The Annals of Statistics
2016, Vol. 44, No. 1, 346–372
DOI: 10.1214/14-AOS1275

GRAPH CONNECTION LAPLACIAN METHODS CAN BE MADE ROBUST TO NOISE

BY NOUREDDINE EL KAROUI[1] AND HAU-TIENG WU[2]

University of California, Berkeley and University of Toronto

Recently, several data analytic techniques based on graph connection Laplacian (GCL) ideas have appeared in the literature. At this point, the properties of these methods are starting to be understood in the setting where the data is observed without noise. We study the impact of additive noise on these methods and show that they are remarkably robust. As a by-product of our analysis, we propose modifications of the standard algorithms that increase their robustness to noise. We illustrate our results in numerical simulations.

1. Introduction. In the last few years, several interesting variants of kernel-based spectral methods have appeared in the applied mathematics literature. These ideas were developed in connection with new types of data, where pairs of objects or measurements of interest have a relationship that is "blurred" by the action of a nuisance parameter. More specifically, we can find this type of data in a wide range of problems, for instance, in the class averaging algorithm for the cryo-electron microscope (cryo-EM) problem [34, 40], in a modern light source imaging technique known as ptychography [27], in graph realization problems [11, 12], in vectorized PageRank [9], in multi-channels image processing [4], etc....

Before we give further details about a specific motivating example, the cryo-EM problem, let us present the main building blocks of the methods we will study. They depend on the following three components:

(1) An undirected graph $\mathtt{G} = (\mathtt{V}, \mathtt{E})$ which describes all observations. The observations are the vertices of the graph $\mathtt{G}$, denoted as $\{V_i\}_{i=1}^n$.

(2) An *affinity function* $w : \mathtt{E} \to \mathbb{R}_+$, satisfying $w_{i,j} = w_{j,i}$, which describes how close two observations are (i and j index our observations). One common choice of $w_{i,j} = w(V_i, V_j)$ is of the form $w_{i,j} = \exp(-m(V_i, V_j)^2/\varepsilon)$, where $m(x, y)$ is a metric measuring how far x and y are.

(3) A *connection function* $r : \mathtt{E} \to \mathsf{G}$, where G is a Lie group, which describes how two samples are related. In its application to the cryo-EM problem, $r_{i,j}$'s can be thought of estimates of our nuisance parameters, which are orthogonal matrices in this example.

Received May 2014; revised September 2014.

[1] Supported by NSF Grant DMS-08-47647 (CAREER).

[2] Supported by AFOSR Grant FA9550-09-1-0643.

MSC2010 subject classifications. Primary 60F99; secondary 53A99.

Key words and phrases. Concentration of measure, random matrices, graph connection Laplacian, vector diffusion maps, spectral geometry, kernel methods.

These three components form *the connection graph* associated with the data, which is denoted as $(\mathtt{G}, w, r)$. They can be either given to the data analyst or have to be estimated from the data, depending on the application.

This fact leads to different connection graph and associated noise models. For example, in the cryo-EM problem, all components of the connection graph $(\mathtt{G}, w, r)$ are determined from the given projection images, where each vertex represents an image [18], Appendix A; in the ptychography problem [27], $\mathtt{G}$ is given by the experimenter, r is established from the experimental setup and w is built up from the diffractive images collected in the experiment. Depending on the application, different metrics, deformations or connections among pairs of observations are considered or estimated from the dataset to present the local information among data; see, for example, [3, 6, 10, 26, 28, 36, 37, 39].

We focus in this paper on the graph connection Laplacian (GCL), and hence we take the Lie Group $\mathsf{G} = O(k)$, where $k \in \mathbb{N}$, and assume that r satisfies $r_{i,j} = r_{j,i}^{-1}$.

1.1. *Motivating example*: *The Cryo-EM problem.* In the cryo-EM problem, the experimenter collects 2-dimensional projection images of a 3-dimensional macro-molecular object of interest, and the goal is to reconstruct the 3-dimensional geometric structure of the macro-molecular object from these projection images. Mathematically, the collected images $\mathcal{X}_{\text{cryoEM}} := \{I_i\}_{i=1}^N \in \mathbb{R}^{m^2}$ can be modeled as the X-ray transform of the potential of the macro-molecular object of interest, denoted as $\psi : \mathbb{R}^3 \to \mathbb{R}_+$. More precisely, in the setting that is usually studied, we have $I_i = X_\psi(R_i)$, where $R_i \in SO(3)$, $SO(3)$ is the 3-dimensional special orthogonal group and X_ψ is the X-ray transform of ψ. The X-ray transform $X_\psi(R_i)$ is a function from $\mathbb{R}^2$ to $\mathbb{R}_+$ and hence can be treated by the data analyst as an image. We refer the reader to [18], Appendix A, for precise mathematical details. (For the rest of the discussion, we write $R_i = [R_i^1 \quad R_i^2 \quad R_i^3]$ in the canonical basis, where R_i^k are three-dimensional unit vectors.)

The experimental process produces data with a high level of noise. Therefore, to solve this inverse problem, that is, to reconstruct ψ from $\{I_i\}_{i=1}^N$, it is a common consensus to preprocess the images to increase the signal-to-noise ratio (SNR) before sending them to the cryo-EM data analytic pipeline. An efficient way to do so is to estimate the projection directions of these images, that is, R_i^3. This direction plays a particular role in the X-ray transform, which is different from the other two directions. If R_i^3's were known, we would cluster the images according to these vectors and, for instance, take the mean of all properly rotationally aligned images in a cluster as a starting point for data-analysis. This would increase the SNR of the projection images. With these "improved" images, we can proceed to estimate R_i for the ith image by applying, for example, the common line algorithm [22], so that the 3-D image can be reconstructed by the inverse X-ray transform [20]. We note that R_i^3 is a unit vector in $\mathbb{R}^3$ and hence lives on the standard sphere S^2.

Conceptually, the problem is rendered difficult by the fact that the X-ray transform $X_\psi(R_i)$ is equivariant under the action of rotations that leave R_i^3 unchanged:

if r_θ is an in-plane rotation, that is, a rotation that leaves R_i^3 unchanged but rotates R_i^1 and R_i^2 by an angle θ, the image $X_\psi(r_\theta R_i)$ is $X_\psi(R_i)$ rotated by the angle θ. In other words, $X_\psi(r_\theta R_i) = r_2(\theta)X_\psi(R_i)$, where $r_2(\theta)$ stands for the 2-dimensional rotation by the angle θ. These in-plane rotations are clearly nuisance parameters if we want to evaluate the projection direction R_i^3.

To measure the distance between R_i^3 and R_j^3, we hence use a rotationally invariant distance, that is, $d_{i,j}^2 = \inf_{\theta\in[0,2\pi)} \|I_i - r_2(\theta)I_j\|_2^2$. More concretely, we look at the Euclidean distance between our two X-ray transforms/images after we have "aligned" them as best as possible. We now think of R_i^3's—the vectors we would like to estimate—as elements of the manifold S^2, equipped with a metric g_ψ, which depends on the macro-molecular object of interest. It turns out that the vector diffusion maps algorithm (VDM), which is based on GCL and which we study in this paper, is effective in producing a good approximation of g_ψ from the local information $d_{i,j}$'s and the rotations we obtain by aligning the various X-ray transforms. This, in turn, implies better clustering of the R_i^3's and improvement in the data-analytic pipeline for cryo-EM problems [34, 40].

1.2. *Motivation for the paper*: *Impact of noise on these procedures*. The point of this paper is to understand how the GCL algorithms perform when the input data is corrupted by noise. The relationship between this method and the connection concept in differential geometry is the following: the projection images I_i form a graph, and we can define the affinity and connection among a pair of images so that the topological structure of the 2-dimensional sphere (S^2, g_ψ) is encoded in the graph. This amounts to using the local geometric information derived from our data to estimate the global information—including the topology—of (S^2, g_ψ).

What is missing from these considerations and the current literature is an understanding of how noise impacts the procedures which are currently used and have mathematical backing in the noise-free context. The aim of our paper is to shed light on the issue of the impact of noise on these interesting and practically useful procedures. We will be concerned in this paper with the impact of adding noise on the observations, collected, for instance, in the way described above.

Note that additive noise may impact all three building blocks of the connection graph associated with the data. First, it might make the graph noisy. For example, in the cryo-EM problem, the standard algorithm builds up the graph from a given noisy data set, where $\{P_i\}_{i=1}^n = \{I_i + N_i\}_{i=1}^n - I_i$ is the signal, and N_i is additive noise, using the nearest neighbors determined by a pre-assigned metric. In other words, we put an edge between two vertices when they are close enough in that metric. Then, clearly, the existence of the noise N_i will likely create a different nearest-neighbor graph from the one that would be built up from the (clean) projection images $\{I_i\}_{i=1}^n$. As we will see in this paper, in some applications, it might be beneficial to consider a complete graph instead of a nearest-neighbor graph.

The second noise source is the way in which w and r are provided or determined from the samples. For example, in the cryo-EM problem, although $\{P_i\}_{i=1}^n$ are points located in a high-dimensional Euclidean space, we determine the affinity and connection between two images by evaluating their rotationally invariant distance. It is clear that when P_i is noisy, the derived affinity and connection will be noisy and likely quite different from the affinity and connection we would compute from the clean dataset $\{I_i\}_{i=1}^n$. On the other hand, in the ptychography problem, the connection is directly determined from the experimental setup, so that it is noise-free, even when our observations are corrupted by additive noise.

In summary, corrupting the observations by additive noise might impact the following elements of the data analysis:

(1) which scheme and metric we choose to construct the graph;
(2) how we build up the affinity function;
(3) how we build up the connection function.

1.3. *More details on GCL methods.* At a high-level, graph connection Laplacian (GCL) methods create a block matrix from the connection graph. The spectral properties of this matrix are then used to estimate properties of the intrinsic structure from which we posit the data is drawn from. This in turns lead to good estimation methods for, for instance, geodesic distance on the manifold, if the underlying intrinsic structure is a manifold. We refer the reader to the Supplementary Material [19] D and references [2, 8, 9, 31, 33] for more information.

Given a $n \times n$ symmetric matrix W, with scalar entries denoted by $w_{i,j}$ and a $nk \times nk$ block matrix G with $k \times k$ block entries denoted by $G_{i,j}$, we define a $nk \times nk$ matrix S with (i, j)-block entries

$$S_{i,j} = w_{i,j} G_{i,j}$$

and a $nk \times nk$ block diagonal matrix D with (i, i)-block entries

$$D_{i,i} = \sum_{j \neq i} w_{i,j} \mathrm{Id}_k,$$

which is assumed to be invertible. Let us call

$$L(W, G) := D^{-1} S \quad \text{and} \quad L_0(W, G) := L(W \circ 1_{i \neq j}, G). \tag{1}$$

In other words, $L_0(W, G)$ is the matrix $L(W, G)$ computed from the weight matrix W where the diagonal weights have been replaced by 0.

GCL terminology. Suppose we are given a connection graph $(\mathtt{G}, w, r)$, and construct the (symmetric) $n \times n$ *affinity matrix* W so that $w_{i,j} = w(i, j)$ and the *connection matrix* G, the $nk \times nk$ block matrix with $k \times k$ block entries $G_{i,j} = r(i, j)$. *The* (*normalized*) GCL associated with the connection graph $(\mathtt{G}, w, r)$ is defined as

$$\mathrm{Id}_{nk} - L(W, G). \tag{GCL}$$

The modified GCL associated with the connection graph $(\mathtt{G}, w, r)$ is defined as

$$\text{(modifGCL)} \qquad \mathrm{Id}_{nk} - L_0(W, G).$$

We note that under our assumptions on r, that is, $r_{i,j} = r_{j,i}^{-1} = r_{i,j}^*$, the connection matrix G is Hermitian. [Since W is symmetric, $L(W, G)$ and $L_0(W, G)$ are similar to Hermitian matrices.]

We are interested in the large eigenvalues of $L(W, G)$ [or, equivalently, the small eigenvalues of the GCL $\mathrm{Id}_{nk} - L(W, G)$] as well as the corresponding eigenvectors. In the case where the data is not corrupted by noise, the GCL's asymptotic properties have been studied in [31, 33], when the underlying intrinsic structure is a manifold. Its so-called synchronization properties have been studied in [2, 9].

1.4. *Organization of the paper.* We develop in Section 2 a theory for the impact of noise on a specific metric motivating this work, the rotationally invariant distance and related quantities. In Section 3, we give results concerning general GCL algorithms and propose modifications to the standard algorithms to render them more robust to noise. We present in Section 4 some numerical results, illustrating in part the points we raised in Sections 2 and 3.

Questions we address. The aim of our study is to understand the impact of additive noise on GCL algorithms. Our main results are Propositions 2.1 and 2.2, which explain the effect of noise on the affinity, and Theorem 2.2, which explains the effect of noise on the connection. These results are derived in the important case of the rotationally invariant distance. They lead to suggestions for modifying the standard GCL algorithms: the methods are more robust when we use a complete graph than when we use a nearest-neighbor graph, the latter being commonly used in practice. One should also use the matrix $L_0(W, G)$ instead of $L(W, G)$ to make the method more robust to noise. After we suggest these modifications, one main result is Theorem 3.1, which shows that even when the signal-to-noise-ratio (SNR) is very small, that is, going to 0 asymptotically, our modifications of the standard algorithm will approximately yield the same spectral results as if we had been working on the GCL matrix computed from noiseless data. Another important result in the paper is Theorem 3.2, which generalizes Theorem 3.1 to a broader class of GCL methods.

Notation. $\mathcal{T}$ denotes a set of linear transforms. Id_k stands for the $k \times k$ identity matrix. If $v \in \mathbb{R}^n$, $D(\{v\})$ is a $nk \times nk$ block diagonal matrix with the (i, i)th block equal to $v_i \mathrm{Id}_k$. We denote by $A \circ B$ the Hadamard, that is, the entry-wise, product of the matrices A and B. $|||M|||_2$ is the largest singular value (a.k.a. operator norm) of the matrix M. $\|M\|_F$ is its Frobenius norm. We use the probabilistic Landau notation o_P and O_P with the standard meaning; see, for example, [38], page 12 for definitions, if needed.

We now turn to the analysis of elements of a specific algorithm, the class averaging algorithm in the cryo-EM problem, with a broadly accepted model of noise contamination to demonstrate the impact of noise on this procedure.

2. Impact of additive noise on the rotationally invariant distance. We assume that we observe noisy versions of the k-dimensional images/objects, $k \geq 2$, we are interested in. If the images in the (unobserved) clean dataset are called $\{S_i\}_{i=1}^n$, we observe

$$I_i = S_i + N_i.$$

Here $\{N_i\}_{i=1}^n$ are pure-noise images/objects. Naturally, after discretization, the images/objects we consider are just data vectors of dimension p—we view S_i and N_i as vectors in $\mathbb{R}^p$. In other words, for a k-dimensional image, we sample p points from the domain $\mathbb{R}^k$ using the *sampling grid* $\mathcal{X} := \{\mathsf{x}_i\}_{i=1}^p \subset \mathbb{R}^k$, and the image is discretized according to these points. We also assume that the random variables N_i's, $i = 1, \dots, n$, are independent.

2.1. *Distance measurement between pairs of images.* We start from a general definition. Take a set of linear transforms $\mathcal{T}^{(k)} \subset O(k)$. Consider the following measurement between two objects/images, $d_{ij} \geq 0$, with

$$d_{ij}^2 = \inf_{\mathsf{o} \in \mathcal{T}^{(k)}} \|I_i - \mathsf{o} \circ I_j\|_2^2,$$

where $\circ$ means that the transform is acting on the pixels. For example, in the continuous setup where I_j is replaced by $f_j \in L^2(\mathbb{R}^k)$, given $\mathsf{o} \in SO(k)$, we have $\mathsf{o} \circ f_j(x) := f_j(\mathsf{o}^{-1}x)$ for all $x \in \mathbb{R}^k$. When $\mathcal{T}^{(k)} = SO(k)$, d_{ij} is called *the rotationally invariant distance* (*RID*).

Difficulties arising from discretization. In the discrete setup of interest in this paper, we assume that $\mathcal{X} = \mathsf{o}^{-1}\mathcal{X}$ for all $\mathsf{o} \in \mathcal{T}^{(k)}$; that is, the linear transform is *exact* (with respect to the grid $\mathcal{X}$), in that it maps the sampling grid onto itself. For concreteness, here is an example of sampling grid and associated exact linear transforms: Let $k = 2$, and take the sampling grid to be the polar coordinates grid. Since we are in dimension 2, we pick m rays of length 1 at angles $2\pi k/m$, $k = 0, \dots, m-1$ and have l equally spaced points on each of those rays. We consider I_i to be the discretization of the function $f_i \in L^2(\mathbb{R}^2)$ which is compactly supported inside the unit disk, at the polar coordinate grid. The set $\mathcal{T}^{(2)}$ consisting of elements of $SO(2)$ with angles $\theta_k = 2\pi \frac{k}{m}$, where $k = 1, \dots, m$, is thus exact and associated to the polar coordinate grid.

The discretization and notation merit further discussion. As a linear transform of the domain $\mathbb{R}^k$, $\mathsf{o} \in \mathcal{T}^{(k)}$ can be represented by a $k \times k$ matrix. On the other hand, in the discretized setup we consider here, we can map $\mathcal{T}^{(k)}$ to a set $\mathcal{T}$ of $p \times p$ matrices O which act on the discretized images I_j. These images are viewed as a set of p-dimensional vectors, denoted as $I_j^\vee$, and O acts on a "flattened" or "vectorized" (i.e., 1-dimensional) version of the k-dimensional object of interest. Note that to each transform o there corresponds a unique $p \times p$ matrix O. In the following, we will use o to denote the transform acting on the pixels, and use O to mean its companion matrix acting on the vectorized version of the object we are interested in. A simple but very important observation is that

$$(\mathsf{O} \circ I_i)^\vee = O I_i^\vee .$$

In other words, we will have $\inf_{\mathsf{O} \in \mathcal{T}^{(k)}} \|I_i - \mathsf{O} \circ I_j\| = \inf_{\mathsf{O} \in \mathcal{T}} \|I_i^\vee - \mathsf{O} I_j^\vee\|$. To simplify the notation, when it is clear from the context, we will use I_j to mean both the discretized object of interest and its vectorized version.

Approximation results. In what follows, we assume that $\mathcal{T}$ always contains Id_p. We study the impact of noise on d_{ij} through a uniform approximation argument. Let us call for $\mathsf{O} \in \mathcal{T}$,

$$d^2_{ij,\text{noisy}}(\mathsf{O}) := \|I_i^\vee - \mathsf{O} I_j^\vee\|^2 \quad \text{and} \quad d^2_{ij,\text{clean}}(\mathsf{O}) := \|S_i^\vee - \mathsf{O} S_j^\vee\|^2 .$$

Essentially we will show that, when $\mathcal{T}$ contains only orthogonal matrices and is not "too large,"

$$\sup_{\mathsf{O} \in \mathcal{T}} \sup_{i \neq j} \left| d^2_{ij,\text{noisy}}(\mathsf{O}) - d^2_{ij,\text{clean}}(\mathsf{O}) - f(i,j) \right| = \mathrm{o}_P(1),$$

where $f(i,j)$ does not depend on O. Our approximations will in fact be much more precise than this, but we will be able to conclude that in these circumstances,

$$\sup_{i \neq j} \left| \inf_{\mathsf{O} \in \mathcal{T}} d^2_{ij,\text{noisy}}(\mathsf{O}) - \inf_{\mathsf{O} \in \mathcal{T}} d^2_{ij,\text{clean}}(\mathsf{O}) - f(i,j) \right| = \mathrm{o}_P(1).$$

We have the following theorem for any given set of transforms $\mathcal{T}$.

THEOREM 2.1. *Suppose that for* $1 \leq i \leq n$, N_i *are independent*, *with* $N_i^\vee \sim \mathcal{N}(0, \Sigma_i)$. *Call*

$$\mathsf{t}_\mathsf{p} := \sup_i \sup_{\mathsf{O} \in \mathcal{T}} \sqrt{\operatorname{trace}\left((\mathsf{O}\Sigma_i\mathsf{O}')^2\right)}$$

and

$$\mathsf{s}_p := \sup_{1 \leq i \leq n} \sup_{\mathsf{O} \in \mathcal{T}} \sqrt{|||\mathsf{O}\Sigma_i\mathsf{O}'|||_2}.$$

Then we have

$$\sup_{\mathsf{O} \in \mathcal{T}} \sup_{i \neq j} \left| d^2_{ij,\text{noisy}}(\mathsf{O}) - d^2_{ij,\text{clean}}(\mathsf{O}) - \operatorname{trace}\left(\Sigma_i + \mathsf{O}\Sigma_j\mathsf{O}'\right) \right|$$

$$= \mathrm{O}_P\left(\sqrt{\log[\operatorname{Card}\{\mathcal{T}\} n^2]}\left(\mathsf{t}_\mathsf{p} + \mathsf{s}_p \sup_{i, \mathsf{O} \in \mathcal{T}} \|\mathsf{O} S_i^\vee\|\right) + \log[\operatorname{Card}\{\mathcal{T}\} n^2] \mathsf{s}_p^2\right).$$

A proof of this theorem can be found in Section A-1 in the Supplementary Material [19]. In light of the previous theorem, we have the following proposition.

PROPOSITION 2.1. *Suppose that for all* $1 \le i \le n$ *and* $\mathsf{O} \in \mathcal{T}$, $|||\mathsf{O}\Sigma_i\mathsf{O}'|||_2 \le \sigma_p^2$, $\sqrt{\mathrm{trace}([\mathsf{O}\Sigma_i\mathsf{O}']^2)/p} \le s_p^2$ *and* $\|\mathsf{O}S_i^\vee\| \le K$, *where* K *is a constant independent of* p. *Then*

$$\sup_{\mathsf{O}\in\mathcal{T}}\sup_{i\neq j}\left|d^2_{ij,\mathrm{noisy}}(\mathsf{O}) - d^2_{ij,\mathrm{clean}}(\mathsf{O}) - \mathrm{trace}(\Sigma_i + \mathsf{O}\Sigma_j\mathsf{O}')\right| = \mathrm{O}_P(u_{n,p}), \tag{2}$$

where $u_{n,p} := \sqrt{\log[\mathrm{Card}\{\mathcal{T}\}n^2]}(\sqrt{p}s_p^2 + K\sigma_p) + \log[\mathrm{Card}\{\mathcal{T}\}n^2]\sigma_p^2$.

It follows that, if $\sqrt{\log[\mathrm{Card}\{\mathcal{T}\}n^2]}\max(\sqrt{p}s_p^2, \sigma_p) \to 0$, *and* $\mathcal{T}$ *contains only orthogonal matrices*,

$$\sup_{\mathsf{O}\in\mathcal{T}}\sup_{i\neq j}\left|d^2_{ij,\mathrm{noisy}}(\mathsf{O}) - d^2_{ij,\mathrm{clean}}(\mathsf{O}) - \mathrm{trace}(\Sigma_i + \Sigma_j)\right| = \mathrm{O}_P(u_{n,p}) = \mathrm{o}_P(1).$$

Furthermore, in this case,

$$\sup_{i\neq j}\left|d^2_{ij,\mathrm{noisy}} - d^2_{ij,\mathrm{clean}} - \mathrm{trace}(\Sigma_i + \Sigma_j)\right| = \mathrm{o}_P(1),$$

where

$$d^2_{ij,\mathrm{noisy}} := \inf_{\mathsf{O}\in\mathcal{T}}\|I_i^\vee - \mathsf{O}I_j^\vee\|^2,$$
$$d^2_{ij,\mathrm{clean}} := \inf_{\mathsf{O}\in\mathcal{T}}\|S_i^\vee - \mathsf{O}S_j^\vee\|^2.$$

The following set of assumptions is natural in light of the previous proposition:

ASSUMPTION G1. $\forall i, \mathsf{O} \in \mathcal{T}$, $|||\mathsf{O}\Sigma_i\mathsf{O}'|||_2 \le \sigma_p^2$, $\sqrt{\mathrm{trace}([\mathsf{O}\Sigma_i\mathsf{O}']^2)/p} \le s_p^2$ and $\|\mathsf{O}S_i^\vee\| \le K$, where K is a constant independent of p. Furthermore, $\sqrt{\log[\mathrm{Card}\{\mathcal{T}\}n^2]}\max(\sqrt{p}s_p^2, \sigma_p) \to 0$ and hence $u_{n,p} \to 0$.

We refer the reader to Proposition C.1 on page 16 in the Supplementary Material [19] for a bound on Card$\{\mathcal{T}\}$ that is relevant to the class averaging algorithm in the cryo-EM problem.

PROOF OF PROPOSITION 2.1. The first two statements are immediate consequences of Theorem 2.1. For the second one, we use the fact that since $\mathsf{O} \in \mathcal{T}$ is orthogonal, $\mathrm{trace}(\mathsf{O}\Sigma_j\mathsf{O}') = \mathrm{trace}(\Sigma_j)$.

Now, if F and G are two functions, we clearly have $|\inf F(x) - \inf G(x)| \le \sup|G(x) - F(x)|$. Indeed, $\forall x$, $F(x) \le G(x) + \sup|G(x) - F(x)|$. Hence, for all x,

$$\inf_x F(x) \le F(x) \le G(x) + \sup\left|G(x) - F(x)\right|,$$

and we conclude by taking inf in the right-hand side. The inequality is proved similarly in the other direction. The results of Theorem 2.1 therefore show that

$$\sup_{i\neq j}\left|d^2_{ij,\text{noisy}} - d^2_{ij,\text{clean}} - \operatorname{trace}(\Sigma_i + \Sigma_j)\right|$$
$$= \mathrm{O}_P\left(\sqrt{\log[\operatorname{Card}\{\mathcal{T}\}n^2]}\left(\sqrt{p}s_p^2 + K\sigma_p\right) + \log[\operatorname{Card}\{\mathcal{T}\}n^2]\sigma_p^2\right)$$

and we get the announced conclusions under our assumptions. □

We now present two examples to show that our assumptions are quite unrestrictive. This will later help to prove that the algorithms we are studying can tolerate very large amounts of noise.

Magnitude of noise: *First example.* Assume that $N_i^\vee \sim p^{-(1/4+\varepsilon)}\mathcal{N}(0, \mathrm{Id}_p)$, where $\varepsilon > 0$. In this case, $\|N_i^\vee\| \sim p^{1/4-\varepsilon} \gg \sup_i \|S_i^\vee\|$ if $\varepsilon < 1/4$. In other words, the norm of the error vector is much larger than the norm of the signal vector. Indeed, asymptotically, the signal to noise ratio $\|S_i^\vee\|/\|N_i^\vee\|$ is 0. Furthermore, $\sigma_p = p^{-(1/4+\varepsilon)}$ and $\sqrt{p}s_p^2 = p^{-2\varepsilon}$. Hence, if $\operatorname{Card}\{\mathcal{T}\} = \mathrm{O}(p^\gamma)$ for some γ, our conditions translate into $\sqrt{\log(np)}\max(p^{-(1/4+\varepsilon)}, p^{-2\varepsilon}) \to 0$. This is of course satisfied provided n is subexponential in p. See Proposition C.1 in [19] for a natural example of $\mathcal{T}$ whose cardinal is polynomial in p.

Magnitude of noise: *Second example.* We now consider the case where Σ_i has one eigenvalue equal to $p^{-\varepsilon}$, and all the others are equal to $p^{-(1/2+\eta)}$, $\varepsilon, \eta > 0$. In other words, the noise is much larger in one direction than in all the others. In this case, $\sigma_p^2 = p^{-\varepsilon}$ and $\operatorname{trace}(\Sigma_i^2) = p^{-2\varepsilon} + (p-1) * p^{-(1+2\eta)} \leq p^{-2\varepsilon} + p^{-2\eta}$. So if once again, $\operatorname{Card}\{\mathcal{T}\} = \mathrm{O}(p^\gamma)$, our conditions translate into $\sqrt{\log(np)}\max(p^{-\varepsilon} + p^{-\eta}, p^{-\varepsilon/2}) \to 0$. This example would also work if the number of eigenvalues equal to $p^{-\varepsilon}$ were $\mathrm{o}(p^{2\varepsilon}/[\log(np)])$, provided $\sqrt{\log(np)}\max(p^{-\eta}, p^{-\varepsilon/2}) \to 0$.

Comment on the conditions on the signal in Assumption G1. At first glance, it might look like the condition $\sup_{i,\mathsf{O}\in\mathcal{T}} \|\mathsf{O}S_i^\vee\| \leq K$ is very restrictive due to the fact that, after discretization, S_i has p pixels. However, it is typically the case that if we start from a function in $L^2(\mathbb{R}^k)$, the discretized and vectorized image $S_i^\vee$ is normalized by the number of pixels p, so that $\|S_i^\vee\|$ is roughly equal to the L^2-norm of the corresponding function. Hence, our condition $\sup_{i,\mathsf{O}\in\mathcal{T}} \|\mathsf{O}S_i^\vee\| \leq K$ is very reasonable.

2.2. *The case of "exact rotations."* We now focus on the most interesting case for our problem, namely the situation where O leaves our sampling grid invariant. We call $\mathcal{T}^{(k)}_{\text{exact}} \subset SO(k)$ the corresponding matrices O and $\mathcal{T}_{\text{exact}}$ the companion $p \times p$ matrices. We note that $\mathcal{T}^{(k)}_{\text{exact}}$ depends on p, but since this is evident, we do not index $\mathcal{T}^{(k)}_{\text{exact}}$ by p to avoid cumbersome notation. From the standpoint of statistical applications, our focus in this paper is mostly on the case $k = 1$ (which corresponds to "standard" kernel methods commonly used in statistical learning) and $k = 2$.

We show in Proposition C.1 in [19] that if $\mathsf{O} \in \mathcal{T}_{\text{exact}}$, O is an orthogonal $p \times p$ matrix. Furthermore, we show in Proposition C.1 in [19] that $\text{Card}\{\mathcal{T}_{\text{exact}}\}$ is polynomial in p. We therefore have the following proposition.

PROPOSITION 2.2. *Let*

$$d^2_{ij,\text{noisy}} := \inf_{\mathsf{O} \in \mathcal{T}^{(k)}_{\text{exact}}} \|I_i - \mathsf{O} \circ I_j\|^2, \qquad d^2_{ij,\text{clean}} := \inf_{\mathsf{O} \in \mathcal{T}^{(k)}_{\text{exact}}} \|S_i - \mathsf{O} \circ S_j\|^2.$$

Suppose N_i are independent with $N_i^\vee \sim \mathcal{N}(0, \Sigma_i)$. When Assumption G1 holds with $\mathcal{T}_{\text{exact}}$ being the set of companion matrices of $\mathcal{T}^{(k)}_{\text{exact}}$, we have

$$\sup_{i \neq j} \left| d^2_{ij,\text{noisy}} - d^2_{ij,\text{clean}} - \text{trace}(\Sigma_i + \Sigma_j) \right| = \mathrm{o}_P(1)$$

and

$$\sup_{\mathsf{O} \in \mathcal{T}^{(k)}_{\text{exact}}} \sup_{i \neq j} \left| d^2_{ij,\text{noisy}}(\mathsf{O}) - d^2_{ij,\text{clean}}(\mathsf{O}) - \text{trace}(\Sigma_i + \Sigma_j) \right| = \mathrm{O}_P(u_{n,p}) = \mathrm{o}_P(1).$$

2.3. *On the transform $\mathsf{O}^*_{ij,\text{noisy}}$.* We now use the notation

$$d_{ij,\text{noisy}}(\mathsf{O}) = \|I_i - \mathsf{O} \circ I_j\| \quad \text{and} \quad d_{ij,\text{clean}}(\mathsf{O}) = \|S_i - \mathsf{O} \circ S_j\|.$$

Naturally, the study of

$$\mathsf{O}^*_{ij,\text{noisy}} = \operatorname*{argmin}_{\mathsf{O} \in \mathcal{T}^{(k)}_{\text{exact}}} d_{ij,\text{noisy}}(\mathsf{O}) \tag{3}$$

is more complicated than the study of $\inf_{\mathsf{O} \in \mathcal{T}^{(k)}_{\text{exact}}} d_{ij,\text{noisy}}(\mathsf{O})$.

We will assume that the clean/noise-free images are nicely behaved when it comes to the $d_{ij,\text{clean}}(\mathsf{O})$ minimization, in that rotations that are near minimizers of $d_{ij,\text{clean}}(\mathsf{O})$ are close to one another. More formally, we assume the following.

ASSUMPTION A0. $\mathcal{T}^{(k)}_{\text{exact}}$ is a subset of $SO(k)$ and contains only exact rotations. Call $\mathsf{O}^*_{ij,\text{clean}} := \operatorname{argmin}_{\mathsf{O} \in \mathcal{T}^{(k)}_{\text{exact}}} d^2_{ij,\text{clean}}(\mathsf{O})$ and call $\mathcal{T}^{(k)}_{ij,\varepsilon} := \{\mathsf{O} \in \mathcal{T}^{(k)}_{\text{exact}} : d^2_{ij,\text{clean}}(\mathsf{O}) \leq d^2_{ij,\text{clean}}(\mathsf{O}^*_{ij,\text{clean}}) + \varepsilon\}$. We assume that

$$\exists \delta_{ij,p} > 0: \qquad \forall \varepsilon < \delta_{ij,p}, \forall \mathsf{O} \in \mathcal{T}^{(k)}_{ij,\varepsilon} \qquad d(\mathsf{O}, \mathsf{O}^*_{ij,\text{clean}}) \leq g_{ij,p}(\varepsilon),$$

for d the canonical metric on the orthogonal group and some positive $g_{ij,p}(\varepsilon)$.

ASSUMPTION A1. $\delta_{ij,p}$ can be chosen independently of i, j and p. Furthermore, there exists a function g such that $g(\varepsilon) \to 0$ as $\varepsilon \to 0$ and $g_{ij,p}(x) \leq g(x)$, if $x \leq \delta_{ij,p} \leq \delta$.

We discuss the meaning of these assumptions after the following theorem.

THEOREM 2.2. *Suppose that the assumptions underlying Theorem 2.1 hold and that Assumptions G1, A0 and A1 hold. Suppose further that $\mathcal{T}_{\text{exact}}^{(k)}$ is the set of exact rotations for our discretization. Then, for any η given, where $0 < \eta < 1$, as p and n go to infinity,*

$$\sup_{i \neq j} d\big(\mathrm{O}_{ij,\text{noisy}}^{*}, \mathrm{O}_{ij,\text{clean}}^{*}\big) = \mathrm{O}_P\big(g\big(u_{n,p}^{1-\eta}\big)\big), \tag{4}$$

where $u_{n,p}$ is defined in (2). (Under Assumption G1, $u_{n,p} \to 0$ as n and p tend to infinity.)

The informal meaning of this theorem is that under regularity assumptions on the set of clean/noise-free images, the optimal rotation computed from the set of noisy images is close to the optimal rotation computed from the set of clean/noise-free images. In other words, this step of the GCL procedure is robust to noise.

Interpretation of Assumptions A0–A1. Assumption A0 guarantees that all near minimizers of $d_{ij,\text{clean}}(\mathrm{O})$ are close to one another and hence the optimum. Our uniform bounds in Proposition 2.2 only guarantee that $\mathrm{O}_{ij,\text{noisy}}^{*}$ is a near minimizer of $d_{ij,\text{clean}}(\mathrm{O})$ and nothing more. If $d_{ij,\text{clean}}(\mathrm{O})$ had near minimizers that were far from the optimum $\mathrm{O}_{ij,\text{clean}}^{*}$, it could very well happen that $\mathrm{O}_{ij,\text{noisy}}^{*}$ end up being close to one of these near minimizers but far from $\mathrm{O}_{ij,\text{clean}}^{*}$, and we would not have the consistency result of Theorem 2.2. Hence, the robustness to noise of this part of the GCL algorithm is clearly tied to some regularity or "niceness" property for the set of clean/noise-free images.

In the cryo-EM problem, these assumptions reflect a fundamental property of a manifold dataset, *its condition number* [29]. Conceptually, the condition number reflects "how difficult it is to reconstruct a manifold" from a finite sample of points from that manifold. Precisely, it is the inverse of the reach of the manifold, which is defined to be the radius of the smallest normal bundle that is homotopic to the manifold. This also highlights the fact that even if we were to run the GCL algorithm on the clean/noise-free dataset, without these assumptions, the results might not be stable and reliable since intrinsically distant points (i.e., distant in the geodesic distance) might be identified as neighbors.

About $\mathcal{T}_{\text{exact}}^{(k)}$ and extensions. We are chiefly interested in this paper about 2-dimensional images and hence about the case $k = 2$; see the cryoEM example. It is then clear that when our polar coordinate grid is fine, $\mathcal{T}_{\text{exact}}^{(k)}$ is also a fine discretization of $SO(2)$ and contains many elements. (More details are given in Section C-3 in [19].) The situation is more intricate when $k \geq 3$, but since it is a bit tangential to the main purpose of the current paper, we do not discuss it further here. We refer the interested reader to Section C-3 in [19] for more details about the case $k \geq 3$.

We also note that our arguments are not tied to using a standard polar coordinate grid for the discretization of the images. For another sampling grid, we would possibly get another $\mathcal{T}^{(k)}_{\text{exact}}$. Our arguments go through when: (a) if $\mathcal{O} \in \mathcal{T}^{(k)}$, the operation $\mathcal{O}\circ$ maps our sampling grid of points onto itself and (b) $\text{Card}\{\mathcal{T}^{(k)}\}$ grows polynomially in p.

2.4. *Extensions and different approaches.* Central to our arguments are strong concentration results for quadratic forms in Gaussian random variables. Naturally, our results extend to other types of random variables for which these concentration properties hold. We refer to [25] and [15] for examples. A natural example in our context would be a situation where $N_i = \Sigma_i^{1/2} X_i$, and X_i has i.i.d. uniformly bounded entries. This is particularly relevant in the case where Σ_i is diagonal, for instance, the interpretation being then that the noise contamination is through the corruption of each individual pixel by independent random variables with possibly different standard deviations. The arguments in Lemma C-1 in [19] handle this case, though the bound is slightly worse than the one in Lemma C-2 in [19] when a few eigenvalues of Σ_i are larger than most of the others. Indeed, the only thing that matters in this more general analysis is the largest eigenvalue of Σ_i, so that in the notation of Assumption G1, $\sqrt{p}s_p^2$ is replaced by $\sqrt{p}\sigma_p^2$. Hence our approximation will require in this more general setting that $\sigma_p = \mathrm{o}(p^{-1/4})$, whereas we have seen in the Gaussian case that we can tolerate a much larger largest eigenvalue.

We also note that we could of course settle for weaker results on concentration of quadratic forms, which would apply to more distributions. For instance, using bounds on $\mathbf{E}(|\|N_i\|^2 - \mathbf{E}(\|N_i\|^2)|^k)$ would change the dependence of results such as Proposition 2.1 on $\mathfrak{b} \triangleq \text{Card}\{\mathcal{T}\}n^2$ from powers of $\log(\mathfrak{b})$ to powers of $\mathfrak{b}^{1/k}$. This is in turn would mean that our results would become tolerant to lower levels of noise but apply to more noise distributions.

3. Robustness theory for general GCL problems. Our aim in this section is to develop a theory that explains the behavior of GCL algorithms in the presence of noise. In particular, it will apply to algorithms of the cryo-EM type. We give in Section 3.1 approximation results that apply to general GCL problems. In Section 2, we study in detail the impact of noise on both the affinity and the connection used in the computation of the GCL when using the rotationally invariant distance (this is particularly relevant for the cryo-EM problem). We put these two sets of results together for a detailed study of GCL algorithms in the presence of noise in Section 3.2. We also propose in Section 3.2 modifications to the standard algorithms that increase the robustness to noise of GCL methods.

3.1. *General approximation results.* We first present two results that apply generally to GCL algorithms and are not related to specific affinity or connection functions.

LEMMA 3.1. *Suppose W and $\tilde{W}$ are $n \times n$ matrices, with scalar entries denoted by $w_{i,j}$ and $\tilde{w}_{i,j}$ and G and $\tilde{G}$ are $nd \times nd$ block matrices, with $d \times d$ blocks denoted by $G_{i,j}$ and $\tilde{G}_{i,j}$. We assume that*

$$\exists\{f_i\}_{i=1}^n, f_i > 0: \qquad \sup_{i,j}\left|\frac{\tilde{w}_{i,j}}{f_i} - w_{i,j}\right| \le \varepsilon \quad \textit{and} \quad \sup_{i,j}\|\tilde{G}_{i,j} - G_{i,j}\|_F \le \eta.$$

Suppose furthermore that there exists $C > 0$ such that $0 \le w_{i,j} \le C$, $\sup_{i,j}\|G_{i,j}\|_F \le C$ and $\sup_{i,j}\|\tilde{G}_{i,j}\|_F \le C$. Then, if $\inf_i \sum_{j\neq i} w_{i,j}/n > \gamma$ and $\gamma > \varepsilon$, we have, with the notation of equation (1),

$$|||L(W,G) - L(\tilde{W},\tilde{G})|||_2 \le \frac{1}{\gamma}C(\eta+\varepsilon) + \frac{\varepsilon}{\gamma(\gamma-\varepsilon)}C^2 .$$

We note that quite remarkably, there are essentially no conditions on f_i's: in particular, $\tilde{w}_{i,j}$ and $w_{i,j}$ could be of completely different magnitudes. The previous lemma also shows that, for the purpose of understanding the large eigenvalues and eigenvectors of $L(W,G)$, we do not need to estimate f_i's: we can simply use $L(\tilde{W},\tilde{G})$, that is, just work with the noisy data.

In the case $f_i = 1$ for all i's, this lemma says that if we can approximate the matrix W well entrywise and each of the individual matrices $G_{i,j}$ well too, data analytic techniques working on the GCL matrix $L(\tilde{W},\tilde{G})$ will do essentially as well as those working on the corresponding matrix for $L(W,G)$ in the spectral sense.

This result is useful because many methods rely on these connection graph ideas, with different input in terms of affinity and connection functions [6, 7, 10–12, 26, 28, 34, 36, 37, 40]. However, it will often be the case that we can approximate $w_{i,j}$—which we think of as measurements we would get if our signals were not corrupted by noise—only up to a constant, which is why we need to allow for the presence of f_i's.

PROOF OF LEMMA 3.1. Let us call $\tilde{W}_f$ the matrix with scalar entries $\tilde{w}_{i,j}/f_i$. We note simply that

$$L(\tilde{W}_f,\tilde{G}) = L(\tilde{W},\tilde{G}).$$

The assumptions of Lemma C-3 in the Supplementary Material [19] apply to $(\tilde{W}_f,\tilde{G})$, and hence we have

$$|||L(W,G) - L(\tilde{W}_f,\tilde{G})|||_2 \le \frac{1}{\gamma}C(\eta+\varepsilon) + \frac{\varepsilon}{\gamma(\gamma-\varepsilon)}C^2.$$

But since $L(\tilde{W}_f, \tilde{G}) = L(\tilde{W}, \tilde{G})$, we also have

$$|||L(W, G) - L(\tilde{W}, \tilde{G})|||_2 \leq \frac{1}{\gamma} C(\eta + \varepsilon) + \frac{\varepsilon}{\gamma(\gamma - \varepsilon)} C^2. \qquad \square$$

In some situations that will be of interest to us below, it is, however, not the case that we can find f_i's such that

$$\exists \{f_i\}_{i=1}^n, f_i > 0: \qquad \sup_{i,j} \left| \frac{\tilde{w}_{i,j}}{f_i} - w_{i,j} \right| \leq \varepsilon.$$

Rather, this approximation is possible only when $i \neq j$, yielding the condition

$$\forall i, \exists f_i > 0: \qquad \sup_{i \neq j} \left| \frac{\tilde{w}_{i,j}}{f_i} - w_{i,j} \right| \leq \varepsilon.$$

This apparently minor difference turns out to have significant consequences, both practical and theoretical. We propose in the following lemma to modify the standard way of the computing the GCL matrix to handle this more general case.

LEMMA 3.2. *We work under the same setup as in Lemma 3.1 and with the same notation. We now assume that multiplicative approximations of the weights is possible only on the off-diagonal elements of our weight matrix*

$$\exists \{f_i\}_{i=1}^n, f_i > 0: \qquad \sup_{i \neq j} \left| \frac{\tilde{w}_{i,j}}{f_i} - w_{i,j} \right| \leq \varepsilon \quad \textit{and} \quad \sup_{i,j} \|\tilde{G}_{i,j} - G_{i,j}\|_F \leq \eta.$$

Suppose furthermore that there exists $C > 0$ such that $0 \leq w_{i,j} \leq C$, $\sup_{i,j} \|G_{i,j}\|_F \leq C$, and $\sup_{i,j} \|\tilde{G}_{i,j}\|_F \leq C$. Then if $\inf_i \sum_{j \neq i} w_{i,j}/n > \gamma$ and $\gamma > \varepsilon$, we have

$$|||L_0(W, G) - L_0(\tilde{W}, \tilde{G})|||_2 \leq \frac{1}{\gamma} C(\eta + \varepsilon) + \frac{\varepsilon}{\gamma(\gamma - \varepsilon)} C^2$$

and

$$|||L(W, G) - L_0(\tilde{W}, \tilde{G})|||_2 \leq \frac{1}{\gamma} C(\eta + \varepsilon) + \frac{\varepsilon}{\gamma(\gamma - \varepsilon)} C^2 + \frac{C^2}{n\gamma}.$$

The lemma is shown in Section A-3 in the Supplementary Material [19].

Comment. Concretely, this lemma means that if we do not include the block diagonal terms in the computation of the GCL obtained from our "noisy data," that is, $(\tilde{W}, \tilde{G})$, we will get a matrix that is very close in spectral norm to the GCL computed from the "clean/noise-free data," that is, (W, G). The significance of this result lies in the fact that recent work in applied mathematics has proposed to use the large eigenvalues and eigenvectors of $L(W, G)$ for various data analytic tasks, such as the estimation of local geodesic distances when the data is thought to be sampled from an unknown manifold.

What our result shows is that even when f_i are arbitrarily large, which we can think of as the situation where the signal to noise ratio in $\tilde{W}$ is basically 0, working with $L_0(\tilde{W}, \tilde{G})$ will allow us to harness the power of these recently developed tools. Naturally, working with $(\tilde{W}, \tilde{G})$ is a much more realistic assumption than working with (W, G) since we expect all our measurements to be noisy in practice. Results based on (W, G) essentially assume that there is no noise in the dataset.

Finally, the previous lemma also suggests that practitioners not use nearest-neighbor information when using GCL methods. Indeed, the nearest-neighbor information is generally different for noisy and noise-free datasets, and incorporating it would damage or destroy the spectral approximate-equivalence results of Lemma 3.2.

We provide a simple extension that may be useful from a practical standpoint in Lemma A-2 in the Supplementary Material [19], Section A-3.1: in the case where $L_0(W, G)$ can be approximated by a sparser matrix, we are able to weaken the requirement of uniform approximation of $G_{i,j}$'s by $\tilde{G}_{i,j}$'s.

3.2. *Consequences for GCL algorithm and other kernel-based methods.*

3.2.1. *Reminders and preliminaries* Recall that in GCL methods performed with the rotationally invariance distance induced by $SO(k)$ (henceforth RID), we mostly care about the spectral properties—especially large eigenvalues and corresponding eigenvectors—of the GCL matrix $L(\tilde{W}, \tilde{G})$, where $\tilde{W}$ is a $n \times n$ matrix, and $\tilde{G}$ is a $nk \times nk$ block-matrix with $k \times k$ blocks defined through

$$\tilde{W}_{i,j} = \exp(-d^2_{ij,\text{noisy}}/\varepsilon), \qquad \tilde{G}_{i,j} = \mathrm{O}^*_{ij,\text{noisy}},$$

where $\mathrm{O}^*_{ij,\text{noisy}}$ is defined in equation (3).

The "good" properties of GCL stem from the fact that the matrix $L(W, G)$, the GCL matrix associated with the clean/noise-free images, has "good" spectral properties. For example, when a manifold structure is assumed, the theoretical work in [31, 33] relates the properties of $L(W, G)$—the matrix obtained in the same manner as above when we replace $d_{ij,\text{noisy}}$ by $d_{ij,\text{clean}}$ and $\mathrm{O}^*_{ij,\text{noisy}}$ by $\mathrm{O}^*_{ij,\text{clean}}$—to the geometric and topological properties of the manifold from which the data is sampled. The natural approximate "sparsity" of the spectrum of these kinds of matrices is discussed in Section D in the Supplementary Material [19].

In practice, the data analyst has to work with $L(\tilde{W}, \tilde{G})$ or variants taking into account only the k nearest neighbors of each datapoint. Hence, it could potentially be the case that $L(\tilde{W}, \tilde{G})$ or its variants do not share many of the good properties of $L(W, G)$. It is therefore natural to study the properties of the standard GCL algorithm applied to noisy data.

We mention that GCL algorithms may apply beyond the case of the rotational invariance distance and $O(k)$, and we explain in Section 3.2.3 how our results apply in this more general context.

3.2.2. *Modified GCL algorithm and rotationally invariant distance.* We now show that our modification to the standard algorithm is robust to noise. More precisely, we show that the modified GCL matrix $L_0(\tilde{W}, \tilde{G})$ is spectrally close to the GCL matrix computed from the noise-free data, $L(W, G)$. We also argue below that it is important to use the full matrix $L_0(\tilde{W}, \tilde{G})$ and not incorporate nearest-neighbor information.

THEOREM 3.1. *Consider the modified GCL matrix $L_0(\tilde{W}, \tilde{G})$ computed from the noisy data and the GCL matrix $L(W, G)$ computed from the noise-free data. Under Assumptions* G1 *and* A0–A1, *we have*, *if* $\operatorname{trace}(\Sigma_i) = \operatorname{trace}(\Sigma_j) = \operatorname{trace}(\Sigma)$ *for all* (i, j),

$$|||L_0(\tilde{W}, \tilde{G}) - L(W, G)|||_2 = \mathrm{o}_P(1),$$

provided there exists $\gamma > 0$, *independent of* n *and* p *such that*

$$\inf_i \sum_{j \neq i} \frac{\exp(-d^2_{ij,\mathrm{clean}}/\varepsilon)}{n} \geq \gamma > 0.$$

The proof is given in Section A-4 in the Supplementary Material [19].

Note that the previous result means that $L_0(\tilde{W}, \tilde{G})$ and $L(W, G)$ are essentially spectrally equivalent: indeed we can use the Davis–Kahan theorem or Weyl's inequality to relate eigenvectors and eigenvalues of $L_0(\tilde{W}, \tilde{G})$ to those of $L(W, G)$; see [5, 35] or [14] for a brief discussion putting all the needed results together; note that L_0 and L are similar to Hermitian matrices. In particular, if the large eigenvalues of $L(W, G)$ are separated from the rest of the spectrum, the eigenvalues of $L_0(\tilde{W}, \tilde{G})$ and corresponding eigenspaces will be close to those of $L(W, G)$.

Is the diagonal modification of the algorithm really needed? It is natural to ask what would have happened if we had not used a diagonal modification to the standard algorithm, that is, if we had worked with $L(\tilde{W}, \tilde{G})$ instead of $L_0(\tilde{W}, \tilde{G})$. It is easy to see that

$$L(\tilde{W}, \tilde{G}) = L_0(\tilde{W}, \tilde{G}) + \mathsf{D},$$

where D is a block diagonal matrix with

$$\mathsf{D}(i, i) = \frac{\tilde{w}_{i,i}}{\sum_{j \neq i} \tilde{w}_{i,j}} \mathrm{Id}_k = \frac{1}{\sum_{j \neq i} \tilde{w}_{i,j}} \mathrm{Id}_k.$$

Under our assumptions,

$$\left|\left|\left| n \exp(-2 \operatorname{trace}(\Sigma)/\varepsilon)\mathsf{D} - D\left(\left\{\left[\frac{\sum_{j \neq i} \exp(-d^2_{ij,\mathrm{clean}}/\varepsilon)}{n}\right]^{-1}\right\}_{i=1}^n\right)\right|\right|\right|_2 = \mathrm{o}_P(1).$$

We also recall that under Assumption G1, $\mathrm{trace}(\Sigma)$ can be as large as $p^{1/2-\eta}$, a very large number in our asymptotics. So in particular, if n is polynomial in p, we have then $n^{-1}\exp(2\,\mathrm{trace}(\Sigma)/\varepsilon) \to \infty$. This implies that

$$L(\tilde{W}, \tilde{G}) = L_0(\tilde{W}, \tilde{G}) + \mathsf{D}$$

is then dominated in spectral terms by D in general. So it is clear that in the high-noise regime, if we had used the standard GCL algorithm, the spectrum of $L(\tilde{W}, \tilde{G})$ could have mirrored that of D—which has little to do in general with the spectrum of $L(W, G)$, which we are trying to estimate—and the noise would have rendered the algorithm ineffective. An exception is the case where D is spectrally close to the identity, in which case the eigenvectors of $L(\tilde{W}, \tilde{G})$ would be close to those of $L_0(\tilde{W}, \tilde{G})$. Even in this case, it is, however, not harmful to use $L_0(\tilde{W}, \tilde{G})$ instead of $L(\tilde{W}, \tilde{G})$. There is therefore no downside to using $L_0(\tilde{W}, \tilde{G})$ instead of $L(\tilde{W}, \tilde{G})$ by default, and potentially there is some upside.

By using the modification we propose, we guarantee that even in the high-noise regime, the spectral properties of $L_0(\tilde{W}, \tilde{G})$ mirror those of $L(W, G)$. We have hence made the GCL algorithm more robust to noise.

On a technical note, the fact that, in the noisy case, our approximation results for the RID distance hold only (up to a scalar) off the diagonal forces us to work with Lemma 3.2 and not Lemma 3.1. If it were the case, for instance, for different affinity functions, that the approximation results held on the diagonal too, we could use the results of Lemma 3.1, and we would not have to do the diagonal modification.

On the use of nearest-neighbor graphs. In practice, variants of the GCL algorithms we have described use nearest-neighbor information to replace $w_{i,j}$ by 0 if $w_{i,j}$ is not among the k largest elements of $\{w_{i,j}\}_{j=1}^n$. In the high-noise setting, the nearest-neighbor information is typically not robust to noise, which is why we propose to use all the $w_{i,j}$'s and avoid the nearest-neighbor variant of the GCL algorithm, even though the latter probably makes more intuitive sense in the noise-free context. A systematic study of the difference between these two variants is postponed to future work. In Section 4, we carry out some numerical experiments to illustrate the lack of robustness to noise of the nearest-neighbor information and the improvements that result from using $L_0(\tilde{W}, \tilde{G})$.

Comparison with previous results in the literature. As far as we know, the study of the impact of high-dimensional additive noise on kernel methods was started in [16]. Compared to this paper, our extension is two-fold: (1) the noise level [i.e., $\mathrm{trace}(\Sigma)$] that is studied in the current paper is much higher than what was studied in [16]. This is partly a result of the fact that the current paper focuses on the Gaussian kernel whereas [16] studied many more kernels. (2) El Karoui [16] focused on standard kernel methods based on the graph Laplacian, that is, $k = 1$, and the connection information is not included in the data analysis. Incorporating this new element creates new difficulties. See also [30] for another study of the influence of noise in a different setup.

3.2.3. *GCL beyond the rotational invariance distance.* The previous analysis has been carried out for the RID and corresponding rotations where we studied the impact of additive noise in Section 2. However, it is clear that our results apply much more broadly. We have the following theorem.

THEOREM 3.2. *Suppose we are given a collection* $\mathsf{d}_{i,j,\text{noisy}}$ *of* (*scalar-valued*, *symmetric in* i, j) *dissimilarities between noisy versions of objects* i *and* j, $1 \leq i, j \leq n$. *Suppose objects* i *and* j *have* (*scalar-valued*, *symmetric in* i, j) *dissimilarity* $\mathsf{d}_{i,j,\text{clean}}$. *Consider the asymptotic regime where* $n \to \infty$, *and suppose that there exists* $\xi_n \in \mathbb{R}$ *such that*

$$\sup_{i \neq j} |\mathsf{d}^2_{i,j,\text{noisy}} - \mathsf{d}^2_{i,j,\text{clean}} - \xi_n| = \mathrm{o}_P(1).$$

Call $\tilde{w}_{i,j} = \exp(-\mathsf{d}^2_{i,j,\text{noisy}}/\nu)$ *and* $w_{i,j} = \exp(-\mathsf{d}^2_{i,j,\text{clean}}/\nu)$ *the corresponding affinities.* ν *is held fixed in our asymptotics*, *though the way affinities are computed may change with* n.

Suppose $\tilde{G}_{i,j}$ *is the connection between noisy versions of objects* i *and* j, *and* $G_{i,j}$ *is the connection between the clean/noise-free version of objects* i *and* j. *Suppose that* $w_{i,j}$, $G_{i,j}$ *and* $\tilde{G}_{i,j}$ *satisfy the assumptions of Lemma* 3.2, *with* ε *and* η *possibly random but* $\mathrm{o}_P(1)$ *and* γ *bounded below as* $n \to \infty$. *Then*

$$|||L(W, G) - L_0(\tilde{W}, \tilde{G})|||_2 = \mathrm{o}_P(1).$$

PROOF. This theorem is just a consequence of Lemma 3.2. Indeed, the affinities are all bounded by 1. Furthermore, we can use $f_i = \exp(-\xi_n/\nu)$, and all the approximation results needed in Lemma 3.2 are true, so the result follows. □

3.2.4. *A situation without robustness to noise.* So far, our work has been quite general and has shown that when the noise is Gaussian (or Gaussian-like) and its covariance Σ_i is such that $\text{trace}(\Sigma_i) = \text{trace}(\Sigma_j)$ for all i, j, GCL algorithms can be made robust to noise.

It has been recognized [13, 15–17] that to study the robustness of various statistical procedures in high-dimension, it is essential to move beyond the Gaussian-like situation and study, for instance, elliptical/scale mixture of Gaussian models. This is largely due to the peculiar geometry of high-dimensional Gaussian and Gaussian-like vectors; see the above references and [23].

If we now write down a model for the noise where $N_i = \lambda_i Z_i$, where Z_i are i.i.d. $\mathcal{N}(0, \Sigma)$, λ_i's are i.i.d. with $\mathbf{E}\lambda_i^2 = 1$ and $\lambda_i \in \mathbb{R}$ is independent of Z_i, it is easy to modify our analysis (assuming, e.g., that λ_i^2 are bounded, though this condition could easily be relaxed) and to realize that our main approximation result in Proposition 2.1 is replaced by

$$\sup_{i \neq j} |d^2_{ij,\text{noisy}} - d^2_{ij,\text{clean}} - [\lambda_i^2 + \lambda_j^2] \,\text{trace}(\Sigma)| = \mathrm{o}_P(1).$$

In this situation, Theorem 2.2 is still valid. However, Theorem 3.1 is not valid anymore. The matrix $L_0(\tilde{W}, \tilde{G})$ can be approximated by a matrix that depends both on the signal and the distribution of the λ_i^2's, and there is no guarantee in general that this matrix will have approximately the same spectral properties as $L(W, G)$ or $L_0(W, G)$, the GCL matrix generated from the noise-free signals. This suggests that even our modification of the original GCL algorithm will not be robust to this "elliptical"-noise contamination.

4. Numerical work. Although the robustness properties of GCL methods were not well studied in the past, these methods have been successfully applied to different problems, for example, [1, 12, 27, 31, 34, 40]. In this section, we show simulated examples to illustrate the practical performance of our theoretical findings about GCL methods. We refer interested readers to the aforementioned papers for details and results of its applications.

To demonstrate the main finding of this paper—that GCL methods are robust to high-levels of noise in the spectral sense—we take the noise to be a random Gaussian vector $Z \sim \mathcal{N}(0, cI_p/p^\alpha)$, where $\alpha \leq 1$ and $c > 0$. Note that the amount of noise, or the trace of the covariance matrix of Z, is $cp^{1-\alpha}$ and will tend to infinity when $p \to \infty$ and $\alpha < 1$.

4.1. *1-dimensional manifold.* Our first example is a dataset sampled from a low-dimensional manifold, which is embedded in a high-dimensional space. This dataset can be viewed as a collection of high-dimensional points which is (locally) parametrized by only few parameters,[3] but in a nonlinear way.

As a concrete example, we take the twisted bell-shaped simple and closed curve, denoted as M, embedded in the first 3 axes of $\mathbb{R}^p$, where $p \gg 2$, via $\iota: [0, 2\pi) \to \mathbb{R}^p$,

$$\begin{aligned}\iota: t \mapsto \big[&\cos(t), (1 - 0.8e^{-8\cos^2 t})\cos(\pi(\cos(t) + 1)/4), \\ &(1 - 0.8e^{-8\cos^2 t})\sin(\pi(\cos(t) + 1)/4), 0, \dots, 0\big] \in \mathbb{R}^p.\end{aligned}$$

M is a 1-dimensional smooth manifold without boundary; that is, no matter how big p is, locally the points on M can be parametrized by only 1 parameter. See Figure 1(A) for an illustration. One interesting such dataset is the 2-D tomography from noisy projections taken at unknown random directions [32].

For our numerical work, we independently sample n points uniformly at random from $[0, 2\pi)$. Due to the nonlinear nature of ι, it is equivalent to nonuniformly sampling n points from M independently. Denote the clean data as $\mathcal{Y} = \{y_i\}_{i=1}^n \subset$ M. The data $\mathcal{X} = \{x_i\}_{i=1}^n$ we analyze is the clean data contaminated by additive noise,

[3] By definition, although locally the manifold resembles Euclidean space near a point, globally it might not. Thus, in general, we can only parametrize the manifold locally. This feature captures the possible nonlinear structure in the data.

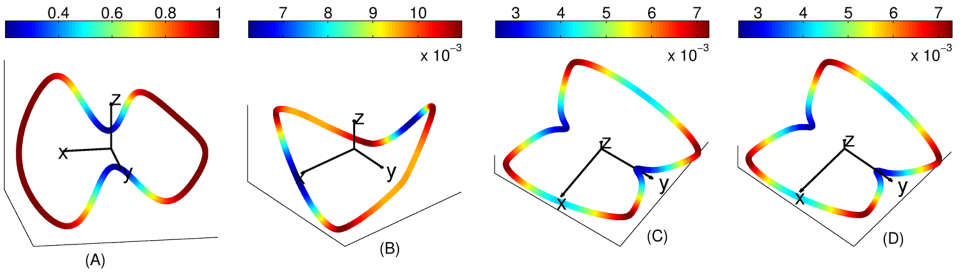

FIG. 1. *Clean samples from the twisted bell-shaped manifold.* (A): *The clean samples. Here we only plot the first* 3 *axes of the high-dimensional data* $\mathcal{Y}$. *The color of each point is a surrogate of the norm of each embedded point—blue means a relative small norm and dark red means a relative large norm*; *the scale above the figure refers to* $\{\|y_i\|_2\}_{i=1}^n$, *that is, the norm of the data vectors in* $\mathbb{R}^p$. (B): *The results of the truncated diffusion maps* (*tDM*), $\Phi_{1,1000,3}$, *when the connection graph is* G^{NN}, *and the diagonal entries are not removed*, *where the number of nearest neighbors is chosen to be* 100; (C): *The result of tDM*, $\Phi_{1,1000,3}$, *when the connection graph is* G *and the diagonal entries are not removed*; (D): *The result of tDM*, $\Phi_{1,1000,3}$, *when the connection graph is* G *and the diagonal entries are removed. Note that without surprise*, *the "parametrization" of the bell-shaped manifold is recovered in* (B), (C) *and* (D). *For* (B), (C), *and* (D) *in Figure* 2, *the scales above the figures refer to the norm of* $\{\Phi_{1,1000,3}(y_i)\}_{i=1}^n$; *these vectors are of course* 3-*dimensional*, *which explains the difference in magnitude of our scales. Indeed*, *in order to recover the local geodesic distance between two close points*, *we need more eigenvectors than* 3. *However*, *for the visualization purpose*, *we have only taken the first* 3 *nontrivial eigenvectors into account here.*

that is, $x_i = y_i + Z_i$, with Z_i i.i.d. with the same distribution as Z. We measure the *signal-to-noise ratio* of the dataset by the quantity $\mathrm{snrdb} := 20 \log \frac{\sqrt{\mathbb{E}X^TX}}{\sqrt{\mathbb{E}Z^TZ}}$. We take $n = p = 1000$ and $\alpha = 1/4$. Note that $\alpha = 1/4$ is the critical value in our analysis beyond which our results do not apply. For concreteness, the snrdb will be -9.25 and -18.73, respectively, when $c = 0.25, 0.4$.

Then, we build up the GCL [in this 1-dimensional manifold with the trivial connection, it is equivalent to the graph Laplacian (GL)] from $\mathcal{X}$ by setting $L(\tilde{W}, \tilde{G})$ [see (1)], where the $n \times n$ affinity matrix $\tilde{W}$ is defined as $\tilde{W}_{i,j} := e^{-\|x_i - x_j\|^2_{\mathbb{R}^p}/m}$, the bandwidth m is the first quartile of all Euclidean distances between pairs of $(x_i, x_j) \in \mathrm{E}$ and the $n \times n$ connection matrix $\tilde{G}$ is defined as $\tilde{G}_{i,j} := 1$ for all $(x_i, x_j) \in \mathrm{E}$. That choice of m is a common in practice.

Note that in practice, it is also common to use a nearest-neighbor (NN) scheme to build up the GCL for the sake of computational efficiency; see the Supplementary Material [19], Section B for details. The associated affinity matrix (resp., connection matrix and GCL) is denoted as $\tilde{W}^{\mathrm{NN}}$ [resp., $\tilde{G}^{\mathrm{NN}}$ and $L(\tilde{W}^{\mathrm{NN}}, \tilde{G}^{\mathrm{NN}})$], where we choose 100 nearest neighbors to construct edges. We have seen in the analysis described earlier in the paper that when $\alpha < 1$, it theoretically helps to remove the diagonal terms of the GCL matrix in order to preserve spectral properties, so we also consider the matrix $L_0(\tilde{W}, \tilde{G})$ for the comparison.

We then evaluate the eigenvalues and eigenvectors of the above three different GCL's. To simplify the notation, we use the same notation to denote the eigenvectors $u_1, u_2, u_3 \ldots \in \mathbb{R}^n$ associated with the eigenvalues $1 = \lambda_1 > \lambda_2 \geq \lambda_3 \geq \cdots \geq 0$. We now show two sets of results to demonstrate the robustness of the GCL methods studied in this paper.

Dimension reduction and data visualization. To achieve this, we may embed the sampled points into $\mathbb{R}^m$ by the *truncated diffusion maps* (tDM) with time $t > 0$ and $m \geq 1$,

$$\Phi_{t,n,m} : x_i \mapsto \big(\lambda_2^t u_2(i), \lambda_3^t u_3(i), \ldots, \lambda_{m+1}^t u_{m+1}(i)\big) \in \mathbb{R}^m,$$

where m is chosen by the user depending on the problem; that is, we map the ith data point to $\mathbb{R}^m$ using the first m nontrivial eigenvectors of the GCL. For the purpose of visualization, we may take $m = 2$ or $m = 3$. For other purposes, we may choose m depending on a given threshold $\delta > 0$. m is chosen so that $|\lambda_{m+1}/\lambda_2|^t > \delta$ and $|\lambda_{m+2}/\lambda_2|^t \leq \delta$. In this example, we choose $t = 1$ and $m = 3$ for the visualization; see also Section B-1 in the Supplementary Material [19] for more details about this simulation. The embedding results of $\mathcal{Y}$, $\Phi_{1,1000,3}$, based on the above different GCL's are shown in Figure 1, and the results from $\mathcal{X}$ with $c = 0.4$ are shown in Figure 2. Ideally, we would expect to recover the "parametrization" of the dataset with the idea that the eigenvectors of the GCL represent a set of new coordinates for the data points, so the high-dimensional dataset can be visualized in this new set of coordinates, or its dimension can be

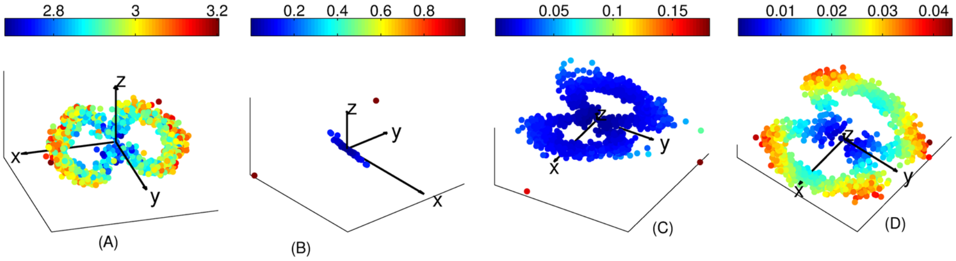

FIG. 2. *Noisy samples from the twisted bell-shaped manifold with* $\alpha = 1/4$ *and* $c = 0.4$. (A): *The noisy samples. We only plot the first* 3 *axes of the data* $\mathcal{X}$; *hence we see only a small fraction of the noise, as* 997 *out of* 1000 *coordinates are not plotted. The color of each point is a surrogate for* $\|x_i\|$, $x_i \in \mathbb{R}^{1000}$. (B): *The results of the truncated diffusion maps* (*tDM*), $\Phi_{1,1000,3}$, *when the connection graph is constructed by the NN scheme and the diagonal entries are not removed, where the number of nearest neighbors is chosen to be* 100. *We can barely see the circular structure in the middle, and there are several big outliers*; (C): *The result of tDM*, $\Phi_{1,1000,3}$, *when the connection graph is complete and the diagonal entries are not removed. Note that when compared with* (B), *the embedding is better in the sense that the "parametrization," the simple and close curve, is better recovered. But we can still observe several outliers*; (D): *The result of tDM*, $\Phi_{1,1000,3}$, *when the connection graph is complete and the diagonal entries are removed. Note that compared with* (C), *the embedding is yet better in the sense that the number of outliers is reduced and the parametrization of the manifold is recovered. Note that for* (B), (C), (D), *the scale above the figures refer to* $\{\|\Phi_{1,1000,3}\|(x_i)\}_{i=1}^n$, *which are* 3-*dimensional vectors. The different scales indicate the presence of outliers.*

reduced. In this specific example, we would expect to find a simple and closed curve out of the noisy dataset which represents the dataset in $\mathbb{R}^3$. Clearly when the dataset is clean, we succeed in the task no matter which GCL we use. However, if the dataset is noisy, at high-noise levels, the embedding might not be that meaningful if we use $L(\tilde{W}^{\mathrm{NN}}, \tilde{G}^{\mathrm{NN}})$ or $L(\tilde{W}, \tilde{G})$. Indeed, as shown in Figure 2, with $L(\tilde{W}^{\mathrm{NN}}, \tilde{G}^{\mathrm{NN}})$ the structure of the dataset is barely recovered; with $L(\tilde{W}, \tilde{G})$, even though we can get the simple closed curve[4] back, there are several outliers which might deteriorate the interpretation. In this noisy case, we can only succeed in the task if we choose $L_0(\tilde{W}, \tilde{G})$, as is discussed in this paper.

Identifying neighbors. Identifying neighbors (in various metrics) of a given data point from a noisy dataset is not only important but also challenging in practice (e.g., it is essential in the class averaging algorithm for the cryo-EM problem to find the correct neighbors when the projection images are noisy). This problem is directly related to local geodesic distance estimation when the dataset is modeled by a manifold. The theoretical properties of diffusion maps and vector diffusion maps make these methods particularly well suited for these tasks [31]. To determine neighbors, we of course need a notion of distance. In addition to the naive L^2 distance between points, we consider the (*truncated*) *diffusion distance* between two points $x_i, x_j \in \mathcal{X}$ by

$$d_{\mathrm{DM},t,n,m}(x_i, x_j) := \|\Phi_{t,n,m}(x_i) - \Phi_{t,n,m}(x_j)\|_{\mathbb{R}^m},$$

where m is determined in the tDM, $\Phi_{t,n,m}$, with m chosen by a given thresholding $\delta > 0$. Then we determine the nearest neighbors of each data point based on these distances, where we choose $t = 1$ and $\delta = 0.2$ for the diffusion distance. More precisely, we first determine 10 nearest neighbors of x_i, denoted as x_{i_j}, $j = 1, \dots, 10$, from the noisy dataset $\mathcal{X}$, for all i. Then, since we know the ground truth, we may check the true relationship between y_i and y_{i_j}, $j = 1, \dots, 10$, that is, $d_{\mathrm{DM},t,n,m}(y_i, y_{i_j})$ for various GCL methods, or $\|y_i - y_{i_j}\|$ if we use L^2 distance. Clearly, if the method preserves nearest-neighbor information, at least approximately, the ranks of the y_{i_j}'s measured in terms of distances to y_i should be small. To quantify the estimation accuracy, we collect the ranks of all estimated nearest neighbors, and plot the cumulative distribution results in Figure 3. In other words, if we call R_{i_j} the rank of y_{i_j} in terms of distance to y_i, we plot the c.d.f. of $\{\{\mathsf{R}_{i_j}\}_{j=1}^{10}\}_{i=1}^{n}$ for the various distances we use. (There are many other methods one could use to do these comparisons, such as using Kendall's τ and variants thereof; see [21]. The one we use here has the benefit of simplicity.) When the dataset is clean, all methods perform in the same way, as is predicted in Theorem D.7 of [19].

[4] The main idea behind tDM is embedding the dataset to a lower dimensional Euclidean space so that the structure underlying the data can be extracted. Please see Section D-5 in the Supplementary Material [19] for details.

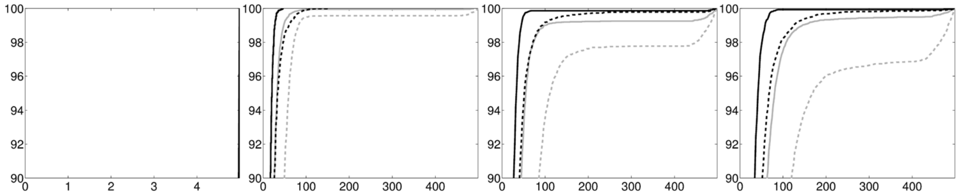

FIG. 3. *The result of nearest-neighbors estimation. In all subfigures, the x-axis is the true rank of an estimated nearest neighbor, and the y-axis is its cumulative distribution. To emphasize the difference, we only show the area ranging from* 90% *to* 100% *in the y-axis. The gray dashed* (*gray, black dashed and black, resp.*) *curve is the cumulative distribution of the true ranks of the estimated nearest neighbors estimated from the ordinary Euclidean distance* [*diffusion distance based on* $L(\tilde{W}^{\mathrm{NN}}, \tilde{G}^{\mathrm{NN}})$, $L(\tilde{W}, \tilde{G})$ *and* $L_0(\tilde{W}, \tilde{G})$, *resp.*]. *From left to right*: *clean samples from the bell shaped manifold, noisy samples with* $\alpha = 1/4$ *and* $c = 0.25, 0.4, 0.5$, *respectively. It is clear that when the noise is large, the result based on the* L^2 *distance is much worse than the others. The result based on* $L(\tilde{W}^{\mathrm{NN}}, \tilde{G}^{\mathrm{NN}})$ *is slightly better, but not that good,* $L(\tilde{W}, \tilde{G})$ *is even better and* $L_0(\tilde{W}, \tilde{G})$ *is the best.*

It is clear from the results that when the noise is large, the result based on the L^2 distance is much worse than the others. The performance based on the diffusion distance from $L(\tilde{W}^{\mathrm{NN}}, \tilde{G}^{\mathrm{NN}})$ is better when the noise level is not big, but still a nonnegligible portion of error exists; the results based on $L(\tilde{W}, \tilde{G})$ and $L_0(\tilde{W}, \tilde{G})$ are much better, while the result based on $L_0(\tilde{W}, \tilde{G})$ is the best.

4.2. *2-dimensional images.* In Section 4.1, we investigated numerically the influence of noise on GCL methods when the connection function is trivial. In this subsection, we discuss an example where the connection function plays an essential role in the analysis. We consider a dataset which contains randomly rotated versions of a set of objects, and the task is to align these objects in addition to classifying them. We encounter these kinds of datasets and problems in, for example, image processing [31, 34, 40], shape analysis [24], phase retrieval problems [1, 27], etc. In [1, 27, 31, 34, 40] and others, the GCL methods have been applied to solve the problem.

To focus specifically on demonstrating the influence of noise on this problem, we work with 2-dimensional images observed in polar coordinates. To make matters simple, our images are defined as functions observed on the unit circle at equally spaced points. We have n_K different clean images. We then randomly and independently rotate these images to create our dataset. We use n_R random rotations for each image. In the end we get $n = n_K n_R$ randomly rotated images $\{S_i\}_{i=1}^n \subset \mathbb{R}^p$. To each image corresponds a rotation $R_i \in SO(2)$, or equivalently an angle. The data $\mathcal{X} = \{I_i\}_{i=1}^n$ we analyze is the clean data contaminated by independent noise, which is i.i.d. sampled from Z; that is, we have $I_i = S_i + Z_i$. We give more precise mathematical and simulation details in Section B-2 in the Supplementary Material [19].

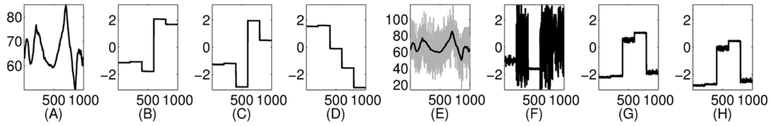

FIG. 4. (A): *One of $n_K = 5$ clean surrogate images*; (B)–(D): *Alignment vectors z computed from clean images*; (E): *A noisy surrogate image.* (F)–(H): *Alignment vectors z computed from noisy images with $c = 6\sigma$ and $\alpha = 1/4$.* (A) *and* (E): *The black curve is a clean surrogate image, and the gray curve is its noisy version*; (B) *and* (F): *The result from the GCL built up from the NN scheme*; (C) *and* (G): *The result from the GCL built from complete connection graph and the diagonal entries are not removed*; (D) *and* (H): *The result from the GCL built from the complete connection graph with the diagonal entries removed. It is clear that when the images are clean, all different GCL's give equivalent results* (*see text for explanation of visual differences which are of no statistical importance*). *But in the presence of noise, the GCL built up from the NN scheme is obviously worse.*

We now build up the GCL $L(\tilde{W}, \tilde{G})$ by setting $\tilde{W}_{i,j} := e^{-d^2_{\mathrm{RID}}(I_i, I_j)/m}$, where m is the first quartile of all nonzero RID distances, and the connection function as $\tilde{G}_{i,j} := \operatorname{argmin}_{R \in \mathcal{T}^{(2)}} \|I_i - R \circ I_j\|$. For comparison purposes, we also take the NN scheme to construct the connection graph, denoted by $L(\tilde{W}^{\mathrm{NN}}, \tilde{G}^{\mathrm{NN}})$, where we choose 100 nearest neighbors—as defined by the RID distance—to construct edges. Thanks to the connection function, we can estimate the rotations R_i applied to the ith image up to a common rotation from the top eigenvector $v_1 \in \mathbb{C}^n$ of the GCL's.

To evaluate the performance of the estimated rotation, we construct a complex vector $u \in \mathbb{C}^n$ whose ith entry is the complex form of the rotation R_i. We then evaluate the difference between the estimated rotation of the ith object and the ground truth by observing the angle of $u(i)^* v(i)$. This quantity shows the discrepancy between the true rotation and the estimated one. For visualization, we plot the vector $z \in \mathbb{R}^n$ where $z(i)$ is the angle of the complex number $u(i)^* v(i)$, which measures estimation error. In Figure 4, the resulting z's with $p = 1000$, $n_K = 5$, $n_R = 200$, $\alpha = 1/4$ and $c = 6\sigma$ are plotted; see the Supplementary Material [19], Section B for the value of σ. Since there are 5 different images, we see a piecewise function with 5 different values when the images are clean/noise-free, indicating that we correctly estimate the rotations R_i's as well as the class membership of the images. The visual dissimilarities between the functions in Figures 4(B), (C) and (D) is due to the fact that all estimation tasks here can be performed only up to a rotation for each of the $n_K = 5$ template images. Further discussion of this example can be found in the Supplementary Material [19], Section B-2.

5. Conclusion. In this paper, we have studied the statistical properties of a recent generalization of kernel methods called GCL methods and in particular their sensitivity to additive noise. We have shown both theoretically and numerically that they can be made tolerant to very high levels of noise. Based on our analysis, we have proposed two modifications of the standard approach that improve performance in the setup we consider. First, practitioners will benefit from not trying to incorporate nearest-neighbor information derived from the affinity function as those tend to be very sensitive to noise. Second, setting the diagonal elements of the affinity matrix to zero increases robustness to noise.

Acknowledgment. The authors thank an anonymous referee for interesting and constructive comments that helped improve the paper.

SUPPLEMENTARY MATERIAL

Supplement to "Graph connection Laplacian methods can be made robust to noise" (DOI: 10.1214/14-AOS1275SUPP; .pdf). We provide detailed proofs and supplementary information in the Supplementary Material.

REFERENCES

[1] ALEXEEV, B., BANDEIRA, A. S., FICKUS, M. and MIXON, D. G. (2014). Phase retrieval with polarization. *SIAM J. Imaging Sci.* **7** 35–66.

[2] BANDEIRA, A. S., SINGER, A. and SPIELMAN, D. A. (2013). A Cheeger inequality for the graph connection Laplacian. *SIAM J. Matrix Anal. Appl.* **34** 1611–1630. MR3138103

[3] BATARD, T. and SOCHEN, N. (2012). Polyakov action on (ρ, G)-equivariant functions application to color image regularization. In *Scale Space and Variational Methods in Computer Vision* 483–494. Springer, Berlin.

[4] BATARD, T. and SOCHEN, N. (2014). A class of generalized Laplacians on vector bundles devoted to multi-channel image processing. *J. Math. Imaging Vision* **48** 517–543. MR3171428

[5] BHATIA, R. (1997). *Matrix Analysis. Graduate Texts in Mathematics* **169**. Springer, New York. MR1477662

[6] BOYER, D. M., LIPMAN, Y., CLAIR, E. S., PUENTE, J., PATEL, B. A., FUNKHOUSER, T., JERNVALL, J. and DAUBECHIES, I. (2011). Algorithms to automatically quantify the geometric similarity of anatomical surfaces. *Proc. Natl. Acad. Sci. USA* **108** 18221–18226.

[7] CHEN, P., LIN, C.-L. and CHERN, I.-L. (2013). A perfect match condition for point-set matching problems using the optimal mass transport approach. *SIAM J. Imaging Sci.* **6** 730–764. MR3038024

[8] CHUNG, F. and KEMPTON, M. (2013). A local clustering algorithm for connection graphs. In *Algorithms and Models for the Web Graph* (A. Bonato, M. Mitzenmacher and P. Pralat, eds.). *Lecture Notes in Computer Science* **8305** 26–43. Springer, Berlin.

[9] CHUNG, F., ZHAO, W. and KEMPTON, M. (2012). Ranking and sparsifying a connection graph. In *Algorithms and Models for the Web Graph* (A. Bonato and J. Janssen, eds.). *Lecture Notes in Computer Science* **7323** 66–77. Springer, Berlin.

[10] COLLINS, A., ZOMORODIAN, A., CARLSSON, G. and GUIBAS, L. J. (2004). A barcode shape descriptor for curve point cloud data. *Computers & Graphics* **28** 881–894.

[11] CUCURINGU, M., LIPMAN, Y. and SINGER, A. (2012). Sensor network localization by eigenvector synchronization over the Euclidean group. *ACM Trans. Sens. Netw.* **8** 19:1–19:42.

[12] CUCURINGU, M., SINGER, A. and COWBURN, D. (2012). Eigenvector synchronization, graph rigidity and the molecule problem. *Inf. Inference* **1** 21–67. MR3311440

[13] DIACONIS, P. and FREEDMAN, D. (1984). Asymptotics of graphical projection pursuit. *Ann. Statist.* **12** 793–815. MR0751274

[14] EL KAROUI, N. (2008). Operator norm consistent estimation of large-dimensional sparse covariance matrices. *Ann. Statist.* **36** 2717–2756. MR2485011

[15] EL KAROUI, N. (2009). Concentration of measure and spectra of random matrices: Applications to correlation matrices, elliptical distributions and beyond. *Ann. Appl. Probab.* **19** 2362–2405. MR2588248

[16] EL KAROUI, N. (2010). On information plus noise kernel random matrices. *Ann. Statist.* **38** 3191–3216. MR2722468

[17] EL KAROUI, N., BEAN, D., BICKEL, P. J., LIM, C. and YU, B. (2013). On robust regression with high-dimensional predictors. *Proc. Natl. Acad. Sci. USA* **110** 14557–14562.

[18] EL KAROUI, N. and WU, H.-T. (2013). Vector diffusion maps and random matrices with random blocks. Available at arXiv:1310.0188.

[19] EL KAROUI, N. and WU, H. (2015). Supplement to "Graph connection Laplacian methods can be made robust to noise." DOI:10.1214/14-AOS1275SUPP.

[20] EPSTEIN, C. L. (2008). *Introduction to the Mathematics of Medical Imaging*, 2nd ed. SIAM, Philadelphia, PA. MR2378706

[21] FAGIN, R., KUMAR, R. and SIVAKUMAR, D. (2003). Comparing top *k* lists. *SIAM J. Discrete Math.* **17** 134–160 (electronic). MR2033311

[22] HADANI, R. and SINGER, A. (2011). Representation theoretic patterns in three dimensional cryo-electron microscopy I: The intrinsic reconstitution algorithm. *Ann. of Math.* (2) **174** 1219–1241. MR2831117

[23] HALL, P., MARRON, J. S. and NEEMAN, A. (2005). Geometric representation of high dimension, low sample size data. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **67** 427–444. MR2155347

[24] HUANG, Q.-X., SU, H. and GUIBAS, L. (2013). Fine-grained semi-supervised labeling of large shape collections. *ACM Transactions on Graphics* (*TOG*) **32** 190.

[25] LEDOUX, M. (2001). *The Concentration of Measure Phenomenon. Mathematical Surveys and Monographs* **89**. Amer. Math. Soc., Providence, RI. MR1849347

[26] LIPMAN, Y., AL-AIFARI, R. and DAUBECHIES, I. (2013). Continuous Procrustes distance between two surfaces. *Comm. Pure Appl. Math.* **66** 934–964. MR3043386

[27] MARCHESINI, S., TU, Y.-C. and WU, H.-T. (2015). Alternating projection, ptychographic imaging and phase synchronization. *Appl. Comput. Harmon. Anal.* DOI:10.1016/j.acha.2015.06.005. Available at arXiv:1402.0550.

[28] MÉMOLI, F. and SAPIRO, G. (2005). A theoretical and computational framework for isometry invariant recognition of point cloud data. *Found. Comput. Math.* **5** 313–347. MR2168679

[29] NIYOGI, P., SMALE, S. and WEINBERGER, S. (2009). Finding the homology of submanifolds with high confidence from random samples. In *Twentieth Anniversary Volume* 1–23. Springer, New York.

[30] SINGER, A. (2011). Angular synchronization by eigenvectors and semidefinite programming. *Appl. Comput. Harmon. Anal.* **30** 20–36. MR2737931

[31] SINGER, A. and WU, H.-T. (2012). Vector diffusion maps and the connection Laplacian. *Comm. Pure Appl. Math.* **65** 1067–1144. MR2928092

[32] SINGER, A. and WU, H.-T. (2013). Two-dimensional tomography from noisy projections taken at unknown random directions. *SIAM J. Imaging Sci.* **6** 136–175. MR3032950

[33] SINGER, A. and WU, H.-T. (2013). Spectral convergence of the connection laplacian from random samples. Submitted.

[34] SINGER, A., ZHAO, Z., SHKOLNISKY, Y. and HADANI, R. (2011). Viewing angle classification of cryo-electron microscopy images using eigenvectors. *SIAM J. Imaging Sci.* **4** 723–759. MR2831077

[35] STEWART, G. W. and SUN, J. G. (1990). *Matrix Perturbation Theory*. Academic Press, Boston, MA. MR1061154

[36] SUN, J., OVSJANIKOV, M. and GUIBAS, L. (2009). A concise and provably informative multiscale signature based on heat diffusion. In *Proceedings of the Symposium on Geometry Processing*, *SGP* '09 1383–1392. Eurographics Association.

[37] TALMON, R., COHEN, I., GANNOT, S. and COIFMAN, R. (2013). Diffusion maps for signal processing: A deeper look at manifold-learning techniques based on kernels and graphs. *IEEE Signal Process*. *Mag*. **30** 75–86.

[38] VAN DER VAART, A. W. (1998). *Asymptotic Statistics*. *Cambridge Series in Statistical and Probabilistic Mathematics* **3**. Cambridge Univ. Press, Cambridge. MR1652247

[39] WANG, F., HUANG, Q. and GUIBAS, L. J. (2013). Image co-segmentation via consistent functional maps. In *The IEEE International Conference on Computer Vision* (*ICCV*).

[40] ZHAO, Z. and SINGER, A. (2013). Rotationally invariant image representation for viewing direction classification in cryo-EM. Available at arXiv:1309.7643.

DEPARTMENT OF STATISTICS
UNIVERSITY OF CALIFORNIA, BERKELEY
BERKELEY, CALIFORNIA 94720
USA
E-MAIL: nkaroui@berkeley.edu

DEPARTMENT OF MATHEMATICS
UNIVERSITY OF TORONTO
TORONTO, ONTARIO M5S 2E4
CANADA
E-MAIL: hauwu@math.toronto.edu