SIAM J. MATH. DATA SCI.
Vol. 5, No. 3, pp. 589–614

Robust Inference of Manifold Density and Geometry by Doubly Stochastic Scaling*

Boris Landa† and Xiuyuan Cheng‡

Abstract. The Gaussian kernel and its traditional normalizations (e.g., row-stochastic) are popular approaches for assessing similarities between data points. Yet, they can be inaccurate under high-dimensional noise, especially if the noise magnitude varies considerably across the data, e.g., under heteroskedasticity or outliers. In this work, we investigate a more robust alternative—the doubly stochastic normalization of the Gaussian kernel. We consider a setting where points are sampled from an unknown density on a low-dimensional manifold embedded in high-dimensional space and corrupted by possibly strong, non–identically distributed, sub-Gaussian noise. We establish that the doubly stochastic affinity matrix and its scaling factors concentrate around certain population forms, and provide corresponding finite-sample probabilistic bounds. We then utilize these results to develop several tools for robust inference under general high-dimensional noise. First, we derive a robust density estimator that reliably infers the underlying sampling density and can substantially outperform the standard kernel density estimator under heteroskedasticity and outliers. Second, we obtain estimators for the pointwise noise magnitudes, the pointwise signal magnitudes, and the pairwise Euclidean distances between clean data points. Lastly, we derive robust graph Laplacian normalizations that accurately approximate various manifold Laplacians, including the Laplace–Beltrami operator, improving over traditional normalizations in noisy settings. We exemplify our results in simulations and on real single-cell RNA-sequencing data. For the latter, we show that in contrast to traditional methods, our approach is robust to variability in technical noise levels across cell types.

Key words. doubly stochastic, Sinkhorn scaling, graph Laplacian, affinity matrix, manifold learning, noise robustness, diffusion maps

MSC codes. 62R07, 62G

DOI. 10.1137/22M1516968

1. Introduction.

1.1. Traditional normalizations of the Gaussian kernel.

Many popular techniques for clustering, manifold learning, visualization, and semisupervised learning begin by learning similarities between observations. The learned similarities then form an affinity matrix that describes a weighted graph, which is further processed and analyzed according to the required task. A popular approach to construct an affinity matrix from the data is to evaluate the Gaussian kernel with pairwise Euclidean distances. Specifically, letting $y_1, \ldots, y_n \in \mathbb{R}^m$ be a

*Received by the editors August 22, 2022; accepted for publication (in revised form) February 22, 2023; published electronically July 20, 2023.
https://doi.org/10.1137/22M1516968
Funding: The work is supported by NSF DMS-2007040 and by NIH grant R01GM131642. The first author acknowledges support by NIH grants UM1DA051410, U54AG076043, and U01DA053628. The second author is also partially supported by NSF (DMS-1818945, DMS-1820827, DMS-2134037) and the Alfred P. Sloan Foundation.

†Program in Applied Mathematics, Yale University, New Haven, CT 06520 USA (boris.landa@yale.edu).
‡Department of Mathematics, Duke University, Durham, NC 27708 USA (xiuyuan.cheng@gmail.com).

collection of given data points, we define the Gaussian kernel $\mathcal{K}_\epsilon : \mathbb{R}^m \times \mathbb{R}^m \to \mathbb{R}_+$ and the resulting kernel matrix $K \in \mathbb{R}^{n\times n}$ as

$$K_{i,j} = \begin{cases} \mathcal{K}_\epsilon(y_i, y_j), & i \neq j, \\ 0, & i = j, \end{cases} \qquad \mathcal{K}_\epsilon(y_i, y_j) = \exp\left\{-\frac{\|y_i - y_j\|_2^2}{\epsilon}\right\} \tag{1.1}$$

for $i, j = 1, \ldots, n$, where $\epsilon > 0$ is a tunable bandwidth parameter. Here we adopt the version of the kernel matrix K whose main diagonal is zeroed-out, namely with no self-loops in the graph described by K. This choice will be further motivated from the viewpoint of noise robustness in section 1.3.

The kernel matrix K is often normalized to attain certain favorable properties. For instance, a popular choice is to divide each row of K by its sum to make it a transition probability matrix. A general family of normalizations that underlies many methods can be expressed by $P^{(\alpha)} \in \mathbb{R}^{n\times n}$ or $\hat{P}^{(\alpha)} \in \mathbb{R}^{n\times n}$ given by

$$\hat{P}^{(\alpha)}_{i,j} = \frac{P^{(\alpha)}_{i,j}}{\sum_{j=1}^n P^{(\alpha)}_{i,j}}, \qquad P^{(\alpha)} = D^{-\alpha} K D^{-\alpha}, \qquad D_{i,i} = \sum_{j=1}^n K_{i,j}, \tag{1.2}$$

where $\alpha \in [0, 1]$ is a parameter of the normalization, and $D \in \mathbb{R}^{n\times n}$ is a diagonal matrix whose main diagonal holds the degrees of the nodes in the weighted graph represented by K. If $\alpha = 0$, then $\hat{P}^{(\alpha)}$ describes the popular row-stochastic normalization $D^{-1}K$, and if $\alpha = 0.5$, then $P^{(\alpha)}$ describes its symmetric variant $D^{-1/2}KD^{-1/2}$. These normalizations have been utilized and extensively studied in the context of clustering [63, 52, 60, 28, 72], nonlinear dimensionality reduction (or manifold learning) [6, 17, 51, 73], image denoising [11, 53, 49, 42, 66], and graph-based signal processing and supervised learning [64, 18, 31, 22, 10].

An important theoretical aspect of various normalizations is the convergence of the corresponding graph Laplacian to a differential operator as $n \to \infty$ and $\epsilon \to 0$, typically under the assumption that the points $y_1, \ldots, y_n$ are sampled from a low-dimensional Riemannian manifold $\mathcal{M}$ embedded in $\mathbb{R}^m$; see [65, 15, 24, 32, 63, 69] and references therein. The family of normalizations in (1.2) was proposed in the diffusion maps paper [17], where it was shown that a population analogue (i.e., a continuous surrogate in the limit $n \to \infty$) of the graph Laplacian $I_n - \hat{P}^{(\alpha)}$ converges to a certain differential operator parametrized by α; see section 3.3 for more details. This operator is particularly appealing from the viewpoint of diffusion on the manifold $\mathcal{M}$; in the case of $\alpha = 1$, this operator is precisely the Laplace–Beltrami operator [29], which determines the solution to the heat equation and encodes the intrinsic geometry of the manifold regardless of the sampling density. Note that the parameter α in (1.2) determines the entrywise power of the degree matrix D, whose diagonal entries are $\sum_{j=1, j\neq i}^n \mathcal{K}_\epsilon(y_i, y_j)$, which can be interpreted as a kernel density estimator evaluated at the sample points $y_1, \ldots, y_n$ [54, 56]. Indeed, α controls the influence of the sampling density on the resulting affinity matrix and its spectral behavior [17].

1.2. The doubly stochastic normalization. An alternative normalization of K that is not covered by $\hat{P}^{(\alpha)}$ or $P^{(\alpha)}$ of (1.2) is the doubly stochastic normalization [81, 82, 48, 79]

$$W_{i,j} = d_i K_{i,j} d_j, \qquad \sum_{j=1}^n W_{i,j} = 1 \tag{1.3}$$

for $i = 1, \ldots, n$, where $d_1, \ldots, d_n > 0$ are the *scaling factors* of K. Since the resulting W is symmetric and stochastic (i.e., each row sums to 1), it is also doubly stochastic. The problem of finding $\mathbf{d} = [d_1, \ldots, d_n] > 0$ is known as *matrix scaling* and has been extensively studied in the literature; see [3, 34] and references therein. In the case of (1.3), the scaling factors are guaranteed to exist and are unique for $n > 2$ since K is fully indecomposable (see Lemma 1 in [42]). Although the scaling factors $\mathbf{d}$ do not admit a closed form solution, they can be found numerically, e.g., by the classical Sinkhorn–Knopp algorithm [67], convex optimization [2], or algorithms specialized for symmetric matrices [40, 79]. The doubly stochastic normalization is also closely related to entropic optimal transport [21, 55]. In particular, $W_{i,j}$ is the optimal transport plan between y_i and y_j according to the squared Euclidean distance loss with an entropic regularization term weighted by ϵ (see Proposition 2 in [42]).

Doubly stochastic affinity matrices proved useful for various tasks such as clustering [82, 5, 78, 44, 1, 23, 13], manifold learning [48, 79, 14], image denoising [50], and graph matching [19], often exhibiting more stable behavior and outperforming traditional normalizations. We note that requiring an affinity matrix to be doubly stochastic is appealing from a geometric perspective since the heat kernel on a compact Riemannian manifold—an operator describing affinities between points according to the intrinsic geometry—is always doubly stochastic [29].

In the context of manifold learning, the doubly stochastic normalization was recently investigated in [48, 79, 14]. Specifically, [48] analyzed a population setting where the scaling equation (1.3) is replaced with a more general family of integral equations parametrized by α (see (2.1) in section 2 for the special case of $\alpha = 0.5$). It was shown that the limiting differential operator can be made the same as for the traditional normalization $\hat{P}^{(\alpha)}$ from (1.2) as $\epsilon \to 0$ for any $\alpha \in [0, 1]$. In a different direction, [79] focused on the case where the manifold is a torus and derived spectral convergence rates for W as $n \to \infty$ and $\epsilon \to 0$, showing that the doubly stochastic normalization admits an improved bias error term compared to the traditional normalization in (1.2) when $\alpha = 0.5$. Lastly, [14] investigated a regularized version of the scaling equation (1.3) where the scaling factors are lower-bounded, and established operator convergence rates of the corresponding graph Laplacian to the same operator as for $\hat{P}^{(0.5)}$ as $n \to \infty$ and $\epsilon \to 0$ on general manifolds. We note that theoretical investigation of the doubly stochastic normalization is typically more challenging than that of more traditional normalizations, particularly due to the implicit and nonlinear nature of the scaling equation (1.3) and its population analogue (which is a nonlinear integral equation).

1.3. The influence of noise. Modern experimental procedures such as single-cell RNA-sequencing (scRNA-seq) [9, 35, 38], cryo-electron microscopy [4, 61, 33], and calcium imaging [47, 57, 45], to name a few, produce large datasets of high-dimensional observations often corrupted by strong noise. In such cases, the classical theoretical setup where data points reside on, or near, a low-dimensional manifold can be highly inaccurate. To model noisy data, we consider

$$y_i = x_i + \eta_i, \tag{1.4}$$

where $x_1, \ldots, x_n$ are the underlying clean observations and $\eta_1, \ldots, \eta_n \in \mathbb{R}^m$ are independent noise random vectors with zero means (to be specified in detail later on).

For identically distributed noise vectors $\{\eta_i\}_{i=1}^n$, the influence of noise on pairwise Euclidean distances and the entries of a kernel matrix was investigated in [25]. Specifically, if the noise magnitudes $\|\eta_i\|_2^2$ concentrate well in high dimension around a global constant c, it was shown that $\|y_i - y_j\|_2^2 \approx \|x_i - x_j\|_2^2 + 2c$ for all $i \neq j$. Consequently, the noisy Gaussian kernel $\mathcal{K}_\epsilon(y_i, y_j)$ is biased for $i \neq j$ by a global multiplicative factor. Since this factor does not exist on the main diagonal of the Gaussian kernel matrix, it is advantageous to zero it out (as done in K from (1.1)). This way, the multiplicative factor cancels out automatically through the traditional normalization $\hat{P}^{(\alpha)}$ (for any α) or $P^{(\alpha)}$ with $\alpha = 0.5$; see [26] for more details.

In many applications, the noise characteristics vary considerably across the data due to heteroskedasticity and outliers. In particular, heteroskedastic noise is prevalent in applications involving count or nonnegative data, typically modeled by, e.g., Poisson, binomial, negative binomial, or gamma distributions, whose variances inherently depend on their means, which can vary substantially across the data. Notable examples for such data are network traffic analysis [62], photon imaging [58], document topic modeling [77], scRNA-seq [30], and high-throughput chromosome conformation capture [36], among many others. Heteroskedastic noise also arises in natural image processing due to spatial pixel clipping [27] and in experimental procedures where conditions vary during data acquisition, such as in spectrophotometry and atmospheric data collection [16, 69]. Besides natural heteroskedasticity, experimental data often include outlier observations with abnormal noise distributions due to, e.g., abrupt deformations or technical errors during acquisition and storage. Consequently, to better understand the advantages of doubly stochastic normalization, it is important to investigate it under general non–identically distributed noise that supports heteroskedasticity and outliers.

If the noise vectors $\{\eta_i\}_{i=1}^n$ are not identically distributed or if $\|\eta_i\|_2^2$ do not concentrate well around a global constant, then the noisy Euclidean distances $\|y_i - y_j\|_2^2$ can be corrupted in a nontrivial way. In such cases, as demonstrated in [42], the Gaussian kernel and several of its traditional normalizations can behave unexpectedly and incorrectly assess the similarities between data points. On the other hand, [42] also shows that the doubly stochastic normalization is robust to non–identically distributed high-dimensional noise. Specifically, under suitable conditions on the noise, and if ϵ and n are fixed while m is growing, W converges pointwise in probability to its clean counterpart, even if the noise magnitudes $\|\eta_i\|_2^2$ are comparable to the signal magnitudes $\|x_i\|_2^2$ and fluctuate considerably. While this result highlights an important advantage of the doubly stochastic normalization, it does not account for the sample size n or the bandwidth ϵ. Hence, the population interpretation of the doubly stochastic normalization under noise in terms of the sampling density and the underlying geometry remains unclear.

1.4. Our results and contributions. In this work, we consider a setting where $x_1, \ldots, x_n$ are sampled from a low-dimensional manifold embedded in $\mathbb{R}^m$, and $\eta_1, \ldots, \eta_n$ are sampled from non–identically distributed sub-Gaussian noise that allows for heteroskedasticity and outliers. Our analysis is carried out in a high-dimensional regime in which the noisy Euclidean distances satisfy $\|y_i - y_j\|_2^2 = \|x_i - x_j\|_2^2 + \|\eta_i\|_2^2 + \|\eta_j\|_2^2 + o(1)$, where $\|\eta_i\|_2^2$ are unknown and can be large, and the $o(1)$ term is vanishing as $m \to \infty$ but is explicitly accounted for. Our main contributions are twofold. First, we characterize the pointwise behavior of the scaling factors $\mathbf{d}$ and the scaled matrix W in terms of the quantities in our setup for large m, n and

small ϵ; see section 2. Second, we build on these results to infer various quantities of interest from the noisy data and provide robust normalizations analogous to (1.2) with appropriate theoretical justification; see section 3. In addition, in section 4 we demonstrate our results on real single-cell RNA-sequencing data, and in section 5 we discuss future research directions. All proofs are deferred to the supplement (supplement.pdf [local/web 360KB]). Below is a detailed account of our results and contributions.

In section 2 we begin by considering the setting of fixed ϵ and large m, n. We establish that d_i and $W_{i,j}$ concentrate around certain quantities that depend explicitly on the clean Gaussian kernel $\mathcal{K}_\epsilon(x_i, x_j)$, the noise magnitudes $\|\eta_i\|_2^2$, and the solution to an integral equation that is the population analogue of (1.3). The associated error term is described via a probabilistic bound that is explicit in m, n, and the sub-Gaussian norm of the noise; see Theorem 2.4. Importantly, this result allows the noise magnitudes $\|\eta_i\|_2^2$ to be large and possibly diverge (stochastically) as $m, n \to \infty$, while the probabilistic errors in $\mathbf{d}$ and W are vanishing. Therefore, the doubly stochastic scaling is robust to the entrywise perturbations of the noise in large samples and high dimension simultaneously. Next, we turn to analyze the solutions to the aforementioned integral equation (see (2.1)) for small ϵ. In particular, we prove a first-order approximation in ϵ that depends on the sampling density and the manifold geometry; see Theorem 2.6. Overall, our results in section 2 show that for small ϵ and sufficiently large m and n, the noisy doubly stochastic affinity matrix $W_{i,j}$ approximates the clean Gaussian kernel $\mathcal{K}_\epsilon(x_i, x_j)$ up to a global constant and a multiplicative bias term that depends inversely on the square root of the sampling densities at x_i and x_j, but not on the noise magnitudes $\|\eta_i\|_2^2$. This is made possible by the scaling factors d_i, which "absorb" the noise magnitudes $\|\eta_i\|_2^2$, thereby correcting the Euclidean distances in the noisy Gaussian kernel. In particular, the scaling factor d_i depend exponentially on the noise magnitude $\|\eta_i\|_2^2$, and inversely on the square root of the sampling density at x_i; see (2.8) in section 2. To the best of our knowledge, these results are new even when no noise is present, as they describe the sample-to-population pointwise behavior of W and the scaling factors $\mathbf{d}$.

In section 3 we proceed by developing several tools for robust inference. First, we construct a robust density estimator by applying a nonlinearity to W and establish its convergence to the true density up to a global constant under appropriate conditions; see (3.1) and Theorem 3.3. We demonstrate that this approach can provide accurate density estimates on a manifold under strong heteroskedastic noise and outliers, whereas the standard kernel density estimator $D_{i,i}$ from (1.2) fails; see Figures 1–4. Second, we show that the scaling factors $\mathbf{d}$ and our robust density estimator can be combined to recover the noise magnitudes $\|\eta_i\|_2^2$, the signal magnitudes $\|x_i\|_2^2$, and the clean Euclidean distances $\|x_i - x_j\|_2^2$ up to small perturbations; see (3.4), (3.5), and Proposition 3.4. We demonstrate that these tools can be useful for detecting outliers, assessing the local quality of data, and identifying near neighbors more accurately; see Figures 5 and 6. Third, by utilizing our robust density estimator and the doubly stochastic matrix W, we provide a family of normalizations that is a robust analogue of the traditional normalizations in (1.2) and establish convergence of the corresponding graph Laplacians to the appropriate family of differential operators; see (3.10), (3.11), and Theorem 3.5. These normalizations can be used to obtain a more robust version of the diffusion maps method [17]. In particular, we demonstrate that in the case of $\alpha = 1$ and high-dimensional heteroskedastic noise, our approach provides a much more accurate approximation to the Laplace–Beltrami

operator than the traditional normalization of (1.2); see Figure 7. Overall, our results in section 3 show that it is possible to recover the sampling density and the manifold geometry under general high-dimensional noise, even when the noise magnitudes are nonnegligible and vary substantially. Importantly, our results show that this recovery is possible even when the ambient dimension grows slowly with the sample size, e.g., $m \propto n^{0.001}$.

Lastly, in section 4 we exemplify the tools derived in section 3 on experimental single-cell RNA-sequencing (scRNA-seq) data with cell type annotations. First, we show that our general-purpose noise magnitude estimator (derived in section 3.2) agrees with a popular model for explaining scRNA-seq data; see Figure 8a. Second, we show that our robust analogue of $\hat{P}^{(\alpha)}$ from (1.2) (derived in section 3.3) describes a more accurate and stable random walk behavior with respect to the ground truth cell types; see Figures 8b and 8c. The reason for this advantage is that different cell types have different levels of technical noise, which are automatically accounted for by our proposed robust normalization.

2. Large sample behavior of doubly stochastic scaling under high-dimensional noise. We consider the setting where the clean points $x_1, \dots, x_n$ are sampled independently from a probability measure $d\nu$ supported on a d-dimensional Riemannian manifold $\mathcal{M} \subset \mathbb{R}^m$. In particular, $d\nu = q(x)d\mu(x)$, where $d\mu(x)$ is the volume form of $\mathcal{M}$ at $x \in \mathcal{M}$ and $q(x)$ is a positive and continuous probability density function on $\mathcal{M}$. We further make the following assumption on $\mathcal{M}$.

Assumption 2.1. $\mathcal{M}$ is compact, smooth, with no boundary, and satisfies $\|x\|_2 \leq 1$ for all $x \in \mathcal{M}$.

A random vector $\eta \in \mathbb{R}^m$ is called sub-Gaussian if $\langle \eta, y \rangle$ is a sub-Gaussian random variable [74] for any $y \in \mathbb{R}^m$, where $\langle \cdot, \cdot \rangle$ is the standard scalar product. For each $x \in \mathcal{M}$, let $\eta(x) \in \mathbb{R}^m$ be a sub-Gaussian random vector with zero mean. Given the clean points $x_1, \dots, x_n$, the noise vectors $\eta_1, \dots, \eta_n$ are sampled independently from $\eta(x_1), \dots, \eta(x_n)$, respectively. Therefore, each clean point x_i is first sampled independently from $\mathcal{M}$ according to the density function $q(x)$, and then each noisy observation y_i is produced by (1.4) according to the realization of the random vector $\eta(x_i)$. Let $\|\eta(x)\|_{\psi_2}$ be the sub-Gaussian norm of $\eta(x)$ [74], given by $\|\eta(x)\|_{\psi_2} = \sup_{\|y\|_2=1} \|\langle \eta(x), y \rangle\|_{\Psi_2}$, where $\|\cdot\|_{\Psi_2}$ in the right-hand side is the sub-Gaussian norm of a random variable [74]. To control the magnitude of the noise, we make the following assumption.

Assumption 2.2. $E := \max_{x \in \mathcal{M}} \|\eta(x)\|_{\psi_2} \leq C/(m^{1/4}\sqrt{\log m})$ for a constant $C > 0$.

For instance, Assumption 2.2 holds if $\eta(x)$ is a multivariate normal with covariance $\Sigma(x) \in \mathbb{R}^{m \times m}$ satisfying $\|\Sigma(x)\|_2 \leq C/(m^{1/4}\sqrt{\log m})$ for all $x \in \mathcal{M}$. Observe that this includes the special case $\Sigma(x) = I_m/\sqrt{m}$, where $I_m \in \mathbb{R}^{m \times m}$ is the identity matrix. In this case, the noise magnitude at $x \in \mathcal{M}$ is $\mathbb{E}\|\eta(x)\|_2^2 = \mathrm{Tr}\{\Sigma(x)\} = 1$, which is equal to or greater than the magnitude of the clean point $\|x\|_2^2$ (by Assumption 2.1). Moreover, in the more extreme case of $\Sigma(x) = I_m/(m^{1/4}\sqrt{\log m})$, the noise magnitude is $\mathbb{E}\|\eta(x)\|_2^2 = \sqrt{m}/\log m$, which is growing in m and can be much larger than $\|x\|_2^2$. Note that the distribution of the noise can vary across $x \in \mathcal{M}$. In particular, we can have regions of $\mathcal{M}$ where $\mathbb{E}\|\eta(x)\|_2^2$ is very large and others where it is very small, allowing for considerable noise heteroskedasticity. Even if $\eta(x)$ is identically distributed across $x \in \mathcal{M}$, the norm $\|\eta(x)\|_2$ is allowed to have a heavy tail that prohibits $\|\eta(x)\|_2$ from concentrating around a global constant. For instance, we can take $\eta(x)$

to be the zero vector with probability $p \in (0,1)$ and sampled uniformly from a bounded subset of $\mathbb{R}^m$ with probability $1-p$. This is a useful model for describing outliers, in which case the magnitudes $\|\eta(x_i)\|_2^2$ can fluctuate substantially over $i = 1, \ldots, n$. Lastly, we emphasize that the coordinates of $\eta(x)$ need not be independent or identically distributed. Figure 2 in section 3 provides a two-dimensional visualization of prototypical noise models covered in our setting.

To analyze the doubly stochastic scaling for large sample size and high dimension, we require the dimension to be at least some fractional power of the sample size; that is, we assume the following.

Assumption 2.3. $m \geq n^\gamma$ for a constant $\gamma > 0$.

We note that the requirement $m \geq n^\gamma$ can be modified to include an arbitrary constant $c > 0$, namely $m \geq cn^\gamma$. This constant was set to 1 for simplicity.

We treat the quantities d, γ, C, $q(x)$, and the geometry of $\mathcal{M}$ (e.g., curvature, reach) as fixed and independent of m, n, ϵ, and E, which can vary. To fix the geometry of $\mathcal{M}$ and make it independent of m, one can consider a manifold that is first embedded in $\mathbb{R}^r$ for fixed $r > 0$, and then embedded in $\mathbb{R}^m$ for any $m > r$ via a rigid transformation (i.e., a composition of rotations, translations, and reflections). Rigid transformations preserve the pairwise Euclidean distances $\{\|x-y\|_2\}_{x,y\in\mathcal{M}}$, thereby making all geometric properties of $\mathcal{M}$ (both intrinsic and extrinsic) independent of m.

To state our results, we introduce the positive function $\rho_\epsilon : \mathcal{M} \to \mathbb{R}_+$ that solves the integral equation

$$\frac{1}{(\pi\epsilon)^{d/2}} \int_{\mathcal{M}} \rho_\epsilon(x) \mathcal{K}_\epsilon(x,y) \rho_\epsilon(y) q(y) d\mu(y) = 1 \tag{2.1}$$

for all $x \in \mathcal{M}$. The integral equation (2.1) and the scaling function $\rho_\epsilon(x)$ can be interpreted as population analogues of the scaling equation (1.3) and the scaling factors $\mathbf{d}$, respectively. Due to the compactness of $\mathcal{M}$ and the positivity and continuity of $\mathcal{K}_\epsilon(x,y)$ and $q(x)$, the results in [8] (see in particular Theorem 5.2 and Corollaries 4.12 and 4.19 therein) guarantee that the solution $\rho_\epsilon(x)$ exists and is a unique positive and continuous function on $\mathcal{M}$ (see also [41]). Note that $\rho_\epsilon(x)$ depends on the kernel bandwidth parameter ϵ. Table 1 summarizes common symbols and notation used in the results of this section.

We now have the following theorem, which characterizes the scaled matrix W and the scaling factors $\mathbf{d}$ for large sample size n and high dimension m.

Theorem 2.4. *Under Assumptions 2.1, 2.2, 2.3, there exist $t_0, m_0(\epsilon), n_0(\epsilon), C'(\epsilon) > 0$, such that for all $m > m_0(\epsilon)$ and $n > n_0(\epsilon)$, we have*

$$W_{i,j} = \frac{\rho_\epsilon(x_i)\mathcal{K}_\epsilon(x_i,x_j)\rho_\epsilon(x_j)}{(n-1)(\pi\epsilon)^{d/2}} (1 + \mathcal{E}_{i,j}), \tag{2.2}$$

$$d_i = \frac{\rho_\epsilon(x_i)}{\sqrt{(n-1)(\pi\epsilon)^{d/2}}} \exp\left(\frac{\|\eta_i\|_2^2}{\epsilon}\right)(1 + \mathcal{E}_{i,i}) \tag{2.3}$$

Table 1
Common symbols and notation.

$\mathcal{M}$	d-dimensional manifold embedded in $\mathbb{R}^m$
d	Intrinsic dimension of $\mathcal{M}$
m	Dimension of the ambient space
n	Number of data points
$q(x)$	Sampling density at $x \in \mathcal{M}$
x_i	Clean data points on $\mathcal{M}$
y_i	Noisy data points in $\mathbb{R}^m$
η_i	Noise vectors in $\mathbb{R}^m$
$\mathcal{K}_\epsilon(\cdot,\cdot)$	Gaussian kernel
ϵ	Kernel bandwidth parameter
$\rho_\epsilon(x)$	Function solving the integral eq. (2.1) at $x \in \mathcal{M}$
K	Noisy $n \times n$ Gaussian kernel matrix
W	Noisy $n \times n$ doubly stochastic affinity matrix
$\mathbf{d}$	Vector of n scaling factors solving (1.3)
E	Maximal sub-Gaussian norm of the noise
$\Delta_\mathcal{M}$	The (negative) Laplace–Beltrami operator on $\mathcal{M}$
$d\mu(x)$	Volume form of $\mathcal{M}$ at $x \in \mathcal{M}$

for all $i, j = 1, \ldots, n$, $i \neq j$, *where* $\rho_\epsilon(x)$ *is the solution to* (2.1), *and*

$$\max_{i,j} |\mathcal{E}_{i,j}| \leq tC'(\epsilon) \max\left\{ E\sqrt{\log m}, E^2\sqrt{m \log m}, \sqrt{\frac{\log n}{n}} \right\}, \tag{2.4}$$

with probability at least $1 - n^{-t}$ *for any* $t > t_0$.

Theorem 2.4 provides explicit asymptotic expressions for the doubly stochastic matrix W and the associated scaling factors $\mathbf{d}$ in terms of the quantities in our setup, as well as a high-probability bound on the relative pointwise errors $\mathcal{E}_{i,j}$ with explicit dependence on m, n, and E. We note that $t_0, m_0(\epsilon), n_0(\epsilon), C'(\epsilon)$ in Theorem 2.4 all depend on d, γ, C, $q(x)$, and on the geometry of $\mathcal{M}$, which are considered as fixed in our setup. The notation $m_0(\epsilon), n_0(\epsilon), C'(\epsilon)$ means that these quantities additionally depend on ϵ.

Observe that under Assumption 2.2, the quantities $E\sqrt{\log m}$ and $E^2\sqrt{m \log m}$ appearing in (2.4) always converge to zero as $m \to \infty$. Moreover, since m is growing with n by Assumption 2.3, all three quantities $E\sqrt{\log m}$, $E^2\sqrt{m \log m}$, and $\sqrt{\log n/n}$ converge to zero as $n \to \infty$. Consequently, Theorem 2.4 implies that if we fix a bandwidth parameter ϵ, then all pointwise errors $\mathcal{E}_{i,j}$ appearing in (2.2) and (2.3) converge almost surely to zero as $n \to \infty$.

According to (2.4), the convergence rate of $\max_{i,j} |\mathcal{E}_{i,j}|$ to zero is bounded by the largest among $E\sqrt{\log m}$, $E^2\sqrt{m \log m}$, and $\sqrt{\log n/n}$. If no noise is present, i.e., $E = 0$, this rate is bounded by $\sqrt{\log n/n}$, which describes the sample-to-population convergence and is independent of the ambient dimension m. In fact, in this case we do not actually require Assumptions 2.2 and 2.3. On the other hand, if noise is present, then the convergence rate depends also on the maximal sub-Gaussian norm E and on the ambient dimension m. Let us suppose for simplicity that $m \propto n$. In this case, as long as $E \leq C/\sqrt{m}$, then the convergence rate of $\max_{i,j} |\mathcal{E}_{i,j}|$ to zero is still bounded by $\sqrt{\log n/n}$. As discussed earlier, a simple example for this case is if $\eta(x)$ is multivariate normal with covariance $\Sigma(x)$ satisfying $\|\Sigma(x)\|_2 \leq C/\sqrt{m}$,

which allows the magnitude of the noise $\mathbb{E}\|\eta(x)\|_2^2$ to be comparable to the signal magnitude $\|x\|_2^2$. If $m \propto n$ and E decays more slowly than $1/\sqrt{m}$, then the bound on $\max_{i,j} |\mathcal{E}_{i,j}|$ becomes dominated by the term $E^2\sqrt{m \log m}$, which converges to zero as $m, n \to \infty$ even though the noise magnitude $\mathbb{E}\|\eta(x)\|_2^2$ can possibly diverge (see the discussion following Assumption 2.2).

The proof of Theorem 2.4 can be found in section SM2 of the supplement and relies on two main ingredients. The first is the decomposition $\mathcal{K}_\epsilon(y_i, y_j) = \exp\{-\|y_i\|_2^2/\epsilon\} \exp\{2\langle y_i, y_j\rangle/\epsilon\}$ $\exp\{-\|y_j\|_2^2/\epsilon\}$, which can be viewed as diagonal scaling of the nonnegative matrix $(\exp\{2\langle y_i, y_j\rangle/\epsilon\})$, together with the fact that the inner products $\langle y_i, y_j\rangle$ concentrate around their clean counterparts $\langle x_i, x_j\rangle$ for $i \neq j$ under high-dimensional sub-Gaussian noise; see Lemma SM1.1 in supplement section SM1. The second ingredient is a refined stability analysis of the scaling factors of a symmetric nonnegative matrix with zero main diagonal; see Lemma SM1.2 in supplement section SM1, which improves upon the analysis in [42]. These two ingredients are combined with a perturbation analysis of the Gaussian kernel and large-sample concentration arguments to prove the results in Theorem 2.4.

Theorem 2.4 asserts that for sufficiently large m and n, $W_{i,j}$ is close to the clean Gaussian kernel $\mathcal{K}_\epsilon(x_i, x_j)$ up to a constant factor and the multiplicative bias term $\rho_\epsilon(x_i)\rho_\epsilon(x_j)$. This bias term is determined by the scaling function $\rho_\epsilon(x)$ that solves (2.1), which depends on the geometry of $\mathcal{M}$ and the density $q(x)$ in a nontrivial way and does not admit a closed form expression in general. Nonetheless, we can provide an explicit approximation to $\rho_\epsilon(x)$ when ϵ is small. To that end, we first assume that $q(x)$ is sufficiently smooth on $\mathcal{M}$; specifically we assume the following.

Assumption 2.5. $q \in \mathcal{C}^6(\mathcal{M})$.

Let $p \geq 1$ and define $\|\cdot\|_{L^p(\mathcal{M},d\mu)}$ as the standard L^p norm on $\mathcal{M}$ with measure $d\mu$, i.e., $\|f\|_{L^p(\mathcal{M},d\mu)} = (\int_{\mathcal{M}} |f_\epsilon(x)|^p d\mu)^{1/p}$ for $f : \mathcal{M} \to \mathbb{R}$. In addition, we denote by $\Delta_{\mathcal{M}}\{f\}(x)$ the negative Laplace–Beltrami operator on $\mathcal{M}$ applied to f and evaluated at $x \in \mathcal{M}$. We now have the following result.

Theorem 2.6. *Under Assumptions 2.1 and 2.5, for any $p \in [1, 4/3)$, there exist constants $\epsilon_0, c_p > 0$ and a function $\omega : \mathcal{M} \to \mathbb{R}$ that depends only on the geometry of $\mathcal{M}$, such that for all $\epsilon \leq \epsilon_0$,*

$$\left\| \rho_\epsilon - q^{-1/2} F_\epsilon \right\|_{L^p(\mathcal{M},d\mu)} \leq c_p \epsilon^2, \tag{2.5}$$

where $\rho_\epsilon(x)$ is the solution to (2.1), and

$$F_\epsilon(x) = 1 - \frac{\epsilon}{8}\left(\omega(x) - \frac{\Delta_{\mathcal{M}}\{q^{-1/2}\}(x)}{\sqrt{q(x)}}\right). \tag{2.6}$$

If $d \leq 5$, then (2.5) holds for any $p \in [1, 2]$, and moreover, there exist $\epsilon_0', c' > 0$, such that for all $\epsilon \leq \epsilon_0'$ and $x \in \mathcal{M}$,

$$\left| \rho_\epsilon(x) - q^{-1/2}(x) F_\epsilon(x) \right| \leq c' \epsilon^{2-d/4}. \tag{2.7}$$

The first part of Theorem 2.6 provides an asymptotic approximation to $\rho_\epsilon(x)$ with an $L^p(\mathcal{M}, d\mu)$ error of $\mathcal{O}(\epsilon^2)$ for $p < 4/3$. This approximation is equal to $q^{-1/2}$ to zeroth-order

with a first-order correction term that depends additionally on the manifold geometry and on the smoothness of the density. The second part of Theorem 2.6 improves upon this result in the case of $d \leq 5$, where the convergence now is in $L^2(\mathcal{M}, d\mu)$ with the same rate of $\mathcal{O}(\epsilon^2)$. Moreover, in this case we have uniform pointwise convergence on $\mathcal{M}$ with rate at least $\mathcal{O}(\epsilon^{2-d/4})$. If $d \leq 3$, then this result implies the first-order pointwise asymptotic approximation $\rho_\epsilon(x) \sim q^{-1/2}(x) F_\epsilon(x)$ uniformly for $x \in \mathcal{M}$. Otherwise, if $d = 4$ or $d = 5$, (2.7) only implies the zeroth-order pointwise asymptotic approximation $\rho_\epsilon(x) \sim q^{-1/2}(x)$ (since the error in the right-hand side of (2.7) becomes $\mathcal{O}(\epsilon)$ or larger). We note that the expression $q^{-1/2} F_\epsilon$ was also used in [14] to construct an approximate solution to (2.1), yet our results here prove the convergence of ρ_ϵ to $q^{-1/2} F_\epsilon$, which did not appear previously.

The proof of Theorem 2.6 can be found in section SM3 of the supplement and is based on the following approach. First, we construct a certain covering of $\mathcal{M}$ to show that the measure of the set $\{x : \rho_\epsilon(x) > t\}$ is upper bounded by c/t^2 for some constant $c > 0$ that depends only on the manifold $\mathcal{M}$ and the density q; see Lemma SM1.3 in supplement section SM1. Then, to establish (2.5), we make use of a technical manipulation of the integral equation (2.1) that relies on the aforementioned Lemma SM1.3, the positive definiteness of the Gaussian kernel (as an integral operator), and the asymptotic expansion developed in [17] (see also Lemma SM1.4 in supplement section SM1). In the special case of $d \leq 5$, the $L_p(\mathcal{M}, d\mu)$ convergence in (2.5) together with Holder's inequality allow us to refine the previous analysis and establish the remaining claims.

By combining Theorems 2.4 and 2.6, we can describe the convergence of d_i and $W_{i,j}$ to population forms that do not depend on the manifold $\mathcal{M}$ (to zeroth-order in ϵ). In particular, if $d \leq 5$, we are guaranteed that in the asymptotic regime where $m, n \to \infty$ and $\epsilon \to 0$ sufficiently slowly, we have

$$d_i \sim \frac{1}{\sqrt{(n-1)(\pi\epsilon)^{d/2} q(x_i)}} \exp\left(\frac{\|\eta_i\|_2^2}{\epsilon}\right), \qquad W_{i,j} \sim \frac{\mathcal{K}_\epsilon(x_i, x_j)}{(n-1)(\pi\epsilon)^{d/2}\sqrt{q(x_i)q(x_j)}}, \tag{2.8}$$

almost surely for all indices $i \neq j$. Hence, if the sampling density on $\mathcal{M}$ is uniform, i.e., $q(x)$ is a constant function, then $W_{i,j}$ approximates the clean Gaussian kernel $\mathcal{K}_\epsilon(x_i, x_j)$ for all $i \neq j$ up to a global constant, even if the noise magnitudes $\|\eta_i\|_2^2$ are large and fluctuate considerably. In this case, the variability of the scaling factors d_i corresponds to the variability of $\|\eta_i\|_2^2$, where large values of d_i correspond to strong noise, and vice versa. If the density is not uniform, then $W_{i,j}$ and d_i are also affected by the variability of the density $q(x_i)$. Nonetheless, this effect can be removed by estimating the density and correcting d_i and $W_{i,j}$ accordingly; see sections 3.1 and 3.2 for more details.

3. Applications to inference of density and geometry.

In this section, we utilize the doubly stochastic scaling (1.3) and the results in the previous section to infer various quantities of interest from the noisy data. All numerical experiments described in this section use the scaling algorithm of [79] to solve (1.3) with a tolerance of 10^{-9}. To simplify the analysis and statements of the results presented in this section, we work under the following assumption that extends the pointwise first-order convergence of $\rho_\epsilon(x)$ in Theorem 2.6 to arbitrary intrinsic dimension d; see Remark 3.2 below.

Assumption 3.1. There exist $\beta \in (0,1]$, $\epsilon_0' > 0$, and $c' > 0$, such that for all $\epsilon \le \epsilon_0'$ and $x \in \mathcal{M}$, $|\rho_\epsilon(x) - q^{-1/2}(x)F_\epsilon(x)| \le c'\epsilon^{1+\beta}$.

Remark 3.2. Assumption 3.1 requires that $\rho_\epsilon(x)$ (the solution to (2.1)) is approximated by $q^{-1/2}(x)F_\epsilon(x)$ uniformly on $\mathcal{M}$ with an error of $\mathcal{O}(\epsilon^{1+\beta})$ for some $\beta \in [0,1)$. According to Theorem 2.6, Assumption 3.1 is immediately satisfied for any $d \le 3$ (with $\beta = 1 - d/4$) under Assumptions 2.1 and 2.5. We conjecture that this property also holds in more general settings and for higher intrinsic dimensions, currently not covered by Theorem 2.6. We therefore rely on Assumption 3.1 to simplify the presentation of our results in this section and state them in more generality for arbitrary intrinsic dimensions. We note that all numerical examples in this section were conducted in settings with $d = 1$ that satisfy Assumption 3.1.

3.1. Robust manifold density estimation. Since the asymptotic expression of the doubly stochastic kernel $W_{i,j}$ in (2.8) is invariant to the noise magnitudes $\|\eta_i\|_2^2$, it is natural to employ $W_{i,j}$ to infer the probability density $q(x_i)$. Recall that the standard kernel density estimator (KDE) using the Gaussian kernel at x_i is given by $D_{i,i}/(n-1) = \sum_{j=1,\, j\neq i}^{n} \mathcal{K}_\epsilon(x_i, x_j)/(n-1)$, which approximates $(\pi\epsilon)^{d/2} q(x_i)$ asymptotically for large n and small ϵ (see [80] and references therein). Clearly, we cannot directly replace the Gaussian kernel in the KDE with W since $\sum_{j=1}^{n} W_{i,j} = 1$. Instead, we propose to employ the nonlinearity $\sum_{j=1}^{n} [W_{i,j}]^s$ for $s > 0$, $s \neq 1$, where $[W_{i,j}]^s$ is the sth power of $W_{i,j}$. Specifically, we define the doubly stochastic kernel density estimator (DS-KDE) as

$$\hat{q}_i = \frac{1}{n-1} \left(\sum_{j=1}^{n} [W_{i,j}]^s \right)^{1/(1-s)} \tag{3.1}$$

for $i = 1, \ldots, n$. We now have the following result.

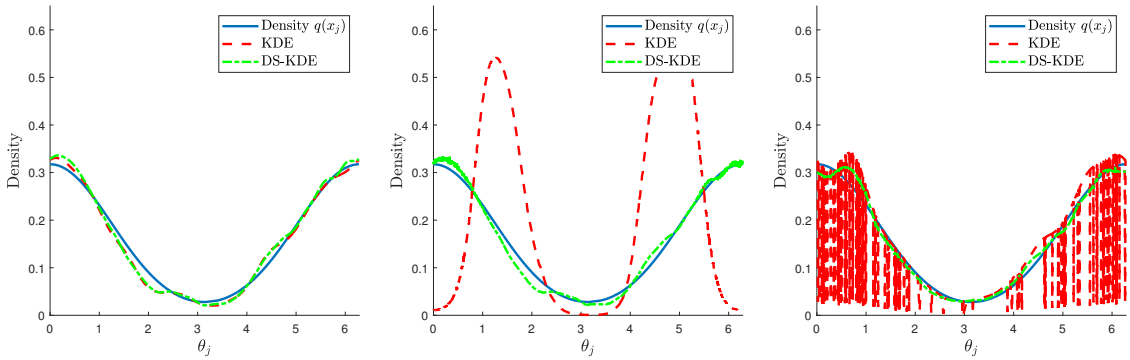
Theorem 3.3. *Fix $s > 0$, $s \neq 1$. Under Assumptions 2.1–3.1, there exist $\epsilon_0, t_0, m_0(\epsilon), n_0(\epsilon)$, $C'(\epsilon) > 0$, such that for all $\epsilon < \epsilon_0$, $m > m_0(\epsilon)$, $n > n_0(\epsilon)$, we have*

$$\hat{q}_i = (\pi\epsilon)^{d/2} s^{d/(2(s-1))} q(x_i) \left[1 + \mathcal{O}(\epsilon) + \mathcal{E}_i^{(1)} \right] \tag{3.2}$$

for all $i = 1, \ldots, n$, where $\max_i |\mathcal{E}_i^{(1)}|$ is upper bounded by the right-hand side of (2.4) with probability at least $1 - n^{-t}$ for any $t > t_0$.

We note that the quantities $\epsilon_0, t_0, m_0(\epsilon), n_0(\epsilon), C'(\epsilon)$ appearing in Theorem 3.3 need not be the same as those in Theorem 2.4, and may additionally depend on s and β, which are considered as fixed constants independent of m, n, E, and ϵ. The proof of Theorem 3.3 can be found in supplement section SM4.

Theorem 3.3 establishes that up to the constant factor $(\pi\epsilon)^{d/2} s^{d/(2(s-1))}$, the DS-KDE $\hat{q}_i$ approximates the density $q(x_i)$ for all $i = 1, \ldots, n$ with a bias error of $\mathcal{O}(\epsilon)$ and a variance error $\mathcal{E}_i^{(1)}$ that has the same behavior as $\mathcal{E}_{i,j}$ in (2.4). In particular, for sufficiently small ϵ and sufficiently large m and n (which depend also on ϵ), the quantity $\hat{q}_i$ can approximate $(\pi\epsilon)^{d/2} s^{d/(2(s-1))} q(x_i)$ with high probability up to an arbitrarily small relative error. Therefore, $\hat{q}_i$ can be serve as a density estimator that is robust to the high-dimensional noise in our setup.



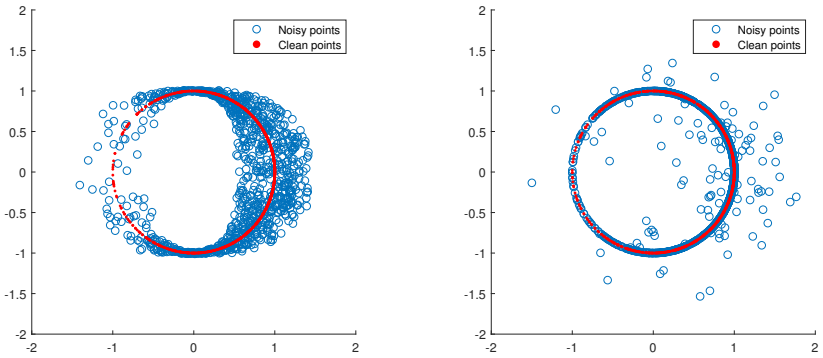
(a) Clean data (b) Varying noise (c) Outlier-type noise

Figure 1. *Example of density estimation on the unit circle using DS-KDE from* (3.1) *versus the standard KDE in clean and noisy scenarios, where* $n = m = 2000$*,* $s = 2$*, and* $\epsilon = 0.1$*. The angles* $\theta_i \in [0, 2\pi)$ *were sampled from* $\mathcal{N}(0, 0.16\pi^2)$ *modulo* 2π*. Panel* (a)*: Clean data. Panel* (b)*: Heteroskedastic noise with smoothly varying magnitude (see Figure* 2a*). Panel* (c)*: Identically distributed outlier type noise (see Figure* 2b*).*

We now demonstrate the advantage of the DS-KDE over the standard KDE via a toy example. We simulated $n = 2000$ points from the unit circle in $\mathbb{R}^2$ and embedded them in $\mathbb{R}^m$ with $m = 2000$ by applying a random orthogonal transformation. The angle of each clean point x_i, denoted by $\theta_i \in [0, 2\pi)$, was sampled from $\mathcal{N}(0, 0.16\pi^2)$ modulo 2π, where $\mathcal{N}(\mu, \sigma^2)$ is the standard univariate normal distribution. The resulting sampling density $q(x)$ on the circle can be seen in Figure 1a. We also depict the outputs of the standard KDE and the DS-KDE with $s = 2$ and $\epsilon = 0.1$. It is evident that without noise, both estimators provide similarly accurate estimates of $q(x_i)$, noting that we normalized the standard KDE by $(\pi\epsilon)^{d/2}$ and the DS-KDE by $(\pi\epsilon)^{d/2} s^{d/(2(s-1))}$.

Next, we simulated two types of high-dimensional noise. First, we added noise η_i sampled uniformly from a ball in $\mathbb{R}^m$ with radius $0.01 + 0.49(1 + \cos(\theta_i))/2$, where θ_i is the angle of x_i on the unit circle. Hence, the expected noise magnitude varies smoothly between 0.01 and 0.5 along the circle; see Figure 2a for a two-dimensional visualization. Figure 1b depicts the standard KDE as well as our robust density estimator $\hat{q}_i$ versus the true density $q(x_i)$. We observe that the standard KDE produces an estimate that is very different from the true density $q(x_i)$, and has more to do with the noise magnitudes $\|\eta_i\|_2^2$ in the data. On the other hand, the DS-KDE is robust to the magnitudes of the noise, and produces an estimate that is nearly as accurate as in the clean case. For the second type of noise, we took each η_i to be the zero vector with probability $p = 0.9$ and sampled it from a multivariate normal with covariance $I_m/(4m)$ with probability $1 - p = 0.1$, thereby simulating identically distributed outlier-type noise; see Figure 2b for a two-dimensional visualization. Figure 1c shows that in this case, the standard KDE suffers from pointwise drops in the estimated density. Essentially, these drops stem from the nonzero realizations of the noise, i.e., the "outliers," whose large noise magnitudes inflate the pairwise Euclidean distances. On the other hand, the DS-KDE produces an estimate that is invariant to the outliers and is very close to $q(x_i)$.

It is interesting to point out that although the DS-KDE is undefined when $s = 1$, the limiting case of $s \to 1$ is interpretable and can be implemented. In particular, according to (3.1), a direct calculation shows that



(a) Noise with smoothly varying magnitude

(b) Outlier type noise

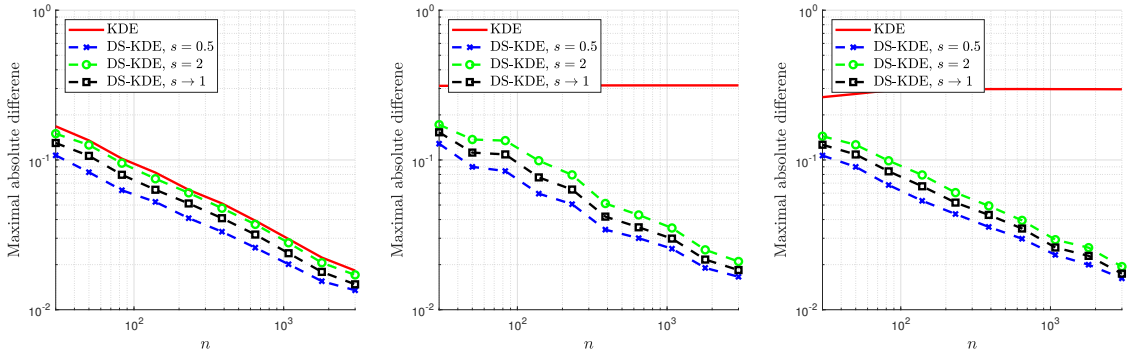
Figure 2. *Two-dimensional visualization of prototypical noise models used in our experiments. The clean data x_i are sampled from the unit circle with nonuniform density as in Figure* 1a*. Panel* (a)*: Heteroskedastic noise with smoothly varying magnitude, where η_i is sampled uniformly from a ball in $\mathbb{R}^m$ with radius* $0.01 + 0.49(1+\cos(\theta_i))/2$ *(θ_i is the angle of x_i on the unit circle). Panel* (b)*: Identically distributed outlier-type noise, where η_i is zero with probability* 0.9 *and sampled from a multivariate normal with covariance $I_m/(4m)$ with probability* 0.1*.*

$$\lim_{s\to 1} \hat{q}_i = \frac{1}{n-1} \exp\left\{ - \sum_{j=1,\, j\neq i}^{n} W_{i,j} \log(W_{i,j}) \right\}. \tag{3.3}$$

The right-hand side of (3.3), up to the factor $1/(n-1)$, is known as the perplexity of the ith row of W, where the expression inside the exponent in (3.3) is the entropy. According to Theorem 3.3, we expect the right-hand side of (3.3) to approximate $(\pi\epsilon)^{d/2} s^{d/(2(s-1))} q(x_i) \to (\pi e \epsilon)^{d/2} q(x_i)$ as $s \to 1$, which provides an explicit relation between the entropy of each row of the doubly stochastic kernel W and the sampled density $q(x_i)$. We mention that Theorem 3.3 does not strictly cover the limit $s \to 1$ since the dependence of the bias and variance errors on s is harder to track and is not made explicit. However, the numerical experiments described below suggest that the conclusions of Theorem 3.3 also hold for $s \to 1$ and that the performance of the density estimator in this case is comparable to other choices of s over a range of bandwidth parameters ϵ.

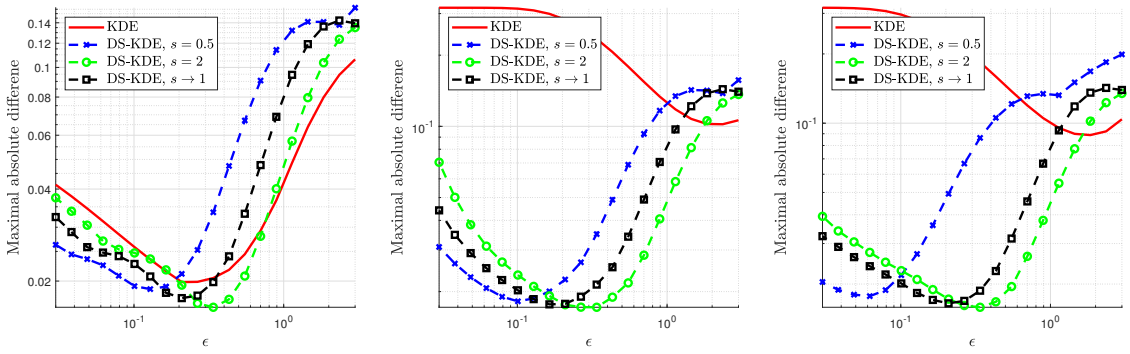
Figure 3 illustrates the maximal density estimation errors (over $i = 1, \ldots, n$) for the standard KDE as well as the DS-KDE as functions of n, for $m = n$ and $s = 0.5$, $s = 2$, and $s \to 1$, where $\epsilon = 0.1$. We used the same noise settings as for Figure 2, and the displayed errors were averaged over 50 randomized trials. In the clean case, the KDE and the DS-KDE perform similarly, where all errors decrease with n at a rate close to $n^{-1/2}$, which agrees with Theorem 3.3 and 2.4 up to a logarithmic factor. In both noisy cases however, the KDE error saturates at a high level and does not decrease further, whereas the DS-KDE errors decrease roughly at the same rate as in the clean case. In particular, the DS-KDE errors for $n = 3000$ are over an order of magnitude smaller than the standard KDE error.

In Figure 4, we depict the maximal density estimation errors for the standard KDE and DS-KDE (3.1) as functions of ϵ for $s = 0.5$, $s = 2$, and $s \to 1$, where $m = n = 2000$. We used the same noise settings as for Figure 2, and the displayed errors were averaged over 10



(a) Clean data (b) Varying noise (c) Outlier-type noise

Figure 3. *Density estimation errors for the standard KDE and the DS-KDE from (3.1) (normalized by the corresponding global constant) versus the number of samples n and several values of s for clean and noisy scenarios, where $m = n$. The density $q(x)$ and noise models are the same as for Figures 1a, 1b, and 1c, respectively; see also Figures 2a and 2b.*



(a) Clean data (b) Varying noise (c) Outlier-type noise

Figure 4. *Density estimation errors for the standard KDE and the DS-KDE from (3.1) (normalized by the corresponding global constant) versus the bandwidth parameter ϵ and several values of s for clean and noisy scenarios, where $m = n = 2000$. The density $q(x)$ and noise models are the same as for Figures 1a, 1b, and 1c, respectively; see also Figures 2a and 2b.*

randomized trials. We observe that in the clean case, all density estimators perform similarly well, attaining errors of about 0.02 for the best values of ϵ, with a small advantage to the DS-KDE with $s = 2$. Yet, in the noisy scenarios, the standard KDE can only achieve an error of about 0.1, which requires using a large bandwidth parameter, while the DS-KDE behaves similarly to the clean case and achieves significantly smaller errors. As expected from Theorem 3.3, we see the prototypical bias-variance trade-off in all noise scenarios, where the error of the DS-KDE is dominated by the bias term $\mathcal{O}(\epsilon)$ for large ϵ, and dominated by the variance error $\max_i |\mathcal{E}_i^{(1)}|$ for small ϵ. However, while the strong noise forces the standard KDE to use a large bandwidth ϵ (proportional to the magnitude of the noise) to achieve the smallest error in the bias-variance trade-off, the DS-KDE does not suffer from this issue and achieves small errors even when the bandwidth is much smaller than the noise magnitudes.

3.2. Recovering noise magnitudes, signal magnitudes, and Euclidean distances. According to the asymptotic expression for the scaling factors d_i in (2.8), we can extract the noise magnitudes $\|\eta_i\|_2^2$ from d_i (up to a global constant) if we know the density $q(x_i)$. Since we do not have access to $q(x_i)$ directly, we replace it with its estimate $\hat{q}_i$ from (3.1) and define

$$\hat{N}_i = \epsilon \log\left(d_i \sqrt{(n-1)\hat{q}_i}\right) \tag{3.4}$$

for $i = 1, \ldots, n$, which serves as an estimator for the noise magnitude $\|\eta_i\|_2^2$. In our setup of high-dimensional noise (Assumptions 2.2 and 2.3), we have $\|y_i\|_2^2 \approx \|x_i\|_2^2 + \|\eta_i\|_2^2$ and $\|y_i - y_j\|_2^2 \approx \|x_i - x_j\|_2^2 + \|\eta_i\|_2^2 + \|\eta_j\|_2^2$; see Lemma SM1.1 and the proof of Theorem 2.4. Hence, we can infer the signal magnitudes $\|x_i\|_2^2$ and the pairwise Euclidean distances $\|x_i - x_j\|_2^2$ according to

$$\hat{S}_i = \|y_i\|_2^2 - \hat{N}_i, \qquad \hat{D}_{i,j} = \|y_i - y_j\|_2^2 - \hat{N}_i - \hat{N}_j, \tag{3.5}$$

respectively, for $i, j = 1, \ldots, n$ with $j \neq i$. Equivalently, $\hat{D}_{i,j}$ from (3.5) can be derived directly from $W_{i,j}$ by canceling out the term $(q(x_i)q(x_j))^{-1/2}$ appearing in (2.8) via the density estimator $\hat{q}_i$, that is,

$$\hat{D}_{i,j} = -\epsilon \log\left\{(n-1)\sqrt{\hat{q}_i} W_{i,j} \sqrt{\hat{q}_j}\right\}. \tag{3.6}$$

Therefore, the corrected distances $\hat{D}_{i,j}$ correspond to the similarities measured by the affinity matrix $\sqrt{\hat{q}_i} W_{i,j} \sqrt{\hat{q}_j}$, which approximates the clean Gaussian kernel (up to a global constant) according to Theorem 3.3 and the results in section 2.

We now have the following result, whose proof can be found in supplement section SM5.

Proposition 3.4. *Under Assumptions 2.1–3.1, there exist $\epsilon_0, t_0, m_0(\epsilon), n_0(\epsilon), C'(\epsilon) > 0$, such that for all $\epsilon < \epsilon_0$, $m > m_0(\epsilon)$, $n > n_0(\epsilon)$, we have*

$$\hat{N}_i = \|\eta_i\|_2^2 + \epsilon \frac{d \log(s)}{4(s-1)} + \mathcal{O}(\epsilon^2) + \mathcal{E}_i^{(2)}, \tag{3.7}$$

$$\hat{S}_i = \|x_i\|_2^2 - \epsilon \frac{d \log(s)}{4(s-1)} + \mathcal{O}(\epsilon^2) + \mathcal{E}_i^{(3)}, \tag{3.8}$$

$$\hat{D}_{i,j} = \|x_i - x_j\|_2^2 - \epsilon \frac{d \log(s)}{2(s-1)} + \mathcal{O}(\epsilon^2) + \mathcal{E}_{i,j}^{(4)}, \tag{3.9}$$

for all $i, j = 1, \ldots, n$, $i \neq j$, where $\max_i |\mathcal{E}_i^{(2)}|$, $\max_i |\mathcal{E}_i^{(3)}|$, and $\max_{i,j} |\mathcal{E}_{i,j}^{(4)}|$ are upper bounded by the right-hand side of (2.4) with probability at least $1 - n^{-t}$ for any $t > t_0$.

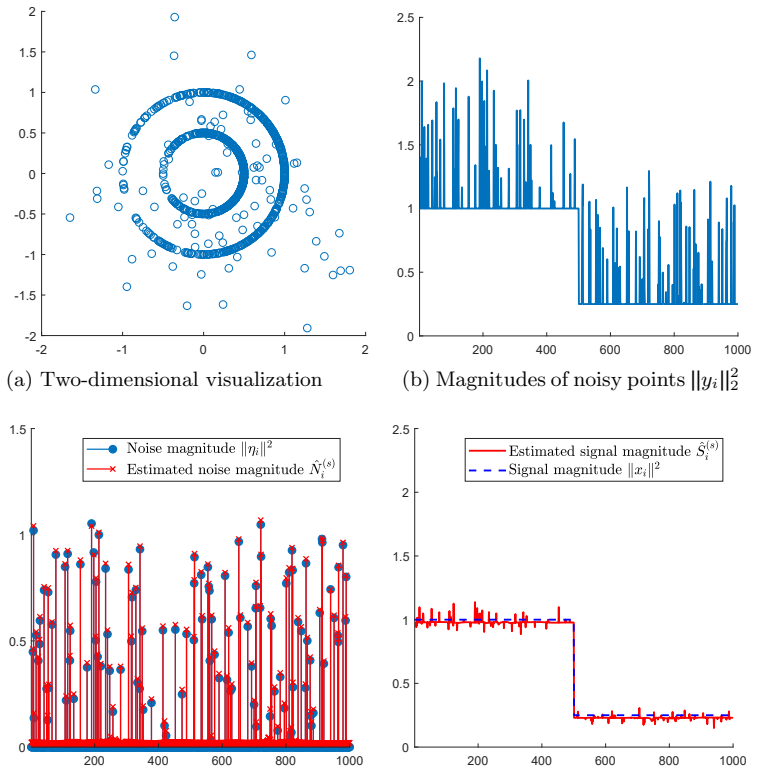
Proposition 3.4 asserts that for sufficiently large m, n and sufficiently small ϵ, the quantities $\hat{N}_i$, $\hat{S}_i$, and $\hat{D}_{i,j}$ can approximate $\|\eta_i\|_2^2$, $\|x_i\|_2^2$, and $\|x_i - x_j\|_2^2$, respectively, up to arbitrarily small errors with high probability. According to (3.7), (3.8), and (3.9), the first error term in these approximations is a global constant that depends explicitly on d, s, and ϵ, and thus can be removed if the intrinsic dimension d is known or can be estimated. Alternatively, if one is only interested in ranking $\|\eta_i\|_2^2$, $\|x_i\|_2^2$, or $\|x_i - x_j\|_2^2$, then the relevant bias error term is improved to $\mathcal{O}(\epsilon^2)$ since ranking is unaffected by a global additive constant. For example, this is the case if one is interested in identifying the points with the largest or smallest noise

magnitudes, or determining the nearest neighbors of each point x_i. The variance error terms $\mathcal{E}^{(2)}$, $\mathcal{E}^{(3)}$, $\mathcal{E}^{(4)}$ have the same behavior as $\mathcal{E}$ from Theorem 2.4 in section 2.

Note that the noise magnitude estimator $\hat{N}_i$ in (3.4) corrects for the effect of the variability of the density $q(x_i)$ on the scaling factors $\mathbf{d}$. However, one does not have to use $\hat{q}_i$ in (3.4) and it can be replaced with the constant 1. In such a case, we would still have an $\mathcal{O}(\epsilon)$ bias error term in each of (3.7), (3.8), and (3.9), but it would depend on $q(x_i)$ (and $q(x_j)$ in the case of (3.9)). Hence, the $\mathcal{O}(\epsilon)$ term would no longer be a global constant that does not influence ranking. Consequently, the main advantage of accounting for the density is to improve the effective bias error term from $\mathcal{O}(\epsilon)$ to $\mathcal{O}(\epsilon^2)$ under ranking.

We begin by demonstrating $\hat{N}_i$ and $\hat{S}_i$ via a toy example. We generated two centered circles in $\mathbb{R}^2$, one with radius 1 and the other with radius 0.5. We independently sampled 500 points from the first circle and 500 from the second circle according to the same (nonuniform) density used for Figure 1a. We then embedded all points in $\mathbb{R}^m$ with $m = 500$ by applying a random orthogonal transformation, and added i.i.d. outlier-type noise η_i taken to be zero with probability 0.9 and sampled from a multivariate normal with covariance $\sigma_i I_m/m$ with probability 0.1, where σ_i is sampled uniformly from $(0, 1)$; see Figure 5a for a two-dimensional visualization of this setup. Figure 5b illustrates the noisy point magnitudes $\|y_i\|_2^2$, where the signal magnitudes are clearly intertwined with the noise. Figure 5c illustrates the noise magnitude estimator $\hat{N}_i$ from (3.4) with $s = 2$ and $\epsilon = 0.1$, for each index $i = 1, \ldots, 1000$. It is evident that $\hat{N}_i$ accurately infers the true noise magnitudes $\|\eta_i\|_2^2$, albeit a small upward shift due to the bias term $\epsilon d \log s/(4(s-1))$ from (3.7). Importantly, $\hat{N}_i$ is invariant to the density $q(x_i)$ and the signal magnitudes $\|x_i\|_2^2$. Similarly, Figure 5d shows that $\hat{S}_i$ accurately recovers $\|x_i\|_2^2$ up to a small global shift and minor fluctuations. Overall, the doubly stochastic scaling allows us to decompose $\|y_i\|_2^2$ into signal and noise parts. In particular, $\hat{N}_i$ can be utilized to identify the noisy points in this setting, while $\hat{S}_i$ reveals that the clean data points can be partitioned into two groups with distinct magnitudes.

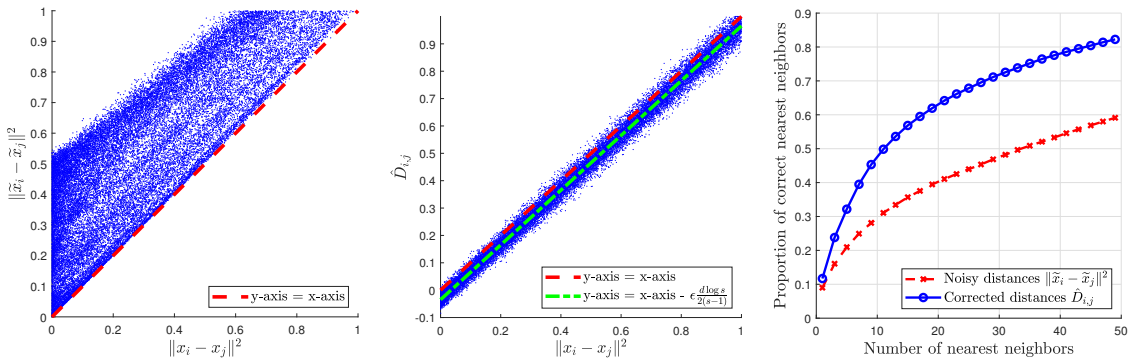
Next, we demonstrate the advantage of correcting the noisy Euclidean distances $\|y_i - y_j\|_2^2$ via $\hat{D}$ from (3.9). Figure 6a shows the noisy distances $\|y_i - y_j\|_2^2$ versus the clean distances $\|x_i - x_j\|_2^2$ in the setup of Figures 1b and 2a, where points are sampled from a circle and corrupted by noise with magnitude that is varying smoothly from 0.01 to 0.5 (see more details in the text of section 3.1). It is evident that the Euclidean distances are strongly corrupted by the variability of the noise magnitude and deviate substantially from the desired behavior, which is the dashed diagonal line (describing perfect correspondence). In Figure 6b we depict the corrected distances $\hat{D}_{i,j}$ computed with $s = 2$ and $\epsilon = 0.1$, which are much closer to the clean distances and concentrate well around a line with slope 1 that is shifted slightly below the desired trend. This shift agrees almost perfectly with the bias term $-\epsilon d \log(s)/(2(s-1)) \approx -0.035$ appearing in (3.9). Lastly, for each $k = 1, \ldots, 50$, we computed the k nearest neighbors of each x_i according to the noisy distances $\|y_i - y_j\|_2^2$ and the corrected distances $\hat{D}_{i,j}$. For each k, Figure 6c shows the proportion of these nearest neighbors that coincide with any of the k true nearest neighbors according to the clean distances $\|x_i - x_j\|_2^2$, averaged over $i = 1, \ldots, 1000$. It is clear that the corrected distances allow for more accurate identification of near neighbors, with more than 80% accuracy for $k = 50$, while the noisy distances provide less than 60% accuracy in that case. Note that the nearest neighbors of each point x_i according to the corrected distances $\hat{D}_{i,j}$ correspond to the largest entries of $\sqrt{\hat{q}_i} W_{i,j} \sqrt{\hat{q}_j}$ in each row



(a) Two-dimensional visualization (b) Magnitudes of noisy points $\|y_i\|_2^2$

(c) Estimated vs. true noise magnitude (d) Estimated vs. true signal magnitude

Figure 5. *Example of estimating the noise magnitudes* $\|\eta_i\|_2^2$ *and signal magnitudes* $\|x_i\|_2^2$ *using* $\hat{N}_i$ *and* $\hat{S}_i$ *from* (3.4) *and* (3.9)*, respectively (panels* (c) *and* (d)*), with* $s = 2$*,* $\epsilon = 0.1$*, for* $i = 1, \ldots, 1000$*. The noisy observations* y_i *are sampled from two concentric circles, each with the nonuniform density of Figure* 1a*, embedded in dimension* $m = 500$ *and corrupted with probability* 0.1 *by multivariate normal noise with covariance* $\sigma_i I_m$*, where* σ_i *is sampled uniformly at random from* $(0, 1)$ *(panels* (a) *and* (b)*).*



(a) Noisy vs. clean distances (b) Corrected vs. clean distances (c) Near neighbor detection

Figure 6. *Influence of noise on pairwise Euclidean distances and on detection accuracy of the nearest neighbors, compared to the corrected distances* $\hat{D}_{i,j}$ *from* (3.9) *with* $s = 2$ *and* $\epsilon = 0.1$*. The data is sampled from a unit circle with nonuniform density, embedded in dimension* $m = 1000$ *and corrupted by noise whose magnitude varies smoothly from* 0.01 *to* 0.5*; see Figures* 1b *and* 2a *and relevant text.*

i. Hence, Figure 6c also describes the advantage of the affinity matrix $\sqrt{\hat{q}_i}W_{i,j}\sqrt{\hat{q}_j}$ over the noisy Gaussian kernel $K_{i,j}$ for encoding similarities between data points.

3.3. Robust weighted manifold Laplacian approximation. In what follows we construct a family of normalizations that is a robust analogue of (1.2) and establish convergence to the associated family of differential operators (see [17]).

Fix $\alpha \in [0,1]$, and define

$$L^{(\alpha)} = \frac{4(I_n - \hat{W}^{(\alpha)})}{\epsilon}, \qquad \hat{W}^{(\alpha)}_{i,j} = \frac{\widetilde{W}^{(\alpha)}_{i,j}}{\sum_{j=1}^{n}\widetilde{W}^{(\alpha)}_{i,j}}, \qquad \widetilde{W}^{(\alpha)}_{i,j} = \frac{W_{i,j}}{[\hat{q}_i\hat{q}_j]^{\alpha-1/2}} \tag{3.10}$$

for all $i,j = 1,\dots,n$, where I_n is the $n \times n$ identity matrix, and $L^{(\alpha)}$ is an appropriately normalized graph Laplacian for $\hat{W}^{(\alpha)}$. The formulas in (3.10) are equivalent to those in (1.2) except that we utilize the robust density estimator $\hat{q}_i$ instead of the standard KDE and further account for the asymptotic approximation of W in (2.8) (leading to the power $\alpha - 0.5$ in the denominator of $\widetilde{W}^{(\alpha)}$ instead of the power α appearing in $P^{(\alpha)}$ of (1.2)). Note that when $\alpha = 0.5$, no normalization of W is actually performed since $\hat{W}^{(0.5)}_{i,j} = W_{i,j}$.

Next, we define the Schrodinger-type differential operator

$$\{T^{(\alpha)}f\}(x) = \frac{\Delta_{\mathcal{M}}\{fq^{1-\alpha}\}(x)}{[q(x)]^{1-\alpha}} - \frac{\Delta_{\mathcal{M}}\{q^{1-\alpha}\}(x)}{[q(x)]^{1-\alpha}}f(x) \tag{3.11}$$

for any $f \in \mathcal{C}^2(\mathcal{M})$, where $\Delta_{\mathcal{M}}$ is the negative Laplace–Beltrami operator on $\mathcal{M}$. If the sampling density is uniform, i.e., $q(x)$ is a constant function, then $T^{(\alpha)}$ reduces to $\Delta_{\mathcal{M}}$ for any α. Otherwise, $T^{(\alpha)}$ depends on the density $q(x)$, except for the special case of $\alpha = 1$, where the density vanishes and $T^{(\alpha)}$ again becomes $\Delta_{\mathcal{M}}$. When $\alpha = 0.5$, $T^{(\alpha)}$ is the Fokker–Planck operator describing Brownian motion via the Langevin equation [51], and when $\alpha = 0$, $T^{(\alpha)}$ describes the limiting operator of the popular random walk graph Laplacian.

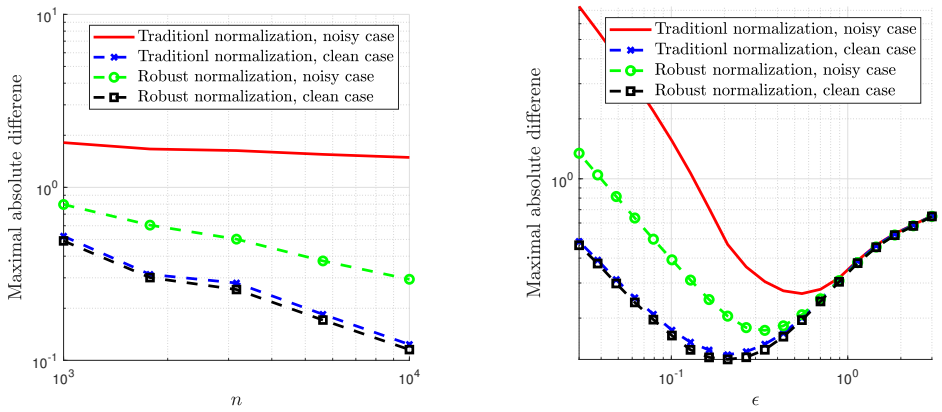
We now have the following result, whose proof can be found in supplement section SM5.

Theorem 3.5. *Fix $f \in \mathcal{C}^3(\mathcal{M})$. Under Assumptions 2.1–3.1, there exist $\epsilon_0, t_0, m_0(\epsilon), n_0(\epsilon)$, $C'(\epsilon) > 0$, such that for all $\epsilon < \epsilon_0$, $m > m_0(\epsilon)$, $n > n_0(\epsilon)$, we have*

$$\sum_{j=1}^{n} L^{(\alpha)}_{i,j} f(x_j) = \{T^{(\alpha)}f\}(x_i) + \mathcal{O}(\epsilon^{\beta}) + \mathcal{E}^{(5)}_i \tag{3.12}$$

for all $i = 1,\dots,n$, where $\max_i |\mathcal{E}^{(5)}_i|$ is upper bounded by the right-hand side of (2.4) with probability at least $1 - n^{-t}$ for any $t > t_0$.

Theorem 3.5 shows that for sufficiently large m,n and sufficiently small ϵ, the matrix $L^{(\alpha)}$ can approximate the action of the operator $T^{(\alpha)}$ pointwise up to an arbitrarily small error with high probability. If $\alpha = 1$, then $L^{(\alpha)}$ approximates the (negative) Laplace–Beltrami operator $\Delta_{\mathcal{M}}$, which encodes the intrinsic geometry of the manifold [17] regardless of the sampling density. If $\alpha = 0.5$, then $L^{(\alpha)} = 4(I_n - W)/\epsilon$ approximates the Fokker–Planck operator, suggesting that the doubly stochastic Markov matrix W simulates Langevin diffusion on $\mathcal{M}$, agreeing with the results in [48, 79, 14]. If $\alpha = 0$, we have $\widetilde{W}^{(\alpha)}_{i,j} = \sqrt{\hat{q}_i}W_{i,j}\sqrt{\hat{q}_j}$, which



(a) Error versus n (b) Error versus ϵ

Figure 7. *Maximal absolute difference between the Laplace–Beltrami operator* $\{\Delta_{\mathcal{M}} f\}(x_i)$ *and its approximation* $\sum_{j=1}^{n} L_{i,j}^{(\alpha)} f(x_j)$ *using the robust graph Laplacian normalization* (3.10) *with* $\alpha = 1$ *and its traditional counterpart, versus the sample size* n *and the bandwidth* ϵ*. The data is sampled according to the setting used for Figures* 1a*,* 1b*, and* 2a*, namely a unit circle with nonuniform density where the points are either clean or corrupted by smoothly varying noise in high dimension.*

corrects for the influence of density on W according to (2.8), thereby approximating the clean Gaussian kernel up to a global constant. In this case, $L^{(\alpha)}$ approximates the same operator as the standard random walk graph Laplacian on the clean data. Theorem 3.5 shows that the popular family of normalizations (1.2) can be made robust to general high-dimensional noise via the doubly stochastic affinity matrix W and our robust density estimator (3.1).

In Figure 7 we demonstrate the advantage of the robust graph Laplacian normalization (3.10) for $\alpha = 1$ over the traditional normalization $4(I_n - \hat{P}^{(1)})/\epsilon$, where $\hat{P}^{(1)}$ is from (1.2). We used the same setting as the one used for Figures 1a and 1b, namely the unit circle with nonuniform density, where the sampled points are either clean or corrupted by smoothly varying noise; see more details in the corresponding text of section 3.1. To quantify the accuracy of the approximation in (3.12), we used the test function $f(\theta(x)) = (\cos(\theta(x)) + \sin(2\theta(x)))/5$, where $\theta(x) \in [0, 2\pi)$ is the angle of a point x on the circle. For $\alpha = 1$, $T^{(\alpha)}$ reduces to the Laplace–Beltrami operator $\Delta_{\mathcal{M}}$, in which case $\{\Delta_{\mathcal{M}}(f)\}(x_i) = (-\cos(\theta_i) - 4\sin(2\theta_i))/5$, where $\theta_i = \theta(x_i)$. We then computed the maximal absolute difference between $\sum_{j=1}^{n} L_{i,j}^{(\alpha)} f(x_j)$ and $\{\Delta_{\mathcal{M}}(f)\}(x_i)$ over $i = 1, \ldots, n$, for both our robust normalization as well as the traditional one. Figure 7a shows these errors versus the sample size n, where $\epsilon = 0.1$, $s = 2$, $m = n$, and we averaged the errors over 30 randomized trials. Figure 7a shows the same errors versus the bandwidth parameter ϵ, where $m = n = 5000$, $s = 2$, and we averaged the results over 20 randomized trials. It is evident that the robust and the traditional graph Laplacian normalizations perform nearly identically in the clean case, both across n and across ϵ, suggesting that the bias and variance errors in (3.12) match those for the traditional normalization, at least in our setting. On the other hand, the robust normalization performs much better in the noisy case whenever the error is dominated by the variance term, i.e., when ϵ is sufficiently small with respect to the sample size n, while having almost identical behavior when the error is dominated by the bias term. Consequently, the robust normalization achieves

smaller errors for any fixed bandwidth in this scenario and allows us to use a smaller optimally tuned bandwidth to obtain a better approximation of the Laplace–Beltrami operator under noise.

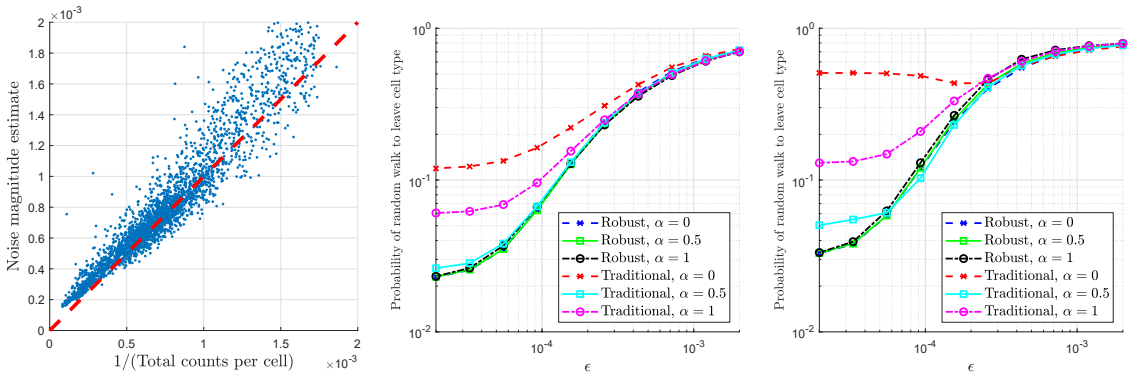
4. Experiments on real single-cell RNA-sequencing data. In this section, we demonstrate our results using real data from single-cell RNA-sequencing (scRNA-seq), which is a revolutionary technology for measuring high-dimensional gene expression profiles of individual cells in diverse populations [70, 46]. In this case, each observation $\widetilde{y}_i \in \mathbb{R}^m$ is a vector of nonnegative integers describing the expression levels of m different genes in the ith cell of the sample. The high resolution of the data—given at the single-cell level—makes it possible to study the similarities between different cells and to characterize different cell populations, which is of paramount importance in immunology and developmental biology. However, one of the main challenges in analyzing scRNA-seq data is the high levels of noise and its nonuniform nature [37, 38, 39].

To demonstrate our results, we used the popular dataset of purified peripheral blood mononuclear cells (PBMC) by [84], where 32733 genes are sequenced over 94654 cells that are annotated experimentally according to 10 known cell types. To preprocess the data, we first randomly subsampled 500 cells from each of the following types: b cells, cd14 monocytes, cd34, cd4 helper, cd56 nk, and cytotoxic t. These cell types are fairly distinguishable one from another, thereby simplifying the interpretability of our subsequent results, while the subsampling makes computations more tractable. We then computed the total expression count for each cell, given by $c_i = \sum_{j=1}^m \widetilde{y}_{i,j}$, where $\widetilde{y}_{i,j}$ denotes the jth entry of $\widetilde{y}_i$, and computed the normalized observations $y_i = \widetilde{y}_i/c_i$. This is a standard normalization in scRNA-seq for making the cell descriptors to be probability vectors, thereby removing the influence of technical variability of counts (also known as "read depth") across cell populations [75, 20]. The doubly stochastic scaling of the Gaussian kernel is then evaluated using the normalized observations $y_1, \ldots, y_n$, $n = 3000$, with a prescribed tolerance of 10^{-6} and a maximum of 10^4 iterations in the algorithm of [79].

In our first experiment, we set out to investigate the accuracy of noise magnitude estimator $\hat{N}_i$ described in (3.4). To validate our noise estimates, we assume the popular Poisson data model $\widetilde{y}_{i,j} \sim \text{Poisson}(\mu_{i,j})$ [59]. In this case, we have $\mathbb{E}[\widetilde{y}_{i,j}] = \text{Var}[\widetilde{y}_{i,j}] = \mu_{i,j}$, and by standard concentration arguments,

$$\|\eta_i\|_2^2 \sim \frac{\sum_{j=1}^m (\widetilde{y}_{i,j} - \mu_{i,j})^2}{c_i^2} \sim \frac{\sum_{j=1}^m \text{Var}[\widetilde{y}_{i,j}]}{c_i^2} = \frac{\sum_{j=1}^m \mu_{i,j}}{c_i^2} \sim \frac{\sum_{j=1}^m \widetilde{y}_{i,j}}{c_i^2} = \frac{1}{c_i}, \tag{4.1}$$

asymptotically as $m \to \infty$ under appropriate delocalization conditions on the Poisson parameters $\{\mu_{i,j}\}_{j=1}^m$. Therefore, we expect the noise magnitude $\|\eta_i\|_2^2$ to be close to $1/c_i$, which is the inverse of the total gene expression counts for cell i. Figure 8a depicts the estimated noise magnitudes $\hat{N}_i$ computed with $\epsilon = 2 \cdot 10^{-5}$ and $s = 2$, versus $1/c_i$ for a prototypical subsampled dataset (with 3000 cells total, 500 from each of six different types). Evidently, the Poisson model suggests that the noise magnitude fluctuates considerably across the data, roughly by an order of magnitude. Of course, the noise magnitude estimator $\hat{N}_i$ is completely



(a) Estimated noise vs Poisson noise

(b) Average transition probability error

(c) Worst-case transition probability error

Figure 8. *Accuracy of the noise magnitude estimate* $\hat{N}_i$ *from* (3.4) *(panel* (a)*) and the Markov matrix* $\hat{W}^{(\alpha)}$ *from* (3.10) *as well as its traditional analogue* $\hat{P}^{(\alpha)}$ *from* (1.2) *(panels* (b) *and* (c)*), evaluated using the annotated single-cell RNA-sequencing data from* [84]*. The noise magnitude estimate* $\hat{N}_i$ *is compared to the anticipated value from the Poisson model* (4.1)*, while the error in the Markov matrix* $\hat{W}^{(\alpha)}$ *and its traditional analogue is assessed by the average probability to transition between distinct cell types (the smaller the better), plotted against the bandwidth parameter* ϵ*.*

oblivious to the Poisson model and does not have access to the total counts c_i (as it is determined solely from the normalized observations). Nonetheless, the estimated noise magnitudes $\hat{N}_i$ concentrate around the red dashed diagonal line, showcasing good agreement with the Poisson model. Note that there seems to be a slight quadratic trend to the estimated noise magnitudes, which is in line with literature suggesting over-dispersion with respect to the standard Poisson [68] (e.g., negative binomial).

In our second experiment, we employ the cell type annotations to assess the accuracy of $\hat{W}^{(\alpha)}_{i,j}$ from (3.10) and its traditional counterpart $\hat{P}^{(\alpha)}$ from (1.2). Since $\hat{W}^{(\alpha)}_{i,j}$ is a transition probability matrix, it describes a random walk over the cells. It is reasonable to assume that a random walk that starts at a certain cell should be unlikely to immediately transition to a cell with a different type. Motivated by the this reasoning, we show in Figure 8b the probability of a cell transitioning to a cell with a different type, averaged over all cells $i = 1, \ldots, 3000$, according to $\hat{W}^{\alpha}_{i,j}$ from (3.10) as well as its traditional counterpart, where we used $s = 2$, $\alpha = 0, 0.5, 1$, and averaged the results over 20 randomized trials (of subsampling cells), plotted against the bandwidth parameter ϵ. Figure 8c is the same as Figure 8b except that we averaged the aforementioned probabilities over the cells in each cell type separately and took the largest of these, namely the worst-case averaged transition probability over the six cell types.

From Figures 8b and 8c it is evident that for large bandwidth parameters ϵ, all normalizations provide similarly undesirable behavior in the form of large probabilities of transition errors, namely probabilities of transitioning between different cell types. As we decrease ϵ, the behavior generally improves across all normalizations, but the errors due to the robust normalizations $\hat{W}^{(\alpha)}$ are consistently smaller than the traditional ones. This advantage of our normalizations is particularly evident over the traditional normalization with $\alpha = 0$, whose worst-case error (for one of the cell types) exceeds 0.4 for all values of ϵ. The traditional

normalization with $\alpha = 0.5$ seems to be more accurate than $\alpha = 0$ or $\alpha = 1$ and provides results very similar to the robust normalizations, albeit slightly larger worst-case errors for small ϵ. Recall that the traditional normalization with $\alpha = 0.5$ is obtained by first performing the symmetric normalization $D_i^{-1/2} W_{i,j} D_i^{-1/2}$, where $D_i = \sum_{j=1}^n K_{i,j}$, and then performing a row-stochastic normalization. These steps are precisely one iteration of the accelerated scaling algorithm described in [79]. This fact can possibly explain the advantage of $\alpha = 0.5$ over the other values of α.

Note that all robust normalizations provide nearly identical probabilities of transition errors, which may initially seem strange in light of the different limiting operators for the corresponding graph Laplacians in Theorem 3.5. However, an important distinction is that the graph Laplacian $L^{(\alpha)}$ describes the first-order behavior $W^{(\alpha)}$ in ϵ, while its zeroth-order behavior is given by $\sum_{j=1}^n \hat{W}_{i,j}^{(\alpha)} f(x_j) \sim f(x_i)$ for large m, n and small ϵ, regardless of α. Hence, we should expect the random walk transition probability errors to be very similar across α. Indeed, this is the case for the robust normalization. On the other hand, the random walk transition probability errors for the traditional normalization differ substantially across α, which is likely due to the strong variability of the noise in the data (see Figure 8a) and the sensitivity of the standard kernel density estimator to such noise.

5. Discussion. The results in this work give rise to several future research directions. On the practical side, to make the tools we developed in section 3 widely applicable, it is desirable to derive procedures for adaptively tuning the bandwidth parameter ϵ and the parameter s in the DS-KDE (3.1). Moreover, for large experimental datasets, the density can vary considerably across the sample space, in which case a global bandwidth parameter is unlikely to provide a satisfactory bias-variance trade-off. Hence, it is worthwhile to tune the bandwidth according to the local density around each point. Techniques for adaptive bandwidth selection have been extensively studied for standard kernel density estimation and traditional graph Laplacian normalizations in the clean case (see [83, 7] and references therein), e.g., by near neighbor distances. However, the adaptation of these techniques to our setting is nontrivial, as the near neighbor distances could be too corrupted for determining the local bandwidth. Therefore, this topic requires substantial analytical and empirical investigation that is left for future work.

On the theoretical side, one important direction is to characterize the constant $C'(\epsilon)$ appearing in (2.4) in terms of ϵ. This would require a more advanced analysis of the stability of the scaling factors $\mathbf{d}$ under perturbations in $\mathbf{W}$ and the prescribed row sums, which is beyond the scope of this work. In addition, we conjecture that the results in Theorem 2.6 can be strengthened, and specifically that $\rho_\epsilon(x) = q^{-1/2}(x) F_\epsilon(x) + \mathcal{O}(\epsilon^2)$ uniformly over $x \in \mathcal{M}$. Currently, Theorem 2.6 only proves an analogous L^p bound for $p \in [1, 4/3)$, while the pointwise bound in Theorem 2.6 is only for $d \leq 5$ and is dominated by a dimension-dependent error that is worse than $\mathcal{O}(\epsilon^2)$. A useful first step in that direction might be to obtain a tighter characterization of ρ_ϵ than in our Lemma SM1.3. However, this will require different theoretical tools and is left for future work. Lastly, properly refined versions of Theorems 2.4 and 2.6 can be combined to describe how to tune the bandwidth parameter ϵ for convergence of the approximation errors described in sections 3.1 and 3.3.

REFERENCES

[1] J. Ah-Pine, *Learning doubly stochastic and nearly idempotent affinity matrix for graph-based clustering*, European J. Oper. Res., 299 (2022), pp. 1069–1078.

[2] Z. Allen-Zhu, Y. Li, R. Oliveira, and A. Wigderson, *Much faster algorithms for matrix scaling*, in 2017 IEEE 58th Annual Symposium on Foundations of Computer Science (FOCS), IEEE, 2017, pp. 890–901.

[3] R. B. Bapat and T. Raghavan, *Nonnegative Matrices and Applications*, Encyclopedia Math. Appl. 64, Cambridge University Press, Cambridge, UK, 1997.

[4] W. T. Baxter, R. A. Grassucci, H. Gao, and J. Frank, *Determination of signal-to-noise ratios and spectral SNRs in cryo-em low-dose imaging of molecules*, J. Struct. Biol., 166 (2009), pp. 126–132.

[5] M. Beauchemin, *On affinity matrix normalization for graph cuts and spectral clustering*, Pattern Recognition Lett., 68 (2015), pp. 90–96.

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

[7] T. Berry and J. Harlim, *Variable bandwidth diffusion kernels*, Appl. Comput. Harmon. Anal., 40 (2016), pp. 68–96.

[8] J. M. Borwein, A. S. Lewis, and R. D. Nussbaum, *Entropy minimization, dad problems, and doubly stochastic kernels*, J. Funct. Anal., 123 (1994), pp. 264–307.

[9] P. Brennecke, S. Anders, J. K. Kim, A. A. Kołodziejczyk, X. Zhang, V. Proserpio, B. Baying, V. Benes, S. A. Teichmann, J. C.Marioni, and M. G. Heisler, *Accounting for technical noise in single-cell RNA-seq experiments*, Nat. Methods, 10 (2013), pp. 1093–1095.

[10] M. M. Bronstein, J. Bruna, Y. LeCun, A. Szlam, and P. Vandergheynst, *Geometric deep learning: Going beyond Euclidean data*, IEEE Signal Process. Mag., 34 (2017), pp. 18–42.

[11] A. Buades, B. Coll, and J.-M. Morel, *A non-local algorithm for image denoising*, in 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05), Vol. 2, IEEE, 2005, pp. 60–65.

[12] J. Calder and N. G. Trillos, *Improved Spectral Convergence Rates for Graph Laplacians on Epsilon-Graphs and k-NN Graphs*, preprint, arXiv:1910.13476, 2019.

[13] M. Chen, M. Gong, and X. Li, *Robust doubly stochastic graph clustering*, Neurocomputing, 475 (2022), pp. 15–25.

[14] X. Cheng and B. Landa, *Bi-stochastically Normalized Graph Laplacian: Convergence to Manifold Laplacian and Robustness to Outlier Noise*, preprint, arXiv:2206.11386, 2022.

[15] X. Cheng and N. Wu, *Eigen-convergence of Gaussian Kernelized Graph Laplacian by Manifold Heat Interpolation*, preprint, arXiv:2101.09875, 2021.

[16] R. N. Cochran and F. H. Horne, *Statistically weighted principal component analysis of rapid scanning wavelength kinetics experiments*, Anal. Chem., 49 (1977), pp. 846–853.

[17] R. R. Coifman and S. Lafon, *Diffusion maps*, Appl. Comput. Harmon. Anal., 21 (2006), pp. 5–30.

[18] R. R. Coifman and M. Maggioni, *Diffusion wavelets*, Appl. Comput. Harmon. Anal., 21 (2006), pp. 53–94.

[19] R. R. Coifman, N. F. Marshall, and S. Steinerberger, *A common variable minimax theorem for graphs*, Found. Comput. Math., 23 (2023), pp. 493–517.

[20] M. B. Cole, D. Risso, A. Wagner, D. DeTomaso, J. Ngai, E. Purdom, S. Dudoit, and N. Yosef, *Performance assessment and selection of normalization procedures for single-cell rna-seq*, Cell Syst., 8 (2019), pp. 315–328.

[21] M. Cuturi, *Sinkhorn distances: Lightspeed computation of optimal transport*, in Advances in Neural Information Processing Systems, 2013, pp. 2292–2300.

[22] M. Defferrard, X. Bresson, and P. Vandergheynst, *Convolutional neural networks on graphs with fast localized spectral filtering*, in Advances in Neural Information Processing Systems, 2016, pp. 3844–3852.

[23] A. Douik and B. Hassibi, *A Riemannian approach for graph-based clustering by doubly stochastic matrices*, in 2018 IEEE Statistical Signal Processing Workshop (SSP), IEEE, 2018, pp. 806–810.

[24] D. B. Dunson, H.-T. Wu, and N. Wu, *Spectral convergence of graph Laplacian and heat kernel reconstruction in L^∞ from random samples*, Appl. Comput. Harmon. Anal., 55 (2021), pp. 282–336.

[25] N. El Karoui, *On information plus noise kernel random matrices*, Ann. Statist., 38 (2010), pp. 3191–3216.
[26] N. El Karoui and H.-T. Wu, *Graph connection Laplacian methods can be made robust to noise*, Ann. Statist., 44 (2016), pp. 346–372.
[27] A. Foi, *Clipped noisy images: Heteroskedastic modeling and practical denoising*, Signal Process., 89 (2009), pp. 2609–2629.
[28] S. Fortunato, *Community detection in graphs*, Phys. Rep., 486 (2010), pp. 75–174.
[29] A. Grigor'yan, *Heat kernels on weighted manifolds and applications*, Cont. Math., 398 (2006), pp. 93–191.
[30] C. Hafemeister and R. Satija, *Normalization and variance stabilization of single-cell RNA-seq data using regularized negative binomial regression*, Genome Biol., 20 (2019), pp. 1–15.
[31] D. K. Hammond, P. Vandergheynst, and R. Gribonval, *Wavelets on graphs via spectral graph theory*, Appl. Comput. Harmon. Anal., 30 (2011), pp. 129–150.
[32] M. Hein, J.-Y. Audibert, and U. Von Luxburg, *From graphs to manifolds—weak and strong pointwise consistency of graph Laplacians*, in International Conference on Computational Learning Theory, Springer, New York, 2005, pp. 470–485.
[33] R. Henderson, *Avoiding the pitfalls of single particle cryo-electron microscopy: Einstein from noise*, Proc. Natl. Acad. Sci. USA, 110 (2013), pp. 18037–18041.
[34] M. Idel, *A Review of Matrix Scaling and Sinkhorn's Normal Form for Matrices and Positive Maps*, preprint, arXiv:1609.06349, 2016.
[35] C. Jia, Y. Hu, D. Kelly, J. Kim, M. Li, and N. R. Zhang, *Accounting for technical noise in differential expression analysis of single-cell RNA sequencing data*, Nucleic Acids Res., 45 (2017), pp. 10978–10988.
[36] T. M. Johanson, H. D. Coughlan, A. T. Lun, N. G. Bediaga, G. Naselli, A. L. Garnham, L. C. Harrison, G. K. Smyth, and R. S. Allan, *Genome-wide analysis reveals no evidence of trans chromosomal regulation of mammalian immune development*, Plos Genet., 14 (2018), e1007431.
[37] P. V. Kharchenko, *The triumphs and limitations of computational methods for scRNA-seq*, Nat. Methods, 18 (2021), pp. 723–732.
[38] J. K. Kim, A. A. Kolodziejczyk, T. Ilicic, S. A. Teichmann, and J. C. Marioni, *Characterizing noise structure in single-cell RNA-seq distinguishes genuine from technical stochastic allelic expression*, Nat. Commun., 6 (2015), pp. 1–9.
[39] T. H. Kim, X. Zhou, and M. Chen, *Demystifying "drop-outs" in single-cell umi data*, Genome Biol., 21 (2020), pp. 1–19.
[40] P. A. Knight, D. Ruiz, and B. Uçar, *A symmetry preserving algorithm for matrix scaling*, SIAM J. Matrix Anal. Appl., 35 (2014), pp. 931–955, https://doi.org/10.1137/110825753.
[41] P. Knopp and R. Sinkhorn, *A note concerning simultaneous integral equations*, Canadian J. Math., 20 (1968), pp. 855–861.
[42] B. Landa, R. R. Coifman, and Y. Kluger, *Doubly stochastic normalization of the Gaussian kernel is robust to heteroskedastic noise*, SIAM J. Math. Data Sci., 3 (2021), pp. 388–413, https://doi.org/10.1137/20M1342124.
[43] B. Landa and Y. Shkolnisky, *The steerable graph Laplacian and its application to filtering image datasets*, SIAM J. Imaging Sci., 11 (2018), pp. 2254–2304, https://doi.org/10.1137/18M1169394.
[44] D. Lim, R. Vidal, and B. D. Haeffele, *Doubly Stochastic Subspace Clustering*, preprint, arXiv:2011.14859, 2020.
[45] S. Lohani, A. H. Moberly, H. Benisty, B. Landa, M. Jing, Y. Li, M. J. Higley, and J. A. Cardin, *Dual color mesoscopic imaging reveals spatiotemporally heterogeneous coordination of cholinergic and neocortical activity*, BioRXiv, 2020.
[46] E. Z. Macosko, A. Basu, R. Satija, J. Nemesh, K. Shekhar, M. Goldman, I. Tirosh, A. R. Bialas, N. Kamitaki, E. M. Martersteck, et al., *Highly parallel genome-wide expression profiling of individual cells using nanoliter droplets*, Cell, 161 (2015), pp. 1202–1214.
[47] W. Q. Malik, J. Schummers, M. Sur, and E. N. Brown, *Denoising two-photon calcium imaging data*, PloS one, 6 (2011), e20490.
[48] N. F. Marshall and R. R. Coifman, *Manifold learning with bi-stochastic kernels*, IMA J. Appl. Math., 84 (2019), pp. 455–482.

[49] F. G. Meyer and X. Shen, *Perturbation of the eigenvectors of the graph Laplacian: Application to image denoising*, Appl. Comput. Harmon. Anal., 36 (2014), pp. 326–334.
[50] P. Milanfar, *Symmetrizing smoothing filters*, SIAM J. Imaging Sci., 6 (2013), pp. 263–284, https://doi.org/10.1137/120875843.
[51] B. Nadler, S. Lafon, R. R. Coifman, and I. G. Kevrekidis, *Diffusion maps, spectral clustering and reaction coordinates of dynamical systems*, Appl. Comput. Harmon. Anal., 21 (2006), pp. 113–127.
[52] A. Y. Ng, M. I. Jordan, and Y. Weiss, *On spectral clustering: Analysis and an algorithm*, in Advances in Neural Information Processing Systems, 2002, pp. 849–856.
[53] J. Pang and G. Cheung, *Graph Laplacian regularization for image denoising: Analysis in the continuous domain*, IEEE Trans. Image Process., 26 (2017), pp. 1770–1785.
[54] E. Parzen, *On estimation of a probability density function and mode*, Ann. Math. Statist., 33 (1962), pp. 1065–1076.
[55] G. Peyré, M. Cuturi, et al., *Computational optimal transport: With applications to data science*, Found. Trends® Mach. Learn., 11 (2019), pp. 355–607.
[56] M. Rosenblatt, *Remarks on some nonparametric estimates of a density function*, Ann. Math. Statist., 27 (1956), pp. 832–837.
[57] A. Rupasinghe, N. Francis, J. Liu, Z. Bowen, P. O. Kanold, and B. Babadi, *Direct extraction of signal and noise correlations from two-photon calcium imaging of ensemble neuronal activity*, Elife, 10 (2021), e68046.
[58] J. Salmon, Z. Harmany, C.-A. Deledalle, and R. Willett, *Poisson noise reduction with non-local PCA*, J. Math. Imaging Vis., 48 (2014), pp. 279–294.
[59] A. Sarkar and M. Stephens, *Separating measurement and expression models clarifies confusion in single-cell RNA sequencing analysis*, Nat. Genet., 53 (2021), pp. 770–777.
[60] P. Sarkar and P. J. Bickel, *Role of normalization in spectral clustering for stochastic blockmodels*, Ann. Stat., 43 (2015), pp. 962–990.
[61] S. H. Scheres, *A Bayesian view on cryo-em structure determination*, J. Mol. Biol., 415 (2012), pp. 406–418.
[62] H. Shen and J. Z. Huang, *Analysis of call centre arrival data using singular value decomposition*, Appl. Stoch. Model. Bus. Ind., 21 (2005), pp. 251–263.
[63] J. Shi and J. Malik, *Normalized cuts and image segmentation*, IEEE Trans. Pattern Anal. Mach. Intell., 22 (2000), pp. 888–905.
[64] D. I. Shuman, S. K. Narang, P. Frossard, A. Ortega, and P. Vandergheynst, *The emerging field of signal processing on graphs: Extending high-dimensional data analysis to networks and other irregular domains*, IEEE Signal Process. Mag., 30 (2013), pp. 83–98.
[65] A. Singer, *From graph to manifold Laplacian: The convergence rate*, Appl. Comput. Harmon. Anal., 21 (2006), pp. 128–134.
[66] A. Singer, Y. Shkolnisky, and B. Nadler, *Diffusion interpretation of nonlocal neighborhood filters for signal denoising*, SIAM J. Imaging Sci., 2 (2009), pp. 118–139, https://doi.org/10.1137/070712146.
[67] R. Sinkhorn and P. Knopp, *Concerning nonnegative matrices and doubly stochastic matrices*, Pacific J. Math., 21 (1967), pp. 343–348.
[68] V. Svensson, *Droplet scRNA-seq is not zero-inflated*, Nat. Biotechnol., 38 (2020), pp. 147–150.
[69] O. Tamuz, T. Mazeh, and S. Zucker, *Correcting systematic effects in a large set of photometric light curves*, Mon. Notices Royal Astron. Soc., 356 (2005), pp. 1466–1470.
[70] F. Tang, C. Barbacioru, Y. Wang, E. Nordman, C. Lee, N. Xu, X. Wang, J. Bodeau, B. B. Tuch, A. Siddiqui, K. Lao, and M. Azim Surani, *mRNA-Seq whole-transcriptome analysis of a single cell*, Nat. Methods, 6 (2009), pp. 377–382.
[71] N. G. Trillos, M. Gerlach, M. Hein, and D. Slepčev, *Error estimates for spectral convergence of the graph Laplacian on random geometric graphs toward the Laplace-Beltrami operator*, Found. Comput. Math., 20 (2020), pp. 827–887.
[72] N. G. Trillos, F. Hoffmann, and B. Hosseini, *Geometric structure of graph Laplacian embeddings*, J. Mach. Learn. Res., 22 (2021), pp. 63–61.
[73] L. Van der Maaten and G. Hinton, *Visualizing data using t-SNE*, J. Mach. Learn. Res., 9 (2008), pp. 2579–2605.

[74] R. Vershynin, *High-Dimensional Probability: An Introduction with Applications in Data Science*, Cambridge Ser. Statist. Probab. Math. 47, Cambridge University Press, Cambridge, UK, 2018.
[75] B. Vieth, S. Parekh, C. Ziegenhain, W. Enard, and I. Hellmann, *A systematic evaluation of single cell RNA-seq analysis pipelines*, Nat. Commun., 10 (2019), pp. 1–11.
[76] U. Von Luxburg, *A tutorial on spectral clustering*, Stat. Comput., 17 (2007), pp. 395–416.
[77] H. M. Wallach, *Topic modeling: Beyond bag-of-words*, in Proceedings of the 23rd International Conference on Machine Learning, 2006, pp. 977–984.
[78] F. Wang, P. Li, A. C. König, and M. Wan, *Improving clustering by learning a bi-stochastic data similarity matrix*, Knowl. Inf. Syst., 32 (2012), pp. 351–382.
[79] C. L. Wormell and S. Reich, *Spectral convergence of diffusion maps: Improved error bounds and an alternative normalization*, SIAM J. Numer. Anal., 59 (2021), pp. 1687–1734, https://doi.org/10.1137/20M1344093.
[80] H.-T. Wu and N. Wu, *Strong uniform consistency with rates for kernel density estimators with general kernels on manifolds*, Inf. Inference J. IMA, 11 (2022), pp. 781–799.
[81] R. Zass and A. Shashua, *A unifying approach to hard and probabilistic clustering*, in Tenth IEEE International Conference on Computer Vision (ICCV'05), Vol. 1, IEEE, 2005, pp. 294–301.
[82] R. Zass and A. Shashua, *Doubly stochastic normalization for spectral clustering*, in Advances in Neural Information Processing Systems, 2007, pp. 1569–1576.
[83] L. Zelnik-Manor and P. Perona, *Self-tuning spectral clustering*, in Advances in Neural Information Processing Systems, 2005, pp. 1601–1608.
[84] G. X. Zheng, J. M. Terry, P. Belgrader, P. Ryvkin, Z. W. Bent, R. Wilson, S. B. Ziraldo, T. D. Wheeler, G. P. McDermott, J. Zhu, et al., *Massively parallel digital transcriptional profiling of single cells*, Nat. Commun., 8 (2017), 14049.