The Curse Revisited: When are Distances Informative for the Ground Truth in Noisy High-Dimensional Data?

Robin Vandaele[1,2,3]   **Bo Kang**[3]   **Tijl De Bie**[3]   **Yvan Saeys**[1,2]

[1]Department of Applied Mathematics, Computer Science and Statistics, Ghent University, Gent, Belgium
[2]Data mining and Modelling for Biomedicine (DaMBi), VIB Inflammation Research Center, Gent, Belgium
[3]IDLab, Department of Electronics and Information Systems, Ghent University, Gent, Belgium

Abstract

Distances between data points are widely used in machine learning applications. Yet, when corrupted by noise, these distances—and thus the models based upon them—may lose their usefulness in high dimensions. Indeed, the small marginal effects of the noise may then accumulate quickly, shifting empirical closest and furthest neighbors away from the ground truth. In this paper, we exactly characterize such effects in noisy high-dimensional data using an asymptotic probabilistic expression. Previously, it has been argued that neighborhood queries become meaningless and unstable when distance concentration occurs, which means that there is a poor relative discrimination between the furthest and closest neighbors in the data. However, we conclude that this is not necessarily the case when we decompose the data in a ground truth—which we aim to recover—and noise component. More specifically, we derive that under particular conditions, empirical neighborhood relations affected by noise are still likely to be truthful even when distance concentration occurs. We also include thorough empirical verification of our results, as well as interesting experiments in which our derived 'phase shift' where neighbors become random or not turns out to be identical to the phase shift where common dimensionality reduction methods perform poorly or well for recovering low-dimensional reconstructions of high-dimensional data with dense noise.

1 Introduction

Motivation The notorious *curse of dimensionality* encompasses various phenomena that occur in high-dimensional data, which complicate their analysis [Indyk and Motwani, 1998, Beyer et al., 1999, Aggarwal et al., 2001, Verleysen and François, 2005, Kuo and Sloan, 2005, Radovanović et al., 2009]. In the particular case of distance functions such as Euclidean, there tends to be little contrast in the distances between different pairs of points. This phenomenon is known as *distance concentration*, and impedes learning and inference from the data through (local and global) neighborhood-based approaches. "In other words, virtually every data point is then as good as any other, and slight perturbations to the query point would result in another data point being chosen as the nearest neighbor" [Beyer et al., 1999]. Therefore, distance concentration in data is commonly regarded as indicative for the distances between points to be meaningless and the empirical neighborhood relations to be unstable.

Notwithstanding the high emphasis on 'distance concentration' in the current literature, *there is an entirely different yet natural possible view on 'when distances are meaningful' when the data is corrupted by noise*, which we formalize in this paper. We assume the common practical case that the observed data is composed of a *ground truth* component $\boldsymbol{X}$, and a *dense noise* component $\boldsymbol{N}$. By 'dense', we mean that each entry of $\boldsymbol{X}$ is likely corrupted by a small nonzero error value. We regard distances as meaningful when the (closest, furthest, k-nearest, . . .) neighborhood relations derived from the observed data $\boldsymbol{X} + \boldsymbol{N}$ likely coincide with those that would have been obtained from $\boldsymbol{X}$, and thus, are informative for the ground truth model underlying the data. Since noise is unavoidable in many real world data due to practical problems in the collection and preparation processes [Zhu and Wu, 2004], we argue that this is a highly natural way to characterize meaningfulness of distances in such data. For example, biological data such as single cell sequenc-

ing data [Zhang et al., 2021] is inherently noisy, due to the imprecise nature of biological experiments [Libralon et al., 2009, Vandaele et al., 2021]. Other high-dimensional examples include noisy images [Buades et al., 2005], climate time series [Ertöz et al., 2003], and neuron activity data [Friedman et al., 2015].

While noise may contribute little to individual dimensions, its overall contribution can be especially harmful to learning from data when it is high-dimensional. When the data dimensionality grows, and *the signal, here: 'the absolute difference between ground truth distances in neighborhood queries'*, cannot cope with the dense noise that is accumulated at the same time, the data will lose its discriminative power for inferring the ground truth. This is formally explored in this paper.

Furthermore, dimensionality reduction methods are commonly used to alleviate the effect of noise on high-dimensional data and facilitate learning. See for example Figure 1, where the distances are much more useful for (topological) inference from a biological cell trajectory data set after a PCA projection. Dimensionality reductions are also used to obtain more meaningful distances in applications such as spectral clustering [Liu and Han, 2004], and even prior to other embedding methods such as t-SNE [Van der Maaten and Hinton, 2008]. This suggests the need of a formal exploration of how dimensionality reductions themselves are susceptible to noise in high dimensions, which we provide in this paper. While we only study this empirically for synthetic examples within the limited scope of this paper, the fact that *our derived 'phase shift' where neighbors become non-random is identical to the phase shift where common dimensionality reductions methods start performing well*, encourages further theoretical and methodological research into this subject.

Note that we will not introduce any novel algorithms in this paper. Yet, we do present and validate novel theoretical results about learning from high-dimensional point cloud data. These results complement previous work on distance concentration, and add to the understanding of counter-intuitive phenomena of high-dimensional data. We argue that *such improved understanding is imperative for the design of better computational methods for the analysis of such data.*

Finally, we emphasize that *the role of distance concentration in this paper is to be interpreted as rather subtle.* In fact, this paper could be (and starting from Section 2 is for a large part) written independent from distance concentration. However, given its prevalence in related work, we found it important to include distance concentration in our motivation and discussion, and point out its differences to the view on 'meaningfulness in distances' analyzed in this paper.
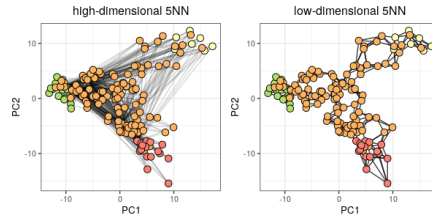

Figure 1: 5NN graphs (edges in black) of a real bifurcating biological cell trajectory (gene expression) data set $\boldsymbol{X} + \boldsymbol{N} \subseteq \mathbb{R}^{1770}$ consisting of four cell groups [Cannoodt et al., 2018, Saelens et al., 2019, Vandaele et al., 2020]. Each dimension quantifies the expression of a particular gene, and the coloring of each point (cell) corresponds to its cell group. The 5NN graphs are visualized through a 2-dimensional PCA embedding of $\boldsymbol{X} + \boldsymbol{N}$. (Left) The edges of the 5NN graph are obtained directly from the distances between the high-dimensional points. (Right) The edges of the 5NN graph are obtained from the 2-dimensional PCA embedding. The possible placement of points is much more constrained in two dimensions. This reduces unwanted behavior caused by noisy high-dimensional distances that impede trajectory inference, such as interconnections between different branches. The resulting lower-dimensional representation will be much more effective for learning the bifurcating model, as common in trajectory inference [Saelens et al., 2019].

Related Work It is well known that many distance measures lose their usefulness for discriminating between neighbors in high-dimensional data. This phenomenon, generally known as distance concentration, has been studied extensively on both a theoretical and experimental level [Beyer et al., 1999, Aggarwal et al., 2001, Durrant and Kabán, 2009, Kabán, 2012, Giannella, 2021]. Its occurrence in data is widely perceived as an indicator that distances between observations are meaningless and neighborhood queries are unstable.

However, we argue that for the abundance of real word data with noise, *'distances are meaningful when they are informative for the ground truth'* is a natural point of view. Nevertheless, to the best of our knowledge, a formal probabilistic analysis of such characterization is lacking, as even extensive studies on the behavior of distances and neighbors in high-dimensional data [Angiulli, 2017] do not include an analysis that explicitly separates a ground truth from a noise component. By maintaining this strategy in this paper however, we make important conclusions that add to the understanding of high-dimensional data. These include that our proposed view on meaningful distances cannot be characterized through distance concentration, and that neighborhood relations may still remain truthful even when distances are dominated by noise.

Contributions

- We provide a probabilistic quantification of the effect of high-dimensional noise on neighboring relations, deriving conditions under which these relations either become highly random or non-random, independent of the magnitude of noise (Theorem 2.1 & Corollary 2.3).
- We provide thorough empirical verification of our theoretical results, and show that our novel yet natural view on meaningful distances is different from distance concentration (Section 3.1).
- We use hyperharmonic series (Example 2.5) to develop experiments that directly link the performance of dimensionality reductions to the randomness of neighborhood relations (Section 3.2).
- We conclude on how our work provides better understanding of learning from noisy high-dimensional point cloud data (Section 4).

2 Quantifying the Effect of Noise on High-Dimensional Neighbors

In the first part of this section (Section 2.1), we provide a probabilistic quantification of the effect of noise on the absolute discrimination between high-dimensional neighbors (Theorem 2.1), and use this to deduce conditions under which empirical neighbors become either highly random or likely truthful (Corollary 2.3). In particular, it will follow that these conditions are independent of the magnitude of noise in the dimensions. Although we commence the analysis assuming we have three fixed points $\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{z} \in \boldsymbol{X}$ in a given ground truth data set $\boldsymbol{X}$, in Section 2.2 we also discuss how our obtained results can be used to derive more general results for $\boldsymbol{X}$, such as Theorem 2.6.

2.1 When Neighbors Become (non-)Random: a Case Study for Three Points

Our setting in this section will be as follows.

- We are given three sequences (which are to be interpreted as vectors) $\boldsymbol{x} = x_1, x_2, \ldots$, $\boldsymbol{y} = y_1, y_2, \ldots$, and $z = z_1, z_2, \ldots$. These correspond to the ground truth—and thus in practice—non-observed points. For $d \in \mathbb{N}^*$, x_d equals the information captured by the d-th dimension of $\boldsymbol{x}$ (analogous for $\boldsymbol{y}, \boldsymbol{z}$). If the model is explained by a finite number of dimensions, we can still regard $\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{z}$ as infinite sequences by letting $x_d = y_d = z_d$ (e.g. $= 0$) for additional dimensions d.
- Rather than observing the vectors $\boldsymbol{x}, \boldsymbol{y}$, and $\boldsymbol{z}$, we observe $\boldsymbol{x} + \boldsymbol{n_x}, \boldsymbol{y} + \boldsymbol{n_y}$, and $\boldsymbol{z} + \boldsymbol{n_z}$. Here, $\boldsymbol{n_x}$ is a realization of a sequence of random noise variables

$\boldsymbol{n_x} = \mathrm{n}_{x_1}, \mathrm{n}_{x_2}, \ldots$ (analogous for $\boldsymbol{y}, \boldsymbol{z}$). We will assume the random variables $\bigcup_{d \in \mathbb{N}^*} \{\mathrm{n}_{x_d}, \mathrm{n}_{y_d}, \mathrm{n}_{z_d}\}$ to be i.i.d, have finite fourth moment μ_4' (measuring the heaviness of the tail of the noise distribution), and be symmetric. While the former two assumptions will be required by the analysis, the latter simply makes it more convenient. Nevertheless, many common random noise distributions such as uniform and normal, are symmetric.

The following result should be interpreted as follows. We are given a query point $\boldsymbol{x}$ from a ground truth data set $\boldsymbol{X}$, and two candidate neighbors $\boldsymbol{y}$ and $\boldsymbol{z}$ of $\boldsymbol{x}$. We want a formula expressing how likely neighborhood relations between $\boldsymbol{x}, \boldsymbol{y}$, and $\boldsymbol{z}$, such as '$\boldsymbol{x}$ is closer to $\boldsymbol{y}$ than to $\boldsymbol{z}$', are preserved after introducing additive noise $\boldsymbol{N}$, i.e., we observe $\boldsymbol{X} + \boldsymbol{N}$ rather than $\boldsymbol{X}$. This formula should be asymptotically valid, i.e., for a sufficiently high dimensionality d of $\boldsymbol{X}$. Intuitively, the resulting probabilities will be in terms of the true distances $\|\boldsymbol{x} - \boldsymbol{y}\|$ and $\|\boldsymbol{x} - \boldsymbol{z}\|$, and noise characteristics, here σ^2 and μ_4'. Indeed, when there is not much difference between $\|\boldsymbol{x} - \boldsymbol{y}\|$ and $\|\boldsymbol{x} - \boldsymbol{z}\|$ (*the signal*), or when σ^2 and μ_4' are large, we expect it to be more difficult to preserve neighborhood relations. The following formula will then be used to derive subsequent insightful results in the rest of this paper.

Theorem 2.1. *Let* $\boldsymbol{x} = x_1, x_2, \ldots$, $\boldsymbol{y} = y_1, y_2, \ldots$ *and* $\boldsymbol{z} = z_1, z_2, \ldots$ *be three sequences of real numbers. Let* $\mathbf{n_x} = \mathrm{n}_{x_1}, \mathrm{n}_{x_2}, \ldots$, $\mathbf{n_y} = n_{y_1}, n_{y_2}, \ldots$, *and* $\mathbf{n_z} = n_{z_1}, n_{z_2}, \ldots$ *be three sequences of jointly i.i.d. symmetric continuous random variables with variance* σ^2 *and finite 4th moment* μ_4'. *For a sequence* $\boldsymbol{s}$, *denote* $\boldsymbol{s}^{(d)}$ *for the vector composed from its first d elements in order. Finally, let*

$$\Delta_\infty(d) := \max \left\{ \left\| \boldsymbol{x}^{(d)} - \boldsymbol{y}^{(d)} \right\|_\infty, \left\| \boldsymbol{x}^{(d)} - \boldsymbol{z}^{(d)} \right\|_\infty \right\}.$$

If

$$\lim_{d \to \infty} \frac{\Delta_\infty(d)}{\sqrt[16]{d}} = 0, \tag{1}$$

then

$$\left| P \left(\left\| \boldsymbol{x}^{(d)} + \mathbf{n}_{\boldsymbol{x}}^{(d)} - \boldsymbol{y}^{(d)} - \mathbf{n}_{\boldsymbol{y}}^{(d)} \right\| \leq \left\| \boldsymbol{x}^{(d)} + \mathbf{n}_{\boldsymbol{x}}^{(d)} - \boldsymbol{z}^{(d)} - \mathbf{n}_{\boldsymbol{z}}^{(d)} \right\| \right) \right.$$
$$\left. - \Phi \left(\zeta^{(d)} \left(\mu_4', \sigma, \boldsymbol{x}, \boldsymbol{y}, \boldsymbol{z} \right) \right) \right| \overset{d \to \infty}{\longrightarrow} 0,$$

where $\zeta^{(d)} : \left(\mathbb{R}^+ \right)^2 \times \left(\mathbb{R}^d \right)^3 \to \mathbb{R} :$

$$\begin{pmatrix} \mu_4' \\ \sigma \\ \boldsymbol{x} \\ \boldsymbol{y} \\ \boldsymbol{z} \end{pmatrix} \mapsto \frac{\|\boldsymbol{x} - \boldsymbol{z}\|^2 - \|\boldsymbol{x} - \boldsymbol{y}\|^2}{\sqrt{2d(\mu_4' + 3\sigma^4) + 8\sigma^2 \left(\|\boldsymbol{x} - \boldsymbol{y}\|^2 + \|\boldsymbol{x} - \boldsymbol{z}\|^2 - \langle \boldsymbol{x} - \boldsymbol{y}, \boldsymbol{x} - \boldsymbol{z} \rangle \right)}},$$
$$\tag{2}$$

and Φ *is the cumulative distribution function of the standard normal distribution.*

Letting

$$
\begin{aligned}
\mathrm{z}(d) &:= \left\| \mathbf{n}_{\boldsymbol{x}}^{(d)} - \mathbf{n}_{\boldsymbol{y}}^{(d)} + \boldsymbol{x}^{(d)} - \boldsymbol{y}^{(d)} \right\|^2 \\
&\quad - \left\| \mathbf{n}_{\boldsymbol{x}}^{(d)} - \mathbf{n}_{\boldsymbol{z}}^{(d)} + \boldsymbol{x}^{(d)} - \boldsymbol{z}^{(d)} \right\|^2 \\
&= \sum_{i=1}^{d} \underbrace{\left[(\mathrm{n}_{x_i} - \mathrm{n}_{y_i} + x_i - y_i)^2 - (\mathrm{n}_{x_i} - \mathrm{n}_{z_i} + x_i - z_i)^2 \right]}_{=:\mathrm{z}_i},
\end{aligned}
\tag{3}
$$

the proof of Theorem 2.1 is based on an application of the central limit theorem (CLT) to quantify the limiting behavior of $\mathrm{z}(d)$. However, the random variables z_i, $i = 1, \ldots, d$, are not necessarily identically distributed, i.e., with the same mean and variance. For this reason, unlike the analysis by [Beyer et al., 1999, Aggarwal et al., 2001], we require special conditions to ensure that the CLT remains applicable in our result. To this end, (1) provides a sufficient condition for Linderberg's condition to be satisfied [Lindeberg, 1922]. A full proof of Theorem 2.1 is provided in Appendix B.

Remark 2.2. *Following the proof in Appendix B, condition (1) can be further weakened to:*

$$
\left(\frac{\Delta_\infty^4(d)}{d + \left\| \boldsymbol{x}^{(d)} - \boldsymbol{y}^{(d)} \right\| \left\| \boldsymbol{x}^{(d)} - \boldsymbol{z}^{(d)} \right\|} \xrightarrow{d \to \infty} 0 \right) \; and \; (\forall \epsilon > 0)
$$

$$
\left(\frac{F_\mathrm{n} \left(-\left(d + \left\| \boldsymbol{x}^{(d)} - \boldsymbol{y}^{(d)} \right\| \left\| \boldsymbol{x}^{(d)} - \boldsymbol{z}^{(d)} \right\| \right)^{\frac{1}{4}} \epsilon \right)}{\times \min \left\{ d \Delta_\infty^4(d), \left\| \boldsymbol{x}^{(d)} - \boldsymbol{y}^{(d)} \right\|^2 \left\| \boldsymbol{x}^{(d)} - \boldsymbol{z}^{(d)} \right\|^2 \right\}}{d + \left\| \boldsymbol{x}^{(d)} - \boldsymbol{y}^{(d)} \right\| \left\| \boldsymbol{x}^{(d)} - \boldsymbol{z}^{(d)} \right\|} \xrightarrow{d \to \infty} 0 \right),
\tag{4}
$$

where F_n is the cumulative distribution function of the (marginal) random noise variable n. By making use of Markov's inequality, it can be straightforwardly shown that (1) $\implies$ (4). While this condition is less insightful than (1), it can be used to easily show that it suffices that $\Delta_\infty(d) = o\left(d^{\frac{1}{4}} \right)$ in the generic case that n is bounded. Nevertheless, in the practical case that $\Delta_\infty(d)$ is bounded, i.e., when newly added dimensions are (eventually) at most as discriminating as the former, condition (1) is trivially satisfied. In Corollary 2.3 we will assume such bound, as it allows for a convenient way to 'symmetrize' the asymptotic growth conditions formalized in this result.

Under the same setting as for Theorem 2.1, i.e., given a query point $\boldsymbol{x}$ and two candidate (closest, furthest, …) neighbors $\boldsymbol{y}$ and $\boldsymbol{z}$ of $\boldsymbol{x}$, we can now study growth conditions on the signal—this being how well we can discriminate between $\boldsymbol{y}$ and $\boldsymbol{z}$ as the ground truth neighbors of $\boldsymbol{x}$—under which the signal 'beats' the noise in high dimensions and vice versa. If the noise beats the signal, then the empirical neighborhood relations, i.e., those derived after additive noise is introduced, thus from the observed data, will be (nearly) completely random. This is expressed by Corollary 2.3.1 below. In the opposite case, the signal beats the noise, and the empirical neighborhood relations will (likely) agree with those that would have been derived without noise, i.e., from the ground truth points $\boldsymbol{x}$, $\boldsymbol{y}$, and $\boldsymbol{z}$. This is expressed by Corollary 2.3.2. The proofs of these results are provided in Appendix B.

Corollary 2.3. *Let $\boldsymbol{x} = x_1, x_2, \ldots$, $\boldsymbol{y} = y_1, y_2, \ldots$ and $\boldsymbol{z} = z_1, z_2, \ldots$ be three sequences of real numbers. Let $\mathbf{n}_{\boldsymbol{x}} = \mathrm{n}_{x_1}, \mathrm{n}_{x_2}, \ldots$, $\mathbf{n}_{\boldsymbol{y}} = \mathrm{n}_{y_1}, \mathrm{n}_{y_2}, \ldots$, and $\mathbf{n}_{\boldsymbol{z}} = \mathrm{n}_{z_1}, \mathrm{n}_{z_2}, \ldots$ be three sequences of jointly i.i.d. symmetric continuous random variables with finite 4th moment μ_4'. Suppose further that $\sup_{d \in \mathbb{N}^*} \Delta_\infty(d) \leq C$ for some constant C. Then the following two statements are true.*

1. *If $\left\| \boldsymbol{x}^{(d)} - \boldsymbol{z}^{(d)} \right\|^2 - \left\| \boldsymbol{x}^{(d)} - \boldsymbol{y}^{(d)} \right\|^2 = o\left(d^{\frac{1}{2}} \right)$,*

$$
\lim_{d \to \infty} P\left(\left\| \boldsymbol{x}^{(d)} + \mathbf{n}_{\boldsymbol{x}}^{(d)} - \boldsymbol{y}^{(d)} - \mathbf{n}_{\boldsymbol{y}}^{(d)} \right\| \leq \right.
$$

$$
\left. \left\| \boldsymbol{x}^{(d)} + \mathbf{n}_{\boldsymbol{x}}^{(d)} - \boldsymbol{z}^{(d)} - \mathbf{n}_{\boldsymbol{z}}^{(d)} \right\| \right) = \frac{1}{2}.
$$

2. *If $d^{\frac{1}{2}} = o\left(\left\| \boldsymbol{x}^{(d)} - \boldsymbol{z}^{(d)} \right\|^2 - \left\| \boldsymbol{x}^{(d)} - \boldsymbol{y}^{(d)} \right\|^2 \right)$,*

$$
\lim_{d \to \infty} P\left(\left\| \boldsymbol{x}^{(d)} + \mathbf{n}_{\boldsymbol{x}}^{(d)} - \boldsymbol{y}^{(d)} - \mathbf{n}_{\boldsymbol{y}}^{(d)} \right\| \right.
$$

$$
\left. \leq \left\| \boldsymbol{x}^{(d)} + \mathbf{n}_{\boldsymbol{x}}^{(d)} - \boldsymbol{z}^{(d)} - \mathbf{n}_{\boldsymbol{z}}^{(d)} \right\| \right) = 1.
$$

Remark 2.4. *The following conclusions—which will all be validated in Section 3—can now be made.*

1. *$\Delta_2(d) := \max \left\{ \left\| \boldsymbol{x}^{(d)} - \boldsymbol{y}^{(d)} \right\|, \left\| \boldsymbol{x}^{(d)} - \boldsymbol{z}^{(d)} \right\| \right\}$ has to grow at least as $d^{\frac{1}{4}}$, otherwise neighbors will become highly random. Indeed, if $\Delta_2(d) = o\left(d^{\frac{1}{4}} \right)$, then condition (4) in Remark 2.3 is satisfied, and Corollary 2.3.1 becomes applicable (note that the assumed bound $\sup_{d \in \mathbb{N}^*} \Delta_\infty(d) \leq C$ is only required for Corollary 2.3.2, see also Appendix B).*

2. *The noise characteristics σ^2 and μ_4' naturally have a direct effect on the usefulness of empirical neighbors for any fixed dimension d, which is intuitively clear, and can be seen from (2). However, under the conditions of Corollary 2.3, their magnitude becomes negligible in large dimensions.*

3. *Even if the distances between the noise vectors are dominant in the distances between the observed data points, neighbors may become non-random, i.e., representative for the ground truth neighbors. For example, if the random noise variable n is uniformly distributed, then the expected distances between two noise vectors grows as $\sqrt{d}$, whereas it is sufficient for the absolute differences between ground truth distances to grow as $\sqrt[4]{d} + \epsilon$ for some $\epsilon > 0$ to ensure that the noise becomes unlikely to effect neighbors in sufficiently high dimensions.*

The following example will prove to be very useful in the experiments (Section 3).

Example 2.5. *Let $\boldsymbol{x} = \boldsymbol{y}$ be the sequences of all zeros. Given $\alpha \in \mathbb{R}_{>2} \cup \{\infty\}$, we define the sequence of reals $\boldsymbol{z}(\alpha) = z_1(\alpha), z_2(\alpha), \ldots$, by letting for $d \in \mathbb{N}^*$,*

$$z_d(\alpha) := \begin{cases} \frac{1}{\sqrt[\alpha]{d}} & \text{if } \alpha \in \mathbb{R}_{>2}; \\ 1 & \text{if } \alpha = \infty. \end{cases}$$

Observe that we have $\Delta_\infty(d) = \left\| \boldsymbol{z}^{(d)}(\alpha) \right\|_\infty = 1$ for all $d \in \mathbb{N}^*$. We furthermore find that for $\alpha \in \mathbb{R}_{>2}$, $\left\| \boldsymbol{z}^{(d)}(\alpha) \right\|^2$ defines a hyperharmonic series*, which for large $d \in \mathbb{N}^*$ can be approximated as*

$$\left\| \boldsymbol{z}^{(d)}(\alpha) \right\|^2 = \sum_{k=1}^{d} \frac{1}{\sqrt[\alpha]{k^2}} \tag{5}$$
$$\stackrel{d \to \infty}{\sim} \int_1^d x^{-\frac{2}{a}} \mathrm{d}x = \left(\frac{\alpha}{\alpha - 2} \right) \left(d^{1 - \frac{2}{\alpha}} - 1 \right),$$

i.e., $\left\| \boldsymbol{z}^{(d)}(\alpha) \right\|^2$ grows as $d^{1-\frac{2}{\alpha}}$. With by convention $\frac{1}{\infty} = 0$, this holds for $\alpha = \infty$ as well. Naturally, in any dimension $\boldsymbol{y}$ is always closer to $\boldsymbol{x}$ than $\boldsymbol{z}$ is. Due to Corollary 2.3, the probability that this remains true under noise in high dimensions satisfies

$$P\left(\left\| \boldsymbol{x}^{(d)} + \mathbf{n}_{\boldsymbol{x}}^{(d)} - \boldsymbol{y}^{(d)} - \mathbf{n}_{\boldsymbol{y}}^{(d)} \right\| \right.$$
$$\left. \leq \left\| \boldsymbol{x}^{(d)} + \mathbf{n}_{\boldsymbol{x}}^{(d)} - \boldsymbol{z}^{(d)} - \mathbf{n}_{\boldsymbol{z}}^{(d)} \right\| \right)$$
$$\stackrel{d \to \infty}{\longrightarrow} \begin{cases} \frac{1}{2} & \alpha < 4; \\ 1 & \alpha > 4; \\ \Phi\left(\sqrt{\frac{2}{\mu_4' + 3\sigma^4}} \right) & \alpha = 4. \end{cases}$$

The last limit can be found from (5) by adapting the proof of Corollary 2.3.2 in Appendix B. Thus, $\alpha = 4$ corresponds to a 'phase shift', where neighbors transition between becoming random or non-random.

In the following section, we discuss how our results can be used to derive more general results for larger data $\boldsymbol{X}$, consisting of more than three points $\boldsymbol{x}, \boldsymbol{y}$, and $\boldsymbol{z}$.

2.2 Randomness in Neighbors for Data Sets of Arbitrary Sizes

In the previous section we restricted to the particular scenario where we have three given points $\boldsymbol{x}$, $\boldsymbol{y}$, and $\boldsymbol{z}$. Naturally, we can also study the effectiveness of deriving neighborhood relations in a data set $\boldsymbol{X}$ of arbitrary size under the effect of noise in high dimensions. The reason for this is that in practice, we deal with a finite number of data points. Therefore, more general mathematical results on preserving neighbors under noise may often be derived directly from the results presented in Section 2.1.

One such example is as follows, providing sufficient growth conditions on the ground truth diameter for the noise to cause neighboring relations to become random, or thus necessary conditions for the noise not to cause this (which is what we want to achieve in practice). The idea here is that when for every point $\boldsymbol{x} \in \boldsymbol{X}$, if in the empirical noisy data it is completely random whether $\boldsymbol{x}$'s true furthest neighbor becomes closer to $\boldsymbol{x}$ than $\boldsymbol{x}$'s true closest neighbor, i.e., this occurs with probability $\frac{1}{2}$, then one can essentially not work effectively with the high-dimensional neighbors.

Theorem 2.6. *Let $\boldsymbol{X} = \boldsymbol{v}_1, \boldsymbol{v}_2, \ldots$ be a sequence of column vectors, for which $\boldsymbol{v}_d \in \mathbb{R}^n$, $n \in \mathbb{N}^*$, and denote by $\boldsymbol{X}^{(d)}$ the matrix in $\mathbb{R}^{n \times d}$ composed of the first d vectors $\boldsymbol{v}_d$ in order. For $d \in \mathbb{N}^*$, and $i = 1, \ldots, n$, we identify the i-th row of the matrix $\boldsymbol{X}^{(d)}$ with the point $\boldsymbol{x}_i^{(d)} \in \boldsymbol{X}^{(d)}$. Let furthermore $\{\mathbf{n}_i = \mathrm{n}_{i_1}, \mathrm{n}_{i_2}, \ldots\}_{i \in \{1, \ldots, n\}}$ be a collection of jointly i.i.d. symmetric continuous random variables with finite 4th moment. For $d \in \mathbb{N}^*$ and each point $\boldsymbol{x}_i^{(d)} \in \boldsymbol{X}^{(d)}$, let $\boldsymbol{x}_{i,\min}^{(d)}$ denote the closest neighbor of $\boldsymbol{x}_i^{(d)}$ in $\boldsymbol{X}^{(d)}$, $\boldsymbol{x}_{i,\max}^{(d)}$ the furthest neighbor of $\boldsymbol{x}_i^{(d)}$ in $\boldsymbol{X}^{(d)}$, and $\boldsymbol{\Delta_2}(d) := \max_{\boldsymbol{x}, \boldsymbol{y} \in \boldsymbol{X}^{(d)}} \|\boldsymbol{x} - \boldsymbol{y}\|$ the diameter of $\boldsymbol{X}^{(d)}$. If $\boldsymbol{\Delta_2}(d) = o\left(d^{\frac{1}{4}}\right)$, then*

$$\sup_{i=1,\ldots,n} P\left(\left\| \boldsymbol{x}_i^{(d)} + \mathbf{n}_i^{(d)} - \boldsymbol{x}_{i,\max}^{(d)} - \mathbf{n}_{\boldsymbol{x}_{i,\max}^{(d)}}^{(d)} \right\| \right.$$
$$\left. \leq \left\| \boldsymbol{x}_i^{(d)} + \mathbf{n}_i^{(d)} - \boldsymbol{x}_{i,\min}^{(d)} - \mathbf{n}_{\boldsymbol{x}_{i,\min}^{(d)}}^{(d)} \right\| \right) \stackrel{d \to \infty}{\longrightarrow} \frac{1}{2}.$$

Proof. Corollary 2.3.1 is valid for every triple of rows in $\boldsymbol{X}$ (see also Remark 2.4.1). Hence, the result follows from the fact that there are only finitely many such triples (note that the row indices of $\boldsymbol{x}_{i,\min}^{(d)}$ and $\boldsymbol{x}_{i,\max}^{(d)}$ in $\boldsymbol{X}^{(d)}$ are allowed to vary with d). □

Theorem 2.6 essentially states the conditions we must avoid for any practical application that relies on the distances between data observations. Conversely, the conditions we should aim for are generally more dependant on the application of interest. Indeed, for many practical purposes it may not be important to preserve *all* neighborhood relations in the data. For example, as will be validated in Section 3.2, for clustering algorithms a sufficient condition for cluster assignments to be likely truthful would be that Corollary 2.3.2 is satisfied for every triple $(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{z})$, where $\boldsymbol{x}$ and $\boldsymbol{y}$ belong to the same (ground truth) cluster and $\boldsymbol{z}$ to a different cluster. Under the effect of noise, any two points from the same ground truth cluster then likely remain closer to each other than to any point from another cluster. However, for particular methods such as single-linkage clustering, this would be too stringent, and more in-depth analysis will be required. Thus,

we will not claim one such generally applicable result in the current paper. Nevertheless, the consensus remains that to overcome the impact of the extra noise, adding dimensions should be accompanied with adding sufficient information to discriminate between important neighbors according to the ground truth.

3 Experimental Results

In this section we conduct experiments that aim to improve one's understanding and intuition about working with distances in noisy high-dimensional data. First, Section 3.1 will be devoted to empirical validation of our theoretical results. Section 3.2 will be devoted to linking the performance of common dimensionality reduction methods as well as spectral clustering to randomness of neighborhood relations. While our empirical observations in Section 3.2 cannot be immediately derived from our theoretical results in their current stage, they point out direct and interesting connections between the performance of machine learning algorithms and the signal-to-noise ratio as formalized in this paper, which encourage further research into this subject. Code for this project is available on `https://github.com/robinvndaele/NoisyDistances`.

3.1 Validation of the Theoretical Results

Validation of Theorem 2.1 We constructed three sets of three sequences x, y, and z containing up to $10\,000$ dimensions. For each set, we let $x = y = 0$. z is used to control ground truth distance growth rates, here measured through the l_2 and l_∞ norm, as follows.

1. **l_2 bounded, l_∞ bounded:** $z = (1, 0, \ldots, 0)$.
2. **l_2 unbounded, l_∞ bounded:** $z = (1, 1, \ldots, 1)$.
3. **l_2 unbounded, l_∞ unbounded:** $z_d = d^{\frac{1}{4} - 0.01}$.

Then for each sequence and in each dimension we added uniform noise n $\sim$ U$[-0.75, 0.75]$, for which $\sigma^2 = \frac{0.75^2}{3}$ and $\mu'_4 = \frac{0.75^4}{5}$. The setup for this experiment is illustrated by Figure 2. Since n is bounded in each dimension, Theorem 2.1 should be applicable to all three cases (see also Remark 2.2). More precisely, for sufficiently large d we should find that $y(d) := \frac{z(d) - \mu(z(d))}{\sigma(z(d))} \overset{\text{approx.}}{\sim} N(0, 1)$, where $z(d)$ is as defined in (3), and $\mu(z(d))$ is the negative nominator and $\sigma(z(d))$ the denominator of $\zeta^{(d)}(\mu'_4, \sigma, x, y, z)$ in (2). We verified this through 5000 samples of $y(d)$ for each of the three ground truth growth rates and various dimension d chosen from a log-scale (replicating the noise outcomes). We used the Shapiro–Wilk test to assess normality. The results are shown in Figure 3, confirming that Theorem 2.1 is indeed applicable to all of the considered growth rates. Since the Shapiro–Wilk test is developed to asses normality, but
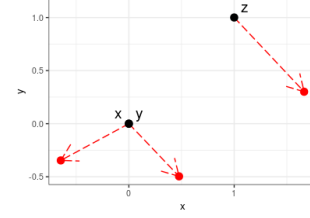

Figure 2: The experiment setup visualized in 2D. We have three points $x = y$ and z, where z controls the true discrimination growth rate. Theorem 2.1 quantifies how likely x will remain closer to y than to z when corrupted by noise in high dimensions, here illustrated by the displacements in red, for various growth rates.
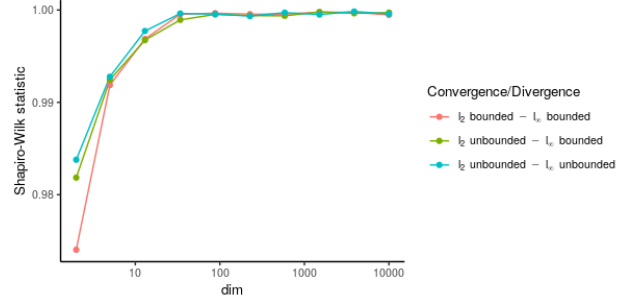

Figure 3: Shapiro–Wilk test statistics to assess normality of y(d) for the various growth rates determined by z, according to the data dimension d. The convergence of the curves to 1 agrees that Theorem 2.1 is applicable to all of the growth rates, and thus allows us to quantify the randomness of neighborhood relations between x, y, and z caused by noise.
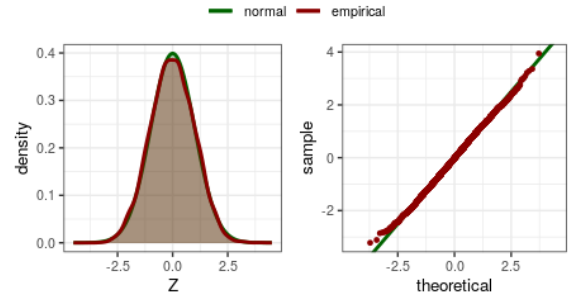

Figure 4: Density plot and Q-Q plot of y($10\,000$) (empirical) compared to the standard normal distribution.

not *standard* normality, Figure 4 shows normality plots of y($10\,000$) for the second set of sequences, confirming the correctness of our calculations.

Validation of Corollary 2.3 We will use Example 2.5 to analyze for which growth rates of the true discrimination between neighbors (the signal), empirical neighbors become random or not. For this, we considered various sets of three sequences x, y, and z, where $x = y = 0$, and $z = z(\alpha)$ controls the growth rate as

determined by α in Example 2.5. We sampled noise using a uniform distribution n $\sim$ U$[-1.25, 1.25]$. A higher magnitude of noise is chosen here to better illustrate that $0 < \Phi_4 := \Phi\left(\sqrt{\frac{2}{\mu_4' + 3\sigma^4}}\right) < 1$ (for $\alpha \neq 4$ the magnitude will not matter in the limit). The setup of this experiment is again visualized by Figure 2. We used 5000 noise replicates to approximate a variety of expected values and probabilities for different growth rates determined by $\alpha \in \{2, 3, 4, 5, 6, \infty\}$. These are illustrated on Figure 5, and defined as follows.

1. The expected distance between two noise vectors, compared to the ground truth diameter growth rates, i.e., of $\|z\|$ (Figure 5, Left).

2. The expected *relative contrast* (Figure 5, Middle) [Aggarwal et al., 2001]:

$$\frac{\max_{p,q \in \{x,y,z\}} \|p + n_p - q - n_q\|}{\min_{p,q \in \{x,y,z\}} \|p + n_p - q - n_q\|} - 1.$$

A relative contrast near 0 indicates distance concentration (discussed in Section 1).

3. The probability (Figure 5, Right)

$$P\left(\|x + n_x - y - n_y\| \leq \|x + n_x - z - n_z\|\right).$$

First, we observe that the distances between the noise vectors is expected to become indefinitely larger than the distances between the ground truth points for $\alpha < \infty$ (Figure 5, Left). Hence, we would intuitively expect the distances between the noise vectors to play a dominant role in the observed empirical distances

for the corresponding growth rates. Second, we observe that also for all considered growth rates determined by $\alpha < \infty$, the expected relative contrast converges to 0 (Figure 5, Middle). This means that in sufficiently high dimensions, the two closest points are expected to be relatively as distant to each other as the two furthest points. If one would interpret this as neighborhood queries to become meaningless and unstable—as argued by [Beyer et al., 1999, Aggarwal et al., 2001]—according to our view discussed in Section 1, this should result in a lot of randomness in the chosen closest neighbor of the noisy observation representing x in the high-dimensional space for all considered $\alpha \in \{2, \ldots, 6\}$. However, as discussed in Example 2.5, this will not be the case whenever $\alpha > 4$, as x will very likely correctly choose y as its neighbor even when this choice is affected by noise in high dimensions. This is confirmed by the empirical probabilities (Figure 5, Right), which furthermore agree with all limits obtained in Example 2.5 from Corollary 2.3.

3.2 Learning with Random Neighbors

Dimensionality Reduction As also discussed in Section 1, dimensionality reductions are commonly applied for preprocessing high-dimensional data that is corrupted by noise. The obtained distances in the lower-dimensional space are then assumed to be more informative for inference and machine learning (Figure 1). This raises the question whether dimensionality reductions can naturally accommodate the effect of noise on high-dimensional neighboring relations.

To investigate this, consider a ground truth data set X of n evenly spaced points on the line segment $\mathcal{L}$ from the origin to $z^{(d)}$ in $\mathbb{R}^d$, where $z = z(\alpha)$ is as defined in Example 2.5 by fixing some $\alpha \in \mathbb{R}_{>2} \cup \{\infty\}$ (Figure 6, Left). Since these points are evenly spaced on $\mathcal{L}$, the growth rate of all squared distances (and the differences between them) will be identical to the growth rate of $\|z^{(d)}\|^2$, up to some constant factor depending on the fixed ground truth ordering of the considered points. Thus, from Corollary 2.3 we find that under the effect of noise in high dimensions, *all* empirical neighbors will become random for $\alpha < 4$, and *all* empirical neighbors will likely remain truthful for $\alpha > 4$.

For a 1D-dimensionality reduction method f, we can now study how well f is able to recover neighboring relations of X from $X + N$, with N a random noise matrix. For this, we look at the correlation between the ordering of points on X and on $f(X + N)$ (Figure 6). Furthermore, we investigate this for six different dimensionality reduction methods that are commonly used for noise or feature size reduction prior to visualization, (topological) inference, or clustering: PCA [Wold et al., 1987, Van der Maaten and Hin-
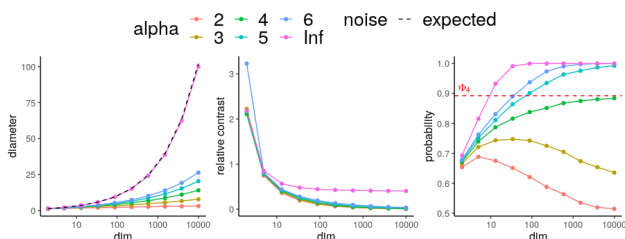

Figure 5: (Left) The expected growth rate of the distance between two noise vectors compared to the ground truth model diameter growth rates determined by α. (Middle) The expected relative contrast converges to 0 for all considered values $\alpha < \infty$, meaning that the two closest points are expected to be relatively as distant to each other as the two furthest points. (Right) The limiting behavior of $P\left(\|x + n_x - y - n_y\| \leq \|x + n_x - z - n_z\|\right)$. Under noise in high dimensions, neighbors will be inferred effectively for $\alpha > 4$, i.e., empirical neighbors will likely be true, whereas for $\alpha < 4$, they become meaningless.

ton, 2008, Street et al., 2018, Cannoodt et al., 2016],
UMAP [McInnes et al., 2018], diffusion maps [Coif-
man and Lafon, 2006, Vandaele et al., 2020, Cannoodt
et al., 2016], robust PCA [Candès et al., 2011] (a
variant of PCA that assumes the data is composed
in a low-rank component $\boldsymbol{X}$ and a *sparse* noise com-
ponent $\boldsymbol{N}$), a basic autoencoder [Kramer, 1991, Vin-
cent et al., 2010] with 5 hidden layers and tanh ac-
tivation, and Isomap [Tenenbaum et al., 2000]. We
evaluated their performances for $n = 25$ points, up
to $d = 10\,000$ dimensions, growth rates determined
by $\alpha \in \{2, 3, 4, 5, 6, \infty\}$, and averaged over 100 noise
replicates from n $\sim$ U$[-1.25, 1.25]$ per dimension. The
autoencoder was built in Python. Other models ran
under standard settings in R (with 10 neighbors in-
stead of 50 for Isomap). Figure 7 shows the results.

We consistently observe that the performances in-
crease by dimension for $\alpha > 4$, and decrease for $\alpha < 4$.
Following the previous results (Figure 5, Right), this
provides empirical evidence that the performance of
these common dimensionality reduction methods is di-
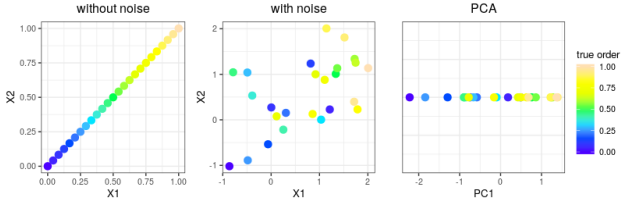rectly affected by whether noise causes randomness in

high-dimensional neighborhood relations or not. This
thus suggests that these methods themselves may be
susceptible to the noise they aim to reduce. These ob-
servations are only contradicted by the autoencoder,
which showed convergence issues for larger dimensions.

Finally, the case $\alpha = 4$ deserves special interest. Since
in a practical setting additional distributional condi-
tions of $\boldsymbol{X}$ will likely result in some non-extreme de-
gree of randomness in the empirical neighborhood re-
lations, we observe that this may be reflected in the
performance of dimensionality reductions as well.

Spectral Clustering Spectral clustering uses the
spectrum of a similarity matrix from the data to per-
form a dimensionality reduction, prior to clustering the
data in fewer dimensions [Filippone et al., 2008]. Natu-
rally, when the dimensionality reduction is affected by
noise, so will the consecutive clustering performance.

To illustrate this, we constructed ground truth clusters
by randomly sampling 25 data points from each of two
10 000-dimensional hyperballs $\mathcal{S}_0 := B\left(\boldsymbol{0}, \frac{\|\boldsymbol{c}(\alpha)\|}{5}\right)$ and
$\mathcal{S}_c := B\left(\boldsymbol{c}(\alpha), \frac{\|\boldsymbol{c}(\alpha)\|}{5}\right)$, where $\boldsymbol{c}(\alpha)$ is defined by

$$c_d(\alpha) = \begin{cases} \frac{5}{\sqrt[\alpha]{d}} & \text{if } d \leq 2; \\ \frac{1}{\sqrt[\alpha]{d}} & \text{if } d > 2. \end{cases}$$

Hence, $\|\boldsymbol{c}^d(\alpha)\|$ grows exactly as $\|\boldsymbol{z}^d(\alpha)\|$ in Example
2.5. The reason that we scale the first two coordinates
is to provide more interpretable visualizations in Fig-
ure 8. Finally, we added noise sampled from the stan-
dard normal distribution $\mathcal{N}(0, 1)$ to each dimension.
We now ask the question how well spectral clustering
is able to recover the ground truth clusters from the
high-dimensional data, as shown in Figure 8.

Figure 6: (Left) $n = 25$ ground truth points, which
make up $\boldsymbol{X}$, are evenly spaced on a linear model
$\mathcal{L} \subseteq \mathbb{R}^d$. (Middle) Rather than observing $\boldsymbol{X}$, we ob-
serve $\boldsymbol{X} + \boldsymbol{N}$ for a random noise matrix $\boldsymbol{N}$. (Right) A
1-dimensional PCA dimensionality reduction aims to
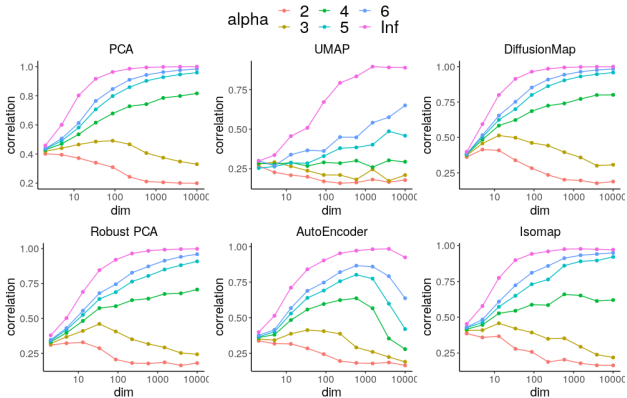retrieve the true ordering of the points on $\mathcal{L}$.

Figure 7: The performance of six common dimen-
sionality reduction methods for recovering the ground
truth ordering of the points on $\mathcal{L}$ under the effect of
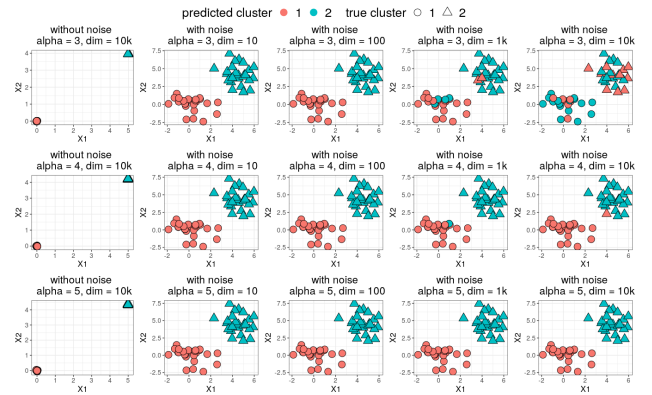noise, by dimension and signal growth rate.

Figure 8: The results of spectral clustering for various
input dimensionalities and values α that control the
separation of two ground truth clusters in high dimen-
sions, with and without Gaussian noise. Each plot is
the restriction of the data to its first two coordinates.

Due to the triangle inequality, any two points in the same ground truth cluster will be closer to each other than to any point in a different cluster, according to the ground truth distances. As a consequence, without noise, spectral clustering (we use the standard settings from the R library Spectrum with a maximum of two clusters) recovers the clusters perfectly from the high-dimensional data (Figure 8, left column). Furthermore, also from the triangle inequality and our results in Section 2.1, it can be shown that when corrupted by noise, points will likely remain closer to points in the same ground truth cluster for $\alpha > 4$, whereas for $\alpha < 4$, points will be nearly equally likely closer to points in other ground truth clusters than their own.

From Figure 8, we see that this directly affects the performance of spectral clustering. For $\alpha = 5$, the clusters are perfectly identified, whereas for $\alpha = 3$, the inferred clusters become increasingly meaningless when the dimensionality of the data from which they are derived grows. Again, $\alpha = 4$ corresponds to a boundary case. Here, the true clusters are identified well, although not perfectly, from the noisy high-dimensional data.

Role of Data Size As common in machine learning applications, we observed that having more data can resolve much of the issues caused by noise. Figure 9 illustrates this for our experiment summarized by Figures 6 & 7, where we now fixed $\alpha = 3$, but varied the data size $n \in \{25, 50, 100\}$. As discussed above, neighboring relations in the high-dimensional data will eventually become random for $\alpha = 3$. For all dimensionality reduction methods, we observe that their performance consistently drops for sufficiently high dimensions (Figure 9). However, for PCA, robust PCA, and diffusion maps, we observe that for larger data sizes, higher performances are reached first, and

the dimension after which they struggle to recover the model gets delayed. For the autoencoder and Isomap, the role of the data size is inconclusive. Interestingly, the performance with UMAP is consistently worse for larger data sizes (with the current settings). These results can vary for the value of α however. For example, in case of the autoencoder model which consistently showed an optimal dimension after which the performance decreases even for $\alpha > 4$ (Figure 7), we observed that larger data sizes may accommodate the noise (see Supplementary Figure 10 in Appendix A).

4 Discussion and Conclusion

Noise can be, but does not have to be, fatal when learning from high-dimensional data based on distances. Although this is not a surprising fact, we provided a first and exact mathematical characterization when such distances become (un)informative under noise. Furthermore, we found that our concept of meaningfulness of distances, i.e., when they are informative for the ground truth, is fundamentally different from distance concentration, and suggests direct connections to the ability of dimensionality reductions to recover the data model. Although we focused on small artificial data sets to validate the results in this (mainly theoretical) paper, they are interesting nevertheless, and encourage further foundational and practical research into learning from noisy high-dimensional data.

Unfortunately, the conditions for distances to be meaningful will be difficult to assess in practice. For example, one can easily derive from our results that when many features are irrelevant to the model, neighborhood relations will become uninformative in the presence of noise. In practice however, we may be unsure whether any features are irrelevant at all. How algorithms may actually benefit from the results presented in this paper, is open to further research. Nevertheless, there is an abundance of high-dimensional data where we cannot effectively recover the structure due to noise, leading to poor subsequent model inference, such as in biological single-cell data analysis. We argue that better understanding the behavior of distances in noisy high-dimensional data—for which we provided, illustrated, and validated theoretical results in this paper—is imperative for one to be able to design better computational methods for their analysis.
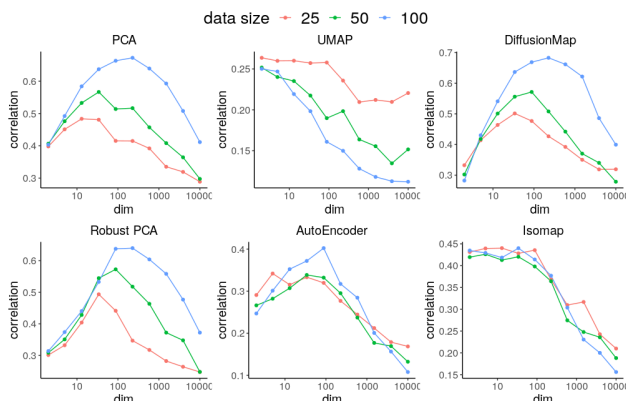

Figure 9: The performance of six common dimensionality reduction methods for recovering the ground truth ordering of the points on $\mathcal{L}$ under the effect of noise, by dimension and data size, for $\alpha = 3$.

Acknowledgments

References

[Aggarwal et al., 2001] Aggarwal, C. C., Hinneburg, A., and Keim, D. A. (2001). On the surprising behavior of distance metrics in high dimensional space. In Van den Bussche, J. and Vianu, V., editors, *Database Theory — ICDT 2001*, pages 420–434, Berlin, Heidelberg. Springer Berlin Heidelberg.

[Angiulli, 2017] Angiulli, F. (2017). On the behavior of intrinsically high-dimensional spaces: Distances, direct and reverse nearest neighbors, and hubness. *J. Mach. Learn. Res.*, 18(1):6209–6268.

[Beyer et al., 1999] Beyer, K., Goldstein, J., Ramakrishnan, R., and Shaft, U. (1999). When is "nearest neighbor" meaningful? In *International conference on database theory*, pages 217–235. Springer.

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

[Candès et al., 2011] Candès, E. J., Li, X., Ma, Y., and Wright, J. (2011). Robust principal component analysis? *Journal of the ACM (JACM)*, 58(3):1–37.

[Cannoodt et al., 2016] Cannoodt, R., Saelens, W., and Saeys, Y. (2016). Computational methods for trajectory inference from single-cell transcriptomics. *European journal of immunology*, 46(11):2496–2506.

[Cannoodt et al., 2018] Cannoodt, R., Saelens, W., Todorov, H., and Saeys, Y. (2018). Single-cell -omics datasets containing a trajectory.

[Coifman and Lafon, 2006] Coifman, R. R. and Lafon, S. (2006). Diffusion maps. *Applied and computational harmonic analysis*, 21(1):5–30.

[Durrant and Kabán, 2009] Durrant, R. J. and Kabán, A. (2009). When is 'nearest neighbour'meaningful: A converse theorem and implications. *Journal of Complexity*, 25(4):385–397.

[Ertöz et al., 2003] Ertöz, L., Steinbach, M., and Kumar, V. (2003). Finding clusters of different sizes, shapes, and densities in noisy, high dimensional data. In *Proceedings of the 2003 SIAM international conference on data mining*, pages 47–58. SIAM.

[Filippone et al., 2008] Filippone, M., Camastra, F., Masulli, F., and Rovetta, S. (2008). A survey of kernel and spectral methods for clustering. *Pattern recognition*, 41(1):176–190.

[Friedman et al., 2015] Friedman, A., Keselman, M. D., Gibb, L. G., and Graybiel, A. M. (2015). A multistage mathematical approach to automated clustering of high-dimensional noisy data. *Proceedings of the National Academy of Sciences*, 112(14):4477–4482.

[Giannella, 2021] Giannella, C. R. (2021). Instability results for euclidean distance, nearest neighbor search on high dimensional gaussian data. *Information Processing Letters*, 169:106115.

[Indyk and Motwani, 1998] Indyk, P. and Motwani, R. (1998). Approximate nearest neighbors: towards removing the curse of dimensionality. In *Proceedings of the thirtieth annual ACM symposium on Theory of computing*, pages 604–613.

[Kabán, 2012] Kabán, A. (2012). Non-parametric detection of meaningless distances in high dimensional data. *Statistics and Computing*, 22(2):375–385.

[Kramer, 1991] Kramer, M. A. (1991). Nonlinear principal component analysis using autoassociative neural networks. *AIChE journal*, 37(2):233–243.

[Kuo and Sloan, 2005] Kuo, F. Y. and Sloan, I. H. (2005). Lifting the curse of dimensionality. *Notices of the AMS*, 52(11):1320–1328.

[Libralon et al., 2009] Libralon, G. L., de Leon Ferreira, A. C. P., Lorena, A. C., et al. (2009). Preprocessing for noise detection in gene expression classification data. *Journal of the Brazilian Computer Society*, 15(1):3–11.

[Lindeberg, 1922] Lindeberg, J. W. (1922). Eine neue herleitung des exponentialgesetzes in der wahrscheinlichkeitsrechnung. *Mathematische Zeitschrift*, 15(1):211–225.

[Liu and Han, 2004] Liu, J. and Han, J. (2004). Spectral clustering. *Advances in neural information processing systems*, 17:1601–1608.

[McInnes et al., 2018] McInnes, L., Healy, J., and Melville, J. (2018). Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*.

[Parzen, 1960] Parzen, E. (1960). *Modern probability theory and its applications*, volume 10. Wiley.

[Radovanović et al., 2009] Radovanović, M., Nanopoulos, A., and Ivanović, M. (2009). Nearest neighbors in high-dimensional data: The emergence and influence of hubs. In *Proceedings of the 26th Annual International Conference on Machine Learning*, ICML '09, pages 865–872, New York, NY, USA. ACM.

[Saelens et al., 2019] Saelens, W., Cannoodt, R., Todorov, H., and Saeys, Y. (2019). A comparison of single-cell trajectory inference methods. *Nature Biotechnology*, 37:1.

[Street et al., 2018] Street, K., Risso, D., Fletcher, R., Das, D., Ngai, J., Yosef, N., Purdom, E., and Dudoit, S. (2018). Slingshot: Cell lineage and pseudo-time inference for single-cell transcriptomics. *BMC Genomics*, 19.

[Tenenbaum et al., 2000] Tenenbaum, J. B., De Silva, V., and Langford, J. C. (2000). A global geometric framework for nonlinear dimensionality reduction. *science*, 290(5500):2319–2323.

[Van der Maaten and Hinton, 2008] Van der Maaten, L. and Hinton, G. (2008). Visualizing data using t-sne. *Journal of machine learning research*, 9(11).

[Vandaele et al., 2021] Vandaele, R., Rieck, B., Saeys, Y., and De Bie, T. (2021). Stable topological signatures for metric trees through graph approximations. *Pattern Recognition Letters*.

[Vandaele et al., 2020] Vandaele, R., Saeys, Y., and Bie, T. D. (2020). Mining topological structure in graphs through forest representations. *Journal of Machine Learning Research*, 21(215):1–68.

[Verleysen and François, 2005] Verleysen, M. and François, D. (2005). The curse of dimensionality in data mining and time series prediction. In *International work-conference on artificial neural networks*, pages 758–770. Springer.

[Vincent et al., 2010] Vincent, P., Larochelle, H., Lajoie, I., Bengio, Y., Manzagol, P.-A., and Bottou, L. (2010). Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion. *Journal of machine learning research*, 11(12).

[Wold et al., 1987] Wold, S., Esbensen, K., and Geladi, P. (1987). Principal component analysis. *Chemometrics and intelligent laboratory systems*, 2(1-3):37–52.

[Zhang et al., 2021] Zhang, R., Atwal, G. S., and Lim, W. K. (2021). Noise regularization removes correlation artifacts in single-cell rna-seq data preprocessing. *Patterns*, 2(3):100211.

[Zhu and Wu, 2004] Zhu, X. and Wu, X. (2004). Class noise vs. attribute noise: A quantitative study. *Artificial intelligence review*, 22(3):177–210.
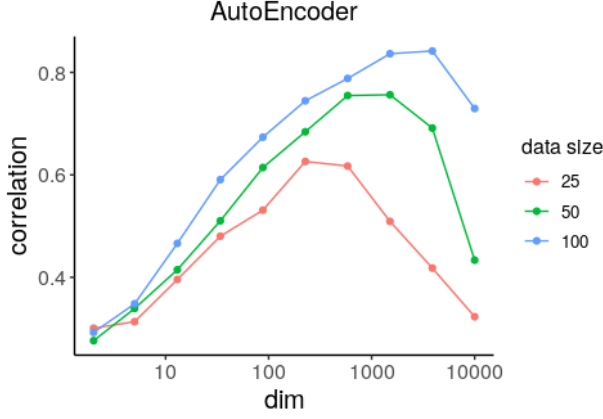
A Supplementary Figures

AutoEncoder

Figure 10: The performance of the autoencoder model for recovering the ground truth ordering of the points on the linear model $\mathcal{L}$ presented in the main paper (see also Figures 6 & 7) under the effect of noise, by dimension and data size, for $\alpha = 5$. Neighboring relations will become increasingly truthful in the noisy high-dimensional data. Nevertheless—for the same hyperparameters and architecture apart from the outer layers (which accommodate the data dimension)—the autoencoder eventually struggles to recover the model from sufficiently large dimensions. However, larger data sizes may provide a temporary solution to this.

B Theorems and Proofs

This part of the appendix contains the mathematical proofs of the results presented in the main paper. The following is the proof of our main theorem that leads to all principal results presented in this paper.

Proof of Theorem 2.1. For $i \in \mathbb{N}^*$, denote $\delta_i(\boldsymbol{x}, \boldsymbol{y}) := x_i - y_i$ and $\delta_i(\boldsymbol{x}, \boldsymbol{z}) := x_i - z_i$. For each $d \in \mathbb{N}^*$, we have

$$
\begin{aligned}
&\mathrm{z}(d) \\
&:= \left\| \mathbf{n}_{\boldsymbol{x}}^{(d)} - \mathbf{n}_{\boldsymbol{y}}^{(d)} + \boldsymbol{x}^{(d)} - \boldsymbol{y}^{(d)} \right\|^2 - \left\| \mathbf{n}_{\boldsymbol{x}}^{(d)} - \mathbf{n}_{\boldsymbol{z}}^{(d)} + \boldsymbol{x}^{(d)} - \boldsymbol{z}^{(d)} \right\|^2 \\
&= \sum_{i=1}^{d} \Bigg[\underbrace{\bigg(\underbrace{\mathrm{n}_{x_i} - \mathrm{n}_{y_i} + \delta_i(\boldsymbol{x}, \boldsymbol{y})}_{=: \mathrm{r}_{x_i, y_i}} \bigg)^2 - \bigg(\underbrace{\mathrm{n}_{x_i} - \mathrm{n}_{z_i} + \delta_i(\boldsymbol{x}, \boldsymbol{z})}_{=: \mathrm{r}_{x_i, z_i}} \bigg)^2 }_{=: \mathrm{z}_i} \Bigg].
\end{aligned}
$$

We have

$$
\mathbb{E}\left(\mathrm{r}_{x_i, y_i} \right) = \delta_i(\boldsymbol{x}, \boldsymbol{y}), \quad \mathrm{Var}\left(\mathrm{r}_{x_i, y_i} \right) = 2\sigma^2,
$$

so that

$$
\mathbb{E}\left(\mathrm{r}_{x_i, y_i}^2 \right) = \mathrm{Var}\left(\mathrm{r}_{x_i, y_i} \right) + \mathbb{E}\left(\mathrm{r}_{x_i, y_i} \right)^2 = 2\sigma^2 + \delta_i^2(\boldsymbol{x}, \boldsymbol{y}),
$$

and thus

$$
\mathbb{E}(\mathrm{z}_i) = \delta_i^2(\boldsymbol{x}, \boldsymbol{y}) - \delta_i^2(\boldsymbol{x}, \boldsymbol{z}).
$$

Since $\mathrm{Cov}(\mathrm{x}, \mathrm{y}) = \mathbb{E}(\mathrm{xy}) - \mathbb{E}(\mathrm{x})\mathbb{E}(\mathrm{y})$ for random variables x, y, by symmetry and the fact that n_{x_i} and n_{y_i} are independent and $\mathbb{E}(\mathrm{n}_{x_i}^{2k+1}) = 0$ for $k \in \mathbb{N}$, we have

$$
\begin{aligned}
\mathrm{Var}\left(\mathrm{r}_{x_i, y_i}^2 \right) &= 2\mathrm{Var}\left(\mathrm{n}_{x_i}^2 \right) - 8\mathrm{Cov}\left(\mathrm{n}_{x_i}^2, \mathrm{n}_{x_i}\mathrm{n}_{y_i} \right) \\
&\quad + 4\mathrm{Var}(\mathrm{n}_{x_i}\mathrm{n}_{y_i}) + 8\delta_i^2(x, y)\mathrm{Var}(\mathrm{n}_{x_i}) \\
&= 2\left(\mu_4' - \mathbb{E}\left(\mathrm{n}_{x_i}^2 \right)^2 \right) + 4\mathbb{E}\left(\mathrm{n}_{x_i}^2 \right)^2 \\
&\quad + 8\sigma^2 \delta_i^2(\boldsymbol{x}, \boldsymbol{y}) \\
&= 2\mu_4' + 2\sigma^4 + 8\sigma^2 \delta_i^2(\boldsymbol{x}, \boldsymbol{y})
\end{aligned}
$$

and analogously

$$
\begin{aligned}
&\mathrm{Cov}\left(\mathrm{r}_{x_i, y_i}^2, \mathrm{r}_{x_i, z_i}^2 \right) \\
&\quad = \mathrm{Var}\left(\mathrm{n}_{x_i}^2 \right) \\
&\quad\quad - 4\mathrm{Cov}\left(\mathrm{n}_{x_i}^2, \mathrm{n}_{x_i}\mathrm{n}_{y_i} \right) \\
&\quad\quad + 2\left(\delta_i(\boldsymbol{x}, \boldsymbol{y}) + \delta_i(\boldsymbol{x}, \boldsymbol{z}) \right) \mathrm{Cov}\left(\mathrm{n}_{x_i}^2, \mathrm{n}_{x_i} \right) \\
&\quad\quad + 4\mathrm{Cov}(\mathrm{n}_{x_i}\mathrm{n}_{y_i}, \mathrm{n}_{x_i}\mathrm{n}_{z_i}) \\
&\quad\quad - 4\left(\delta_i(\boldsymbol{x}, \boldsymbol{y}) + \delta_i(\boldsymbol{x}, \boldsymbol{z}) \right) \mathrm{Cov}(\mathrm{n}_{x_i}\mathrm{n}_{y_i}, \mathrm{n}_{x_i}) \\
&\quad\quad + 4\delta_i(\boldsymbol{x}, \boldsymbol{y})\delta_i(\boldsymbol{x}, \boldsymbol{z})\mathrm{Var}(\mathrm{n}_{x_i}) \\
&\quad = \mu_4' - \mathbb{E}\left(\mathrm{n}_{x_i}^2 \right)^2 + 4\sigma^2 \delta_i(\boldsymbol{x}, \boldsymbol{y})\delta_i(\boldsymbol{x}, \boldsymbol{z}) \\
&\quad = \mu_4' - \sigma^4 + 4\sigma^2 \delta_i(\boldsymbol{x}, \boldsymbol{y})\delta_i(\boldsymbol{x}, \boldsymbol{z}).
\end{aligned}
$$

It thus holds that

$$
\begin{aligned}
&\mathrm{Var}(\mathrm{z}_i) \\
&\quad = \mathrm{Var}\left(\mathrm{r}_{x_i, y_i}^2 \right) + \mathrm{Var}\left(\mathrm{r}_{x_i, z_i}^2 \right) \\
&\quad\quad - 2\mathrm{Cov}\left(\mathrm{r}_{x_i, y_i}^2, \mathrm{r}_{x_i, z_i}^2 \right) \\
&\quad = 2\mu_4' + 6\sigma^4 \\
&\quad\quad + 8\sigma^2 \left(\delta_i^2(\boldsymbol{x}, \boldsymbol{y}) + \delta_i^2(\boldsymbol{x}, \boldsymbol{z}) - \delta_i(\boldsymbol{x}, \boldsymbol{y})\delta_i(\boldsymbol{x}, \boldsymbol{z}) \right).
\end{aligned}
$$

We conclude that

$$
\mu\left(\mathrm{z}(d) \right) = \sum_{i=1}^{d} \mathbb{E}(\mathrm{z}_i) = \left\| \boldsymbol{x}^{(d)} - \boldsymbol{y}^{(d)} \right\|^2 - \left\| \boldsymbol{x}^{(d)} - \boldsymbol{z}^{(d)} \right\|^2,
$$

and

$$
\begin{aligned}
\sigma^2\left(\mathrm{z}(d)\right) &= \sum_{i=1}^{d} \mathrm{Var}(\mathrm{z}_i) \\
&= 2d\left(\mu_4' + 3\sigma^4\right) \\
&\quad + 8\sigma^2\left(\left\|\boldsymbol{x}^{(d)} - \boldsymbol{y}^{(d)}\right\|^2 + \left\|\boldsymbol{x}^{(d)} - \boldsymbol{z}^{(d)}\right\|^2\right. \\
&\qquad\qquad \left. - \left\langle \boldsymbol{x}^{(d)} - \boldsymbol{y}^{(d)}, \boldsymbol{x}^{(d)} - \boldsymbol{z}^{(d)} \right\rangle\right).
\end{aligned}
$$

Also observe that

$$
\begin{aligned}
& \left\|\boldsymbol{x}^{(d)} - \boldsymbol{y}^{(d)}\right\|^2 + \left\|\boldsymbol{x}^{(d)} - \boldsymbol{z}^{(d)}\right\|^2 \\
& \quad - \left\langle \boldsymbol{x}^{(d)} - \boldsymbol{y}^{(d)}, \boldsymbol{x}^{(d)} - \boldsymbol{z}^{(d)} \right\rangle \\
& \geq \left\|\boldsymbol{x}^{(d)} - \boldsymbol{y}^{(d)}\right\|^2 + \left\|\boldsymbol{x}^{(d)} - \boldsymbol{z}^{(d)}\right\|^2 \\
& \quad - \left\|\boldsymbol{x}^{(d)} - \boldsymbol{y}^{(d)}\right\| \left\|\boldsymbol{x}^{(d)} - \boldsymbol{z}^{(d)}\right\| \\
& = \left(\left\|\boldsymbol{x}^{(d)} - \boldsymbol{y}^{(d)}\right\| - \left\|\boldsymbol{x}^{(d)} - \boldsymbol{z}^{(d)}\right\|\right)^2 \\
& \quad + \left\|\boldsymbol{x}^{(d)} - \boldsymbol{y}^{(d)}\right\| \left\|\boldsymbol{x}^{(d)} - \boldsymbol{z}^{(d)}\right\| \\
& \geq \left\|\boldsymbol{x}^{(d)} - \boldsymbol{y}^{(d)}\right\| \left\|\boldsymbol{x}^{(d)} - \boldsymbol{z}^{(d)}\right\|,
\end{aligned}
$$

which is used for Remark 2.2 in the main paper. Now for $\epsilon > 0$, let

$$
\mathbb{A}_{i,\epsilon} := \left\{(\mathrm{n}_{x_i}, \mathrm{n}_{y_i}, \mathrm{n}_{z_i}) : |\mathrm{z}_i - \mathbb{E}(\mathrm{z}_i)| > \epsilon\sigma\left(\mathrm{z}(d)\right)\right\},
$$

and

$$
C_d(\epsilon) := \frac{1}{12}\left(\sqrt{\underbrace{36\Delta_\infty^2(d) + 24\epsilon\sigma\left(\mathrm{z}(d)\right)}_{=:D}} - 6\Delta_\infty(d)\right).
$$

For any $\epsilon > 0$ and $M \geq \sqrt{\frac{6}{\epsilon}}$, by (1), we find that for d sufficiently large

$$
\begin{aligned}
C_d(\epsilon) &\geq \frac{\sqrt{24\epsilon} - \frac{6}{M}}{12}\sqrt{\sigma\left(\mathrm{z}(d)\right)} \tag{6} \\
&\geq \frac{2\sqrt{6\epsilon} - \sqrt{6\epsilon}}{12}\sqrt{\sigma\left(\mathrm{z}(d)\right)} = \frac{\sqrt{6\epsilon}}{12}\sqrt{\sigma\left(\mathrm{z}(d)\right)}, \tag{7}
\end{aligned}
$$

so that in particular $\lim_{d\to\infty} C_d(\epsilon) = +\infty$. If now

$$
\max\{|\mathrm{n}_{x_i}|, |\mathrm{n}_{y_i}|, |\mathrm{n}_{z_i}|\} \leq C_d(\epsilon),
$$

we find that for $i = 1, \ldots, d$,

$$
\begin{aligned}
& |\mathrm{z}_i - \mathbb{E}(\mathrm{z}_i)| \\
& \leq |\mathrm{n}_{y_i}|^2 + |\mathrm{n}_{z_i}|^2 + 2|\mathrm{n}_{x_i}||\mathrm{n}_{y_i}| + 2|\mathrm{n}_{x_i}||\mathrm{n}_{z_i}| \\
& \quad + 2(|\mathrm{n}_{x_i}| + |\mathrm{n}_{y_i}| + |\mathrm{n}_{z_i}|)\Delta_\infty(d) \\
& \leq 6C_d(\epsilon)^2 + 6\Delta_\infty(d)C_d(\epsilon) \\
& \leq \frac{1}{24}\left(72\Delta_\infty^2(d) + 24\epsilon\sigma\left(\mathrm{z}(d)\right) - 12\Delta_\infty(d)\sqrt{D}\right) \\
& \quad + \frac{1}{2}\left(\Delta_\infty(d)\sqrt{D} - 6\Delta_\infty^2(d)\right) \\
& = \epsilon\sigma\left(\mathrm{z}(d)\right).
\end{aligned}
$$

This shows that

$$
\mathbb{A}_{i,\epsilon} \subseteq \{(\mathrm{n}_{x_i}, \mathrm{n}_{y_i}, \mathrm{n}_{z_i}) : \max\{|\mathrm{n}_{x_i}|, |\mathrm{n}_{y_i}|, |\mathrm{n}_{z_i}|\} > C_d(\epsilon)\}
$$
$$
=: \widetilde{\mathbb{A}}_{i,\epsilon}.
$$

Observe that for every $i, k, l \in \mathbb{N}^*$, we have

$$
\begin{aligned}
& \mathbb{E}\left(\mathrm{n}_{x_i}^{2k+1}\mathrm{n}_{y_i}^l \mathbb{1}_{\widetilde{\mathbb{A}}_{i,\epsilon}}\right) \\
& \quad = \mathbb{E}\left(\mathrm{n}_{x_i}^{2k+1}\right)\mathbb{E}\left(\mathrm{n}_{y_i}^l\right) \\
& \qquad - \mathbb{E}\left(\mathrm{n}_{x_i}^{2k+1}\mathbb{1}_{\left(\widetilde{\mathbb{A}}_{i,\epsilon}\right)^c}\right)\mathbb{E}\left(\mathrm{n}_{y_i}^l\mathbb{1}_{\left(\widetilde{\mathbb{A}}_{i,\epsilon}\right)^c}\right) \\
& \quad = 0.
\end{aligned}
$$

Again, due to symmetry, we thus have

$$
\begin{aligned}
& \mathrm{E}\left(\left(\mathrm{r}_{x_i,y_i}^2 - \mathrm{r}_{x_i,z_i}^2\right)\mathbb{1}_{\widetilde{\mathbb{A}}_{i,\epsilon}}\right) \\
& \quad = \left(\delta_i^2(\boldsymbol{x}, \boldsymbol{y}) - \delta_i^2(\boldsymbol{x}, \boldsymbol{z})\right)P\left(\widetilde{\mathbb{A}}_{i,\epsilon}\right).
\end{aligned}
$$

Furthermore, we have

$$
\begin{aligned}
\mathbb{E}\left(\mathrm{r}_{x_i,y_i}^4\mathbb{1}_{\widetilde{\mathbb{A}}_{i,\epsilon}}\right) &= 2\mathbb{E}\left(\mathrm{n}_{x_i}^4\mathbb{1}_{\widetilde{\mathbb{A}}_{i,\epsilon}}\right) + 6\mathbb{E}\left(\mathrm{n}_{x_i}^2\mathrm{n}_{y_i}^2\mathbb{1}_{\widetilde{\mathbb{A}}_{i,\epsilon}}\right) \\
& \quad + 12\delta_i^2(\boldsymbol{x}, \boldsymbol{y})\mathbb{E}\left(\mathrm{n}_{x_i}^2\mathbb{1}_{\widetilde{\mathbb{A}}_{i,\epsilon}}\right) \\
& \quad + \delta_i(\boldsymbol{x}, \boldsymbol{y})^4 P\left(\widetilde{\mathbb{A}}_{i,\epsilon}\right),
\end{aligned}
$$

and

$$
\begin{aligned}
& \mathbb{E}\left(\left(\mathrm{r}_{x_i,y_i}^2\mathrm{r}_{x_i,z_i}^2\right)\mathbb{1}_{\widetilde{\mathbb{A}}_{i,\epsilon}}\right) \\
& \quad = \mathbb{E}\left(\mathrm{n}_{x_i}^4\mathbb{1}_{\widetilde{\mathbb{A}}_{i,\epsilon}}\right) + 3\mathbb{E}\left(\mathrm{n}_{x_i}^2\mathrm{n}_{y_i}^2\mathbb{1}_{\widetilde{\mathbb{A}}_{i,\epsilon}}\right) \\
& \qquad + 2\left(\delta_i(\boldsymbol{x}, \boldsymbol{y}) + \delta_i(\boldsymbol{x}, \boldsymbol{z})\right)^2\mathbb{E}\left(\mathrm{n}_{x_i}^2\mathbb{1}_{\widetilde{\mathbb{A}}_{i,\epsilon}}\right) \\
& \qquad + \delta_i(\boldsymbol{x}, \boldsymbol{y})^2\delta_i(\boldsymbol{x}, \boldsymbol{z})^2 P\left(\widetilde{\mathbb{A}}_{i,\epsilon}\right).
\end{aligned}
$$

Putting things together, we have

$$
\begin{aligned}
&\mathbb{E}\left((\mathrm{z}_i - \mathbb{E}(\mathrm{z}_i))^2 \mathbb{1}_{\mathbb{A}_{i,\epsilon}}\right) \\
&\leq \mathbb{E}\left(\mathrm{z}_i - \mathbb{E}(\mathrm{z}_i))^2 \mathbb{1}_{\widetilde{\mathbb{A}}_{i,\epsilon}}\right) \\
&= \mathbb{E}\left(\left(\mathrm{r}_{x_i,y_i}^2 - \mathrm{r}_{x_i,z_i}^2\right)^2 \mathbb{1}_{\widetilde{\mathbb{A}}_{i,\epsilon}}\right) \\
&\quad - 2\mathbb{E}(\mathrm{x}_i)\mathbb{E}\left(\left(\mathrm{r}_{x_i,y_i}^2 - \mathrm{r}_{x_i,z_i}^2\right)\mathbb{1}_{\widetilde{\mathbb{A}}_{i,\epsilon}}\right) \\
&\quad + \mathbb{E}(\mathrm{z}_i)^2 P\left(\widetilde{\mathbb{A}}_{i,\epsilon}\right) \\
&= \mathbb{E}\left(\mathrm{r}_{x_i,y_i}^4 \mathbb{1}_{\widetilde{\mathbb{A}}_{i,\epsilon}}\right) + \mathbb{E}\left(\mathrm{r}_{x_i,z_i}^4 \mathbb{1}_{\widetilde{\mathbb{A}}_{i,\epsilon}}\right) \\
&\quad - 2\mathbb{E}\left(\left(\mathrm{r}_{x_i,y_i}^2 \mathrm{r}_{x_i,z_i}^2\right)\mathbb{1}_{\widetilde{\mathbb{A}}_{i,\epsilon}}\right) \\
&\quad + \mathbb{E}(\mathrm{z}_i)P\left(\widetilde{\mathbb{A}}_{i,\epsilon}\right)\left(\mathbb{E}(\mathrm{z}_i) - 2\left(\delta_i^2(\boldsymbol{x},\boldsymbol{y}) - \delta_i^2(\boldsymbol{x},\boldsymbol{z})\right)\right) \\
&= 2\mathbb{E}\left(\mathrm{n}_{x_i}^4 \mathbb{1}_{\widetilde{\mathbb{A}}_{i,\epsilon}}\right) + 6\mathbb{E}\left(\mathrm{n}_{x_i}^2 \mathrm{n}_{y_i}^2 \mathbb{1}_{\widetilde{\mathbb{A}}_{i,\epsilon}}\right) \\
&\quad + 8\left(\delta_i^2(\boldsymbol{x},\boldsymbol{y}) + \delta_i^2(\boldsymbol{x},\boldsymbol{z}) - \delta_i(\boldsymbol{x},\boldsymbol{y})\delta_i(\boldsymbol{x},\boldsymbol{z})\right) \\
&\qquad\qquad\qquad\qquad \times \mathbb{E}\left(\mathrm{n}_{x_i}^2 \mathbb{1}_{\widetilde{\mathbb{A}}_{i,\epsilon}}\right) \\
&\quad + \delta_i^2(\boldsymbol{x},\boldsymbol{y})\delta_i^2(\boldsymbol{x},\boldsymbol{z})P\left(\widetilde{\mathbb{A}}_{i,\epsilon}\right).
\end{aligned}
$$

Now since for each $j,k,l \in \mathbb{N}$, it holds that

$$
\begin{aligned}
&\mathbb{E}\left(\mathrm{n}_{x_i}^{2j}\mathrm{n}_{y_i}^{2k}\mathrm{n}_{z_i}^{2l}\mathbb{1}_{\widetilde{\mathbb{A}}_{i,\epsilon}}\right) \\
&\leq \mathbb{E}\left(\mathrm{n}_{x_i}^{2j}\mathbb{1}_{|\mathrm{n}_{x_i}|>C_d(\epsilon)}\right)\mathbb{E}\left(\mathrm{n}_{y_i}^{2k}\right)\mathbb{E}\left(\mathrm{n}_{z_i}^{2l}\right) \\
&\quad + \mathbb{E}\left(\mathrm{n}_{y_i}^{2k}\mathbb{1}_{|\mathrm{n}_{y_i}|>C_d(\epsilon)}\right)\mathbb{E}\left(\mathrm{n}_{x_i}^{2j}\right)\mathbb{E}\left(\mathrm{n}_{z_i}^{2l}\right) \\
&\quad + \mathbb{E}\left(\mathrm{n}_{z_i}^{2l}\mathbb{1}_{|\mathrm{n}_{z_i}|>C_d(\epsilon)}\right)\mathbb{E}\left(\mathrm{n}_{x_i}^{2j}\right)\mathbb{E}\left(\mathrm{n}_{y_i}^{2k}\right),
\end{aligned}
$$

we find that

$$
\begin{aligned}
&\mathbb{E}\left((\mathrm{z}_i - \mathbb{E}(\mathrm{z}_i))^2 \mathbb{1}_{\mathbb{A}_{i,\epsilon}}\right) \\
&\leq 2\mathbb{E}\left(\mathrm{n}_{x_i}^4 \mathbb{1}_{|\mathrm{n}_{x_i}|>C_d(\epsilon)}\right) \\
&\quad + 4\mu_4' P\left(|\mathrm{n}_{x_i}| > C_d(\epsilon)\right) \\
&\quad + 12\sigma^2 \mathbb{E}\left(\mathrm{n}_{x_i}^2 \mathbb{1}_{|\mathrm{n}_{x_i}|>C_d(\epsilon)}\right) \\
&\quad + 6\sigma^4 P\left(|\mathrm{n}_{x_i}| > C_d(\epsilon)\right) \\
&\quad + 8\left(\delta_i^2(\boldsymbol{x},\boldsymbol{y}) + \delta_i^2(\boldsymbol{x},\boldsymbol{z}) - \delta_i(\boldsymbol{x},\boldsymbol{y})\delta_i(\boldsymbol{x},\boldsymbol{z})\right) \\
&\qquad \times \left(\mathbb{E}\left(\mathrm{n}_{x_i}^2 \mathbb{1}_{|\mathrm{n}_{x_i}|>C_d(\epsilon)}\right) + 2\sigma^2 P\left(|\mathrm{n}_{x_i}| > C_d(\epsilon)\right)\right) \\
&\quad + 3\delta_i^2(\boldsymbol{x},\boldsymbol{y})\delta_i^2(\boldsymbol{x},\boldsymbol{z})P\left(|\mathrm{n}_{x_i}| > C_d(\epsilon)\right) \\
&\leq 2\mathbb{E}\left(\mathrm{n}_{x_i}^4 \mathbb{1}_{|\mathrm{n}_{x_i}|>C_d(\epsilon)}\right) + 12\sigma^2 \mathbb{E}\left(\mathrm{n}_{x_i}^2 \mathbb{1}_{|\mathrm{n}_{x_i}|>C_d(\epsilon)}\right) \\
&\quad + 8\left(\delta_i^2(\boldsymbol{x},\boldsymbol{y}) + \delta_i^2(\boldsymbol{x},\boldsymbol{z}) - \delta_i(\boldsymbol{x},\boldsymbol{y})\delta_i(\boldsymbol{x},\boldsymbol{z})\right) \\
&\qquad \times \mathbb{E}\left(\mathrm{n}_{x_i}^2 \mathbb{1}_{|\mathrm{n}_{x_i}|>C_d(\epsilon)}\right) \\
&\quad + \left(4\mu_4' + 6\sigma^4 + 16\max\left\{1,\sigma^2\right\}M_i\right) \\
&\qquad \times P\left(|\mathrm{n}_{x_i}| > C_d(\epsilon)\right),
\end{aligned}
$$

where

$$
\begin{aligned}
M_i :={}& \delta_i^2(\boldsymbol{x},\boldsymbol{y}) + \delta_i^2(\boldsymbol{x},\boldsymbol{z}) - \delta_i(\boldsymbol{x},\boldsymbol{y})\delta_i(\boldsymbol{x},\boldsymbol{z}) \\
& + \delta_i^2(\boldsymbol{x},\boldsymbol{y})\delta_i^2(\boldsymbol{x},\boldsymbol{z}).
\end{aligned}
$$

Summing over $i = 1,\ldots,d$, we find that

$$
\begin{aligned}
&\frac{1}{\sigma^2(\mathrm{z}(d))}\sum_{i=1}^d \mathbb{E}\left(\mathrm{n}_{x_i}^4 \mathbb{1}_{|\mathrm{n}_{x_i}|>C_d(\epsilon)}\right) \\
&\leq \frac{d\mathbb{E}\left(\mathrm{n}_{x_1}^4 \mathbb{1}_{|\mathrm{n}_{x_1}|>C_d(\epsilon)}\right)}{\sigma^2\left(2d(\mu_4' + 3\sigma^4)\right)} \overset{d\to\infty}{\longrightarrow} 0,
\end{aligned}
$$

since μ_4' is finite and $C_d(\epsilon) \overset{d\to\infty}{\longrightarrow} +\infty$. Analogously, we have

$$
\frac{1}{\sigma^2(\mathrm{z}(d))}\sum_{i=1}^d \mathbb{E}\left(\mathrm{n}_{x_i}^2 \mathbb{1}_{|\mathrm{n}_{x_i}|>C_d(\epsilon)}\right) \overset{d\to\infty}{\longrightarrow} 0.
$$

We also have

$$
\begin{aligned}
&\frac{1}{\sigma^2(\mathrm{z}(d))}\sum_{i=1}^d \left(\delta_i^2(\boldsymbol{x},\boldsymbol{y}) + \delta_i^2(\boldsymbol{x},\boldsymbol{z}) - \delta_i(\boldsymbol{x},\boldsymbol{y})\delta_i(\boldsymbol{x},\boldsymbol{z})\right) \\
&\qquad\qquad\qquad \times \mathbb{E}\left(\mathrm{n}_{x_i}^2 \mathbb{1}_{|\mathrm{n}_{x_i}|>C_d(\epsilon)}\right) \\
&\leq \frac{1}{8\sigma^2}\mathbb{E}\left(\mathrm{n}_{x_1}^2 \mathbb{1}_{|\mathrm{n}_{x_1}|>C_d(\epsilon)}\right) \overset{d\to\infty}{\longrightarrow} 0.
\end{aligned}
$$

Finally, we have that

$$
\frac{1}{\sigma^2(\mathrm{z}(d))}\sum_{i=1}^d M_i P\left(|\mathrm{n}_{x_i}| > C_d(\epsilon)\right) \overset{d\to\infty}{\longrightarrow} 0.
$$

Indeed, given the observation above, it suffices to show that

$$
P\left(|\mathrm{n}_{x_1}| > C_d(\epsilon)\right)\frac{\sum_{i=1}^d \delta_i^2(\boldsymbol{x},\boldsymbol{y})\delta_i^2(\boldsymbol{x},\boldsymbol{z})}{\sigma^2(\mathrm{z}(d))} \overset{d\to\infty}{\longrightarrow} 0.
$$

Using Markov's inequality, (1), and (6), this follows from the fact that for d sufficiently large

$$
\begin{aligned}
&P\left(|\mathrm{n}_{x_1}| > C_d(\epsilon)\right)\frac{\sum_{i=1}^d \delta_i^2(\boldsymbol{x},\boldsymbol{y})\delta_i^2(\boldsymbol{x},\boldsymbol{z})}{\sigma^2\left(\mathrm{z}^{(d)}\right)} \\
&\leq \frac{\mathbb{E}\left(|\mathrm{n}_{x_1}|\right)d\Delta_\infty^4(d)}{C_d(\epsilon)\sigma^2(\mathrm{z}(d))} \\
&\leq \frac{12\mathbb{E}\left(|\mathrm{n}_{x_1}|\right)d\Delta_\infty^4(d)}{\sqrt{6}\epsilon\sigma^{\frac{5}{2}}(\mathrm{z}(d))} \overset{d\to\infty}{\longrightarrow} 0.
\end{aligned}
$$

We conclude that

$$
\lim_{d\to\infty}\frac{1}{\sigma^2(\mathrm{z}(d))}\mathbb{E}\left((\mathrm{z}_i - \mathbb{E}(\mathrm{z}_i))^2 \mathbb{1}_{\mathbb{A}_{i,\epsilon}}\right) = 0,
$$

and this for every $\epsilon > 0$. Hence, Linderberg's condition is satisfied, so that we may apply the central limit theorem to $\mathrm{z}(d)$, i.e.,

$$
\frac{\mathrm{z}(d) - \mu(\mathrm{z}(d))}{\sigma(\mathrm{z}(d))} \overset{d\to\infty}{\longrightarrow}_p \mathcal{N}(0,1).
$$

Since convergence in probability to a continuous distribution function necessarily implies uniform convergence [Parzen, 1960], we find that

$$
\lim_{d \to \infty} \left| P\left(\mathrm{z}(d) \leq 0 \right) - \Phi\left(-\frac{\mu\left(\mathrm{z}(d)\right)}{\sigma\left(\mathrm{z}(d)\right)} \right) \right|
$$

$$
= \lim_{d \to \infty} \left| P\left(\frac{\mathrm{z}(d) - \mu\left(\mathrm{z}(d)\right)}{\sigma\left(\mathrm{z}(d)\right)} \leq -\frac{\mu\left(\mathrm{z}(d)\right)}{\sigma\left(\mathrm{z}(d)\right)} \right) \right.
$$

$$
\left. -\Phi\left(-\frac{\mu\left(\mathrm{z}(d)\right)}{\sigma\left(\mathrm{z}(d)\right)} \right) \right| = 0,
$$

which concludes the proof. $\qquad\square$

Theorem 2.1 can now be used to easily derive Corollary 2.3 as follows.

Proof of Corollary 2.3. 1. This is an immediate consequence of Theorem 2.1.

2. By assumption, it holds that

$$
\frac{d}{\left(\left\| \boldsymbol{x}^{(d)} - \boldsymbol{z}^{(d)} \right\|^2 - \left\| \boldsymbol{x}^{(d)} - \boldsymbol{y}^{(d)} \right\|^2 \right)^2} \xrightarrow{d \to \infty} 0.
$$

Furthermore, we have

$$
0 \leq \frac{\left\| \boldsymbol{x}^{(d)} - \boldsymbol{y}^{(d)} \right\|^2 + \left\| \boldsymbol{x}^{(d)} - \boldsymbol{z}^{(d)} \right\|^2 - \left\langle \boldsymbol{x}^{(d)} - \boldsymbol{y}^{(d)}, \boldsymbol{x}^{(d)} - \boldsymbol{z}^{(d)} \right\rangle}{\left(\left\| \boldsymbol{x}^{(d)} - \boldsymbol{z}^{(d)} \right\|^2 - \left\| \boldsymbol{x}^{(d)} - \boldsymbol{y}^{(d)} \right\|^2 \right)^2}
$$

$$
\leq \frac{\left(\left\| \boldsymbol{x}^{(d)} - \boldsymbol{y}^{(d)} \right\| + \left\| \boldsymbol{x}^{(d)} - \boldsymbol{z}^{(d)} \right\| \right)^2}{\left(\left\| \boldsymbol{x}^{(d)} - \boldsymbol{z}^{(d)} \right\| - \left\| \boldsymbol{x}^{(d)} - \boldsymbol{y}^{(d)} \right\| \right)^2}
$$

$$
\times \left(\left\| \boldsymbol{x}^{(d)} - \boldsymbol{z}^{(d)} \right\| + \left\| \boldsymbol{x}^{(d)} - \boldsymbol{y}^{(d)} \right\| \right)^2
$$

$$
= \frac{1}{\left(\left\| \boldsymbol{x}^{(d)} - \boldsymbol{z}^{(d)} \right\| - \left\| \boldsymbol{x}^{(d)} - \boldsymbol{y}^{(d)} \right\| \right)^2}.
$$

Now choose any $M > 0$. We know that for d sufficiently large

$$
\left\| \boldsymbol{x}^{(d)} - \boldsymbol{z}^{(d)} \right\| - \left\| \boldsymbol{x}^{(d)} - \boldsymbol{y}^{(d)} \right\|
$$

$$
= \frac{\left\| \boldsymbol{x}^{(d)} - \boldsymbol{z}^{(d)} \right\|^2 - \left\| \boldsymbol{x}^{(d)} - \boldsymbol{y}^{(d)} \right\|^2}{\left\| \boldsymbol{x}^{(d)} - \boldsymbol{z}^{(d)} \right\| + \left\| \boldsymbol{x}^{(d)} - \boldsymbol{y}^{(d)} \right\|}
$$

$$
\geq \frac{2MC\sqrt{d}}{2\Delta_\infty(d)\sqrt{d}} \geq M,
$$

so that $\frac{1}{\left(\left\| \boldsymbol{x}^{(d)} - \boldsymbol{z}^{(d)} \right\| - \left\| \boldsymbol{x}^{(d)} - \boldsymbol{y}^{(d)} \right\| \right)^2} \xrightarrow{d \to \infty} 0$. Hence, $\zeta^{(d)}\left(\mu_4', \sigma, \boldsymbol{x}^{(d)}, \boldsymbol{y}^{(d)}, \boldsymbol{z}^{(d)} \right) \xrightarrow{d \to \infty} \infty$. The result now follows from Theorem 2.1.

$\qquad\square$