] H.-P. Kriegel, P. Kröger, and A. Zimek, "Clustering high-dimensional data: A survey on subspace clustering, pattern-based clustering, and correlation clustering," *ACM Trans. Knowl. Discov. Data*, vol. 3, no. 1, pp. 1:1–1:58, 2009.
[3] A. K. Jain, "Data clustering: 50 years beyond k-means," *Pattern Recognition Letters*, vol. 31, no. 8, pp. 651–666, 2010.
[4] M. Charikar, S. Guha, E. Tardos, and D. B. Shmoys, "A constant-factor approximation algorithm for the k-median problem," *J. of Computer and System Sciences*, vol. 65, no. 1, pp. 129–149, 2002.
[5] B. G. Lindsay, "Mixture models: theory, geometry, and applications," *NSF-CBMS regional conference series in probability and statistics*, vol. 5, pp. i–163, 1995.
[6] T. Hastie, R. Tibhsirani, and R. Friedman, *The Elements of Statistical Learning*. Springer, 2009.
[7] M. Ester, H.-P. Kriegel, J. Sander, and X. Xu, "A density-based algorithm for discovering clusters in large spatial databases with noise." in *Proc. of ACM SIGKDD*, vol. 96, no. 34, 1996, pp. 226–231.
[8] S. T. Roweis and L. K. Saul, "Nonlinear dimensionality reduction by locally linear embedding," *Science*, vol. 290, no. 5500, pp. 2323–2326, 2000.
[9] J. Shi and J. Malik, "Normalized cuts and image segmentation," *IEEE Trans. Pattern Analysis & Machine Intelligence*, vol. 22, no. 8, pp. 888–905, 2000.
[10] M. Belkin and P. Niyogi, "Laplacian eigenmaps for dimensionality reduction and data representation," *Neural Comput.*, vol. 15, no. 6, pp. 1373–1396, 2003.
[11] R. R. Coifman and S. Lafon, "Diffusion maps," *Applied and Computational Harmonic Analysis*, vol. 21, no. 1, pp. 5–30, 2006.
[12] R. Kannan, S. Vempala, and A. Vetta, "On clusterings: Good, bad and spectral," *J. of the ACM*, vol. 51, no. 3, pp. 497–515, 2004.
[13] I. S. Dhillon, Y. Guan, and B. Kulis, "Kernel k-means: Spectral clustering and normalized cuts," in *Proc. of ACM SIGKDD*, 2004, pp. 551–556.
[14] B. Schölkopf, A. Smola, and K.-R. Müller, "Nonlinear component analysis as a kernel eigenvalue problem," *Neural Comput.*, vol. 10, no. 5, pp. 1299–1319, 1998.
[15] S. X. Yu and J. Shi, "Multiclass spectral clustering," in *Proc. Int. Conf. on Computer Vision*, 2003.
[16] Y. Lipman, D. Cohen-Or, D. Levin, and H. Tal-Ezer, "Parameterization-free projection for geometry reconstruction," *ACM Trans. on Graphics (Proc. of SIGGRAPH)*, vol. 26, no. 3, pp. 22:1–22:6, 2007.
[17] H. Huang, D. Li, H. Zhang, U. Ascher, and D. Cohen-Or, "Consolidation of unorganized point clouds for surface reconstruction," *ACM Trans. on Graphics (Proc. of SIGGRAPH Asia)*, vol. 28, no. 5, pp. 176:1–176:7, 2009.
[18] H. Huang, S. Wu, M. Gong, D. Cohen-Or, U. Ascher, and H. R. Zhang, "Edge-aware point set resampling," *ACM Trans. on Graphics*, vol. 32, no. 1, pp. 9:1–9:12, 2013.
[19] H. Huang, S. Wu, D. Cohen-Or, M. Gong, H. Zhang, G. Li, and B. Chen, "L1-medial skeleton of point cloud," *ACM Trans. on Graphics (Proc. of SIGGRAPH)*, vol. 32, no. 4, pp. 65:1–65:8, 2013.
[20] S. Wu, H. Huang, M. Gong, M. Zwicker, and D. Cohen-Or, "Deep points consolidation," *ACM Trans. on Graphics (Proc. of SIGGRAPH Asia)*, vol. 34, no. 6, pp. 176:1–176:13, 2015.
[21] M. Hein and M. Maier, "Manifold denoising," in *Advances in Neural Information Processing Systems*, 2007, pp. 561–568.
[22] W. Wang and M. A. Carreira-Perpinan, "Manifold blurring mean shift algorithms for manifold denoising," in *Proc. IEEE Conf. on Computer Vision & Pattern Recognition*, 2010, pp. 1759–1766.
[23] D. Gong, F. Sha, and G. G. Medioni, "Locally linear denoising on image manifolds." in *International Conference on Artificial Intelligence and Statistics*, 2010.

[24] Y. Wang, Y. Jiang, Y. Wu, and Z.-H. Zhou, "Spectral clustering on multiple manifolds," *IEEE Transactions on Neural Networks*, vol. 22, no. 7, pp. 1149–1161, 2011.
[25] B. Wang and Z. Tu, "Sparse subspace denoising for image manifolds," in *Proc. IEEE Conf. on Computer Vision & Pattern Recognition*, 2013, pp. 468–475.
[26] S. Deutsch, A. Ortega, and G. Medioni, "Manifold denoising based on spectral graph wavelets," in *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, 2016, pp. 4673–4677.
[27] E. J. Candès, X. Li, Y. Ma, and J. Wright, "Robust principal component analysis?" *Journal of the ACM (JACM)*, vol. 58, no. 3, p. 11, 2011.
[28] M. E. Tipping and C. M. Bishop, "Probabilistic principal component analysis," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 61, no. 3, pp. 611–622, 1999.
[29] D. Arthur and S. Vassilvitskii, "k-means++: The advantages of careful seeding," *Proc. ACM-SIAM Symp. on Discrete algorithms*, pp. 1027–1035, 2007.
[30] B. J. Frey and D. Dueck, "Clustering by passing messages between data points," *Science*, vol. 315, no. 5814, pp. 972–976, 2007.
[31] U. Von Luxburg, "A tutorial on spectral clustering," *Statistics and computing*, vol. 17, no. 4, pp. 395–416, 2007.
[32] F. Murtagh and P. Legendre, "Ward's hierarchical clustering method: clustering criterion and agglomerative algorithm," *J. of Classification*, vol. 31, no. 3, pp. 274–295, 2014.
[33] P. Franti, O. Virmajoki, and V. Hautamaki, "Fast agglomerative clustering using a k-nearest neighbor graph," *IEEE Trans. Pattern Analysis & Machine Intelligence*, vol. 28, no. 11, pp. 1875–1881, 2006.
[34] T. Zhang, R. Ramakrishnan, and M. Livny, "Birch: an efficient data clustering method for very large databases," in *Proc. of ACM SIGMOD*, vol. 25, no. 2, 1996, pp. 103–114.
[35] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine learning in Python," *J. of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.
[36] J. B. Tenenbaum, V. De Silva, and J. C. Langford, "A global geometric framework for nonlinear dimensionality reduction," *Science*, vol. 290, no. 5500, pp. 2319–2323, 2000.
[37] I. Borg and P. J. Groenen, *Modern multidimensional scaling: Theory and applications*. Springer Science & Business Media, 2005.
[38] A. S. Georghiades, P. N. Belhumeur, and D. J. Kriegman, "From few to many: Illumination cone models for face recognition under variable lighting and pose," *IEEE Trans. Pattern Analysis & Machine Intelligence*, vol. 23, no. 6, pp. 643–660, 2001.
[39] Y. Wang, Y.-X. Wang, and A. Singh, "A deterministic analysis of noisy sparse subspace clustering for dimensionality-reduced data," in *Proc. IEEE Int. Conf. on Machine Learning*, 2015, pp. 1422–1431.

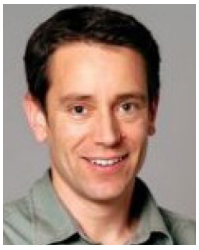
Shihao Wu is a Ph.D. candidate in the Computer Graphics Group in the University of Bern. He received his M.S. degree in South China University of Technology, and B.S. degree in the South China Normal University. His research interests include computer graphics, geometric modeling, machine learning, and point set processing.

Peter Bertholet is a Ph.D. candidate in the Computer Graphics Group in the University of Bern. He also received his M.S. and B.S. degrees in computer science and mathematics at the University of Bern. His research interests include computer graphics and computational geometry with range sensor data.

Hui Huang* (corresponding author) is a distinguished professor of Shenzhen University, where she directs the Visual Computing Research Center in College of Computer Science and Software Engineering. She received her PhD in Applied Math from The University of British Columbia in 2008 and another PhD in Computational Math from Wuhan University in 2006. She is the recipient of NSFC Excellent Young Researcher program and Guangdong Technology Innovation Leading Talent award in 2015. Her research interests are in computer graphics and scientific computing, focusing on point-based modeling, geometric analysis, 3D acquisition and creation.

Daniel Cohen-Or is a professor in the School of Computer Science. He received his B.S. in both mathematics and computer science in 1985, and M.S. in computer science in 1986 from Ben-Gurion University, and Ph.D. from the Department of Computer Science in 1991 at State University of New York at Stony Brook. He received the 2005 Eurographics Outstanding Technical Contributions Award. His research interests are in computer graphics, in particular, synthesis, processing and modeling techniques.

Minglun Gong is a professor of Computer Science at the Memorial University of Newfoundland, Canada. He obtained his M.S. from the Tsinghua University in 1997, and Ph.D. from the University of Alberta in 2003. His research interests cover various topics in the broad area of visual computing, including computer graphics, computer vision, visualization, image processing, and pattern recognition.



Matthias Zwicker is a professor at the Department of Computer Science, University of Maryland, College Park, where he holds the Reginald Allan Hahne Endowed E-nnovate chair. He obtained a PhD from ETH in Zurich, Switzerland. His research focus is on signal processing for high-quality rendering, point-based methods for rendering and modeling, and data-driven modeling and animation.