HOW DO KERNEL-BASED HIGH-DIMENSIONAL SENSOR FUSION ALGORITHMS BEHAVE UNDER HIGH DIMENSIONAL NOISE?

XIUCAI DING AND HAU-TIENG WU

Abstract. We study the behavior of two kernel based sensor fusion algorithms, nonparametric canonical correlation analysis (NCCA) and alternating diffusion (AD), under the nonnull setting that the clean datasets collected from two sensors are modeled by a common low dimensional manifold embedded in a high dimensional Euclidean space and the datasets are corrupted by high dimensional noise. We establish the asymptotic limits and convergence rates for the eigenvalues and some related edge statistics of the associated kernel matrices assuming that the sample dimension and sample size are comparably large, where NCCA and AD are conducted using the Gaussian kernel. It turns out that both the asymptotic limits and convergence rates depend on the signal-to-noise ratio (SNR) of each sensor and selected bandwidths. On one hand, we show that if NCCA and AD are directly applied to the noisy point clouds without any sanity check, it may generate artificial information that mislead scientists' interpretation. On the other hand, we prove that if the bandwidths are selected adequately, both NCCA and AD can be made robust to high dimensional noise when the SNRs are relatively large.

1. Introduction

A long lasting challenge in data science is to adequately quantify the system of interest by assembling available information from datasets collected from different sensors. This problem is commonly referred to as the *sensor fusion* problem [20, 21, 32, 49]. There are several challenges when researchers fuse sensors. For example, the datasets might be noisy and high dimensional, the sensor types might be heterogeneous, different datasets might not be properly aligned, to name but a few. The most typical algorithm in handling this problem is the canonical correlation analysis (CCA) [26] and its descendants [1, 25, 27, 37].

In the modern data analysis era, we face more challenges. Due to the advance of sensor development and growth of the complexities of problems, researchers may need to take the nonlinear structure of the datasets into account to better understand the datasets. To handle this nonlinear structure, several nonlinear sensor fusion algorithms have been developed, for example, nonparametric canonical correlation analysis (NCCA) [40], alternative diffusion (AD) [33, 46] and its generalization [43], time coupled diffusion maps [39], multiview diffusion maps [35], etc. See [43] for a recent and more thoughtful list of available tools in this direction. The main idea beyond these developments is that the nonlinear structure is modeled by various nonlinear geometric structures, and the algorithms are designed to preserve and capture this nonlinear structure. Such ideas and algorithms have been successfully applied to many real world problems, like audio-visual voice activity

detection [11], the study of the sequential audio-visual correspondence [10], automatic sleep stage annotation from two electroencephalogram signals [36], seismic event modeling [34], fetal electrocardiogram analysis [43] and IQ prediction from two fMRI paradigms [48], which is a far from complete list.

While these kernel-based sensor fusion algorithms have been developed and applied for a while, there are still several gaps toward a solid practical application and sound theoretical understanding of these tools. One important gap is understanding how the inevitable noise, particularly when the data dimension is high, impacts the kernel-based sensor fusion algorithms. For example, can we be sure if the obtained fused information really informative, particularly when the datasets are noisy or when one sensor is broken? How does noise impact the information captured by these kernel based sensor? However, to our knowledge, the developed kernel-based sensor fusion algorithms do not take care of how the noise interacts with the algorithm, and most theoretical supports are mainly based on the nonlinear data structure without considering the impact of high dimensional noise, except a recent effort in the null case [8]. In this paper, we focus on studying how high dimensional noise impacts two kernel-based sensor fusion algorithms, NCCA and AD, in the non-null case; that is, when the common information (or signal) exists.

We briefly recall the NCCA and AD algorithms. Consider two noisy point clouds, $\mathcal{X} = \{\mathbf{x}_i\}_{i=1}^n$ and $\mathcal{Y} = \{\mathbf{y}_j\}_{j=1}^n$, where $\mathbf{x}_i \in \mathbb{R}^{p_1}$ and $\mathbf{y}_j \in \mathbb{R}^{p_2}$. We are interested in understanding whether there exists a common component between the two point clouds. Here, the notion of *common component* means that there exists a common geometric structure shared by the two point clouds. For some bandwidth $h_1, h_2 > 0$ and some fixed constant $\upsilon > 0$, we consider two $n \times n$ *affinity matrices*, $\mathbf{W}_1$ and $\mathbf{W}_2$, defined as

$$\mathbf{W}_1(i,j) = \exp\left(-\upsilon \frac{\|\mathbf{x}_i - \mathbf{x}_j\|_2^2}{h_1}\right) \quad \text{and} \quad \mathbf{W}_2(i,j) = \exp\left(-\upsilon \frac{\|\mathbf{y}_i - \mathbf{y}_j\|_2^2}{h_2}\right), \quad (1)$$

where $i, j = 1, \ldots, n$. Here, $\mathbf{W}_1$ and $\mathbf{W}_2$ are related to the point cloud $\mathcal{X}$ and $\mathcal{Y}$ respectively. Denote the associated *degree matrices* $\mathbf{D}_1$ and $\mathbf{D}_2$, which are diagonal matrices such that

$$\mathbf{D}_1(i,i) = \sum_{j=1}^n \mathbf{W}_1(i,j) \quad \text{and} \quad \mathbf{D}_2(i,i) = \sum_{j=1}^n \mathbf{W}_2(i,j), \ i = 1, 2, \cdots, n\,. \quad (2)$$

Moreover, denote the transition matrices $\mathbf{A}_1, \mathbf{A}_2$ as

$$\mathbf{A}_1 = \mathbf{D}_1^{-1}\mathbf{W}_1 \quad \text{and} \quad \mathbf{A}_2 = \mathbf{D}_2^{-1}\mathbf{W}_2\,.$$

The NCCA and AD matrices are defined as

$$\mathbf{N} = \mathbf{A}_1\mathbf{A}_2^\top \quad \text{and} \quad \mathbf{A} = \mathbf{A}_1\mathbf{A}_2\,, \quad (3)$$

respectively. Usually, the top few eigenpairs of $\mathbf{N}$ and $\mathbf{A}$ are used as features of the common component shared by two sensors. We shall emphasize that in general $\mathbf{N}$ and $\mathbf{A}$ are not diagonalizable, but theoretically we can obtain the top few eigenpairs without a problem under the common manifold model [43, 46] since asymptotically $\mathbf{N}$ and $\mathbf{A}$ both converge to self-adjoint operators. To avoid this trouble, researchers also consider singular value decomposition (SVD) of $\mathbf{N}$ and $\mathbf{A}$. Note that in the current paper, for simplicity, we focus our study on the Gaussian kernels. More general kernel functions will be our future topics. Another important fact is that usually we are interested in the case when $\mathcal{X}$ and $\mathcal{Y}$ are *aligned*; that is, $\mathbf{x}_i$ and

$\mathbf{y}_i$ are sampled from the same system at the same time. However, the algorithm is general and can be applied to various combinations of two datasets of the same size.

1.1. **Some related works.** In this subsection, we summarize some related results and give an overview of our results. Our focus are the NCCA and AD matrices in (3), which are essentially products of transition matrices. We first pause to summarize the results of the affinity and transition matrices when there is only one sensor. On one hand, in the noiseless setting, the spectral properties have been studied in lots of work, for example, [2, 4, 12, 19, 22, 23, 44, 45], to name but a few. In summary, under the manifold model, researchers show that the GL converges to the Laplace–Beltrami operator in various settings with properly chosen bandwidth. On other hand, when the low dimensional manifold is contaminated by high dimensional noise, the spectral properties have been investigated in [3, 5, 8, 9, 14, 18, 30] for the pure noise setting and in [7, 14, 15, 16] for the nonnull setting. These works show that when the point cloud, for example $\mathcal{X}$, only contains noise, the eigenvalues of the affinity and transition matrices are governed by a low-rank perturbed Gram matrix when the bandwidth $h_1 = p_1$. Moreover, when the point cloud has a large signal to noise ratio (SNR), the spectral properties of GL constructed from the noisy observation will be close to that constructed from the signal only. Finally, the bandwidth also plays an important role. For a more comprehensive review and sophisticated study on the spectral properties of the affinity and transition matrices for an individual point cloud, we refer the readers to [7, Sections 1.2 and 1.3].

For the NCCA and AD matrix, on one hand, in the noiseless setting, there have been several results under the common manifold model when noise does not exist []. On the other hand, when the signals are corrupted by high dimensional noise, the only existing literature is the recent effort [8] under the null setting that both sensors only capture high dimensional white noise. In this null setup, it turns out that except for a few larger outliers, when $h_1 = p_1$ and $h_2 = p_2$, the edge eigenvalues of $n^2\mathbf{A}$ or $n^2\mathbf{N}$ converge to some deterministic limit depending on the free convolution (c.f. Definition 2.2) of two Marchenko-Pastur (MP) laws [38]. However, in the nonnull setting when both of the sensors contain a common signal, to our knowledge, there does not exist any theoretical study.

1.2. **An overview of our results.** The main contribution of this paper is a comprehensive study of NCCA and AD under the non-null case in the high dimensional setup. This result can be viewed as a continuation of the null case study in [8]. We focus on the setup that the signal is modeled by a low dimensional manifold. It turns out that this problem can be recast as studying the algorithm under the commonly applied spiked model, which will be made clear later. In addition to providing a theoretical justification based on the kernel random matrix theory, we propose a method to choose the bandwidth adaptively. Moreover, peculiar and counter-intuitive results will be presented, which emphasizes the importance of carefully applying these algorithms in practice. In Section 3, we investigate the eigenvalues of the NCCA and AD matrices when $h_1 = p_1$ and $h_2 = p_2$. This choice is inherited from the common setup in the kernel random matrix literature. The behavior of the eigenvalues varies according to both SNRs of the point clouds. The formal definition of SNR will be given in (8) and (9) after necessary notations are introduced.

Heuristically, we now regard the SNRs as n^{ζ_1} and $n^{\zeta_2}, \zeta_1, \zeta_2 \geq 0$, respectively. When both sensors capture signals with relatively small SNRs (i.e., $\zeta_1 < 1$ and $\zeta_2 < 1$), except for a few number of outliers, the eigenvalues of the normalized matrices $n^2\mathbf{N}$ and $n^2\mathbf{A}$ will converge to the quantiles of a free multiplicative convolution of two scaled and shifted MP laws. Moreover, both the number of outliers and the convergence rates rely on SNRs; see Theorem 3.1 for more details. Furthermore, if one of the sensors has large SNR, for example $0 \leq \zeta_2 < 1 \leq \zeta_1$, the eigenvalues of $n\mathbf{N}$ and $n\mathbf{A}$ will be close to a matrix which is a mixture of the signal part and noise part; see Theorem 3.2 for more details. We emphasize that this result warns us that if we directly apply NCCA and AD without any sanity check, it may result in a misleading conclusion. Even though the clean eigenfunctions do not contain useful information, the noisy ones suggest the existence of common information, which is clearly wrong. Next, when both SNRs are larger, that is $\zeta_1, \zeta_2 \geq 1$, which is the most important case in practice, the eigenvalues will be close to the clean NCCA and AD matrices and only a few eigenvalues and eigenfunctions are needed to extract the information of the underlying common manifold; see Theorem 3.3 for more details. Finally, we mention that the classic bandwidth choices $h_k = p_k$ for $k = 1, 2$ are inappropriate for the sensor fusion purpose when ζ_1, ζ_2 are large. Indeed, in this case, since the bandwidth is too small compared with the signal strength, $\mathbf{N} \approx \mathbf{I}, \mathbf{A} \approx \mathbf{I}$, we obtain limited information about the signal; see (41) for more details. In Section 4, we consider bandwidths that are compatible with the signal strength; that is, $h_k = p_k + n^{\zeta_k}$, for $k = 1, 2$. When $0 \leq \zeta_k < 1$, the signals are weak so that the change of bandwidths will not help and the results are similar to the setting when $h_k = p_k$. However, when the signals are stronger (e.g. $\zeta_k \geq 1$), NCCA and AD become non-trivial and informative so that the NCCA and AD matrices will be close to those that come from clean signals. See Theorem 4.1 for more details. In practice, ζ_1, ζ_2 are usually unknown and hard to be estimated especially the manifold has a complicated nonlinear structure. To address this issue, the bandwidths h_1 and h_2 can be chosen adaptively according to an algorithm developed in our previous work [7], which result in the same results as $h_k = p_k + n^{\zeta_k}$; see Corollary **??** for more details. In summary, when $\zeta_1, \zeta_2 \geq 1$, by choosing suitable h_1 and h_2 using our scheme (42) with Algorithm 1 of [7], we claim that this approach makes the NCCA and AD robust against the high dimensional noise.

The paper is organized as follows. In Section 2, we introduce the mathematical framework and some random matrix theory background. In Section 3, we state the main results of this paper for the classic choice of bandwidth. In Section 4, we state the main results for the modified bandwidth and adaptively chosen bandwidth. In Appendix **??**, we explain how the general nonlinear model can be reduced to (5). In Appendix 5, we offer the technical proofs of the main results. In Appendix A, we provide and prove some preliminary results which will be used in the technical proofs.

Conventions. The fundamental large parameter is n and we always assume that p_1 and p_2 are comparable to and depend on n. We use C to denote a generic positive constant, whose value may change from one line to the next. Similarly, we use ϵ, τ, δ, etc., to denote generic small positive constants. If a constant depends on a quantity a, we use $C(a)$ or C_a to indicate this dependence. For two quantities a_n and b_n depending on n, the notation $a_n = \mathrm{O}(b_n)$ means that $|a_n| \leq C|b_n|$ for

some constant $C > 0$, and $a_n = \mathrm{o}(b_n)$ means that $|a_n| \le c_n |b_n|$ for some positive sequence $c_n \downarrow 0$ as $n \to \infty$. We also use the notations $a_n \lesssim b_n$ if $a_n = \mathrm{O}(b_n)$, and $a_n \asymp b_n$ if $a_n = \mathrm{O}(b_n)$ and $b_n = \mathrm{O}(a_n)$. Moreover, we will frequently use the following notion of *stochastic domination* [17, Chapter 6.3]. Let $\mathsf{X} = \big\{\mathsf{X}^{(n)}(u) : n \in \mathbb{N},\ u \in \mathsf{U}^{(n)}\big\}$ and $\mathsf{Y} = \big\{\mathsf{Y}^{(n)}(u) : n \in \mathbb{N},\ u \in \mathsf{U}^{(n)}\big\}$ be two families of nonnegative random variables, where $\mathsf{U}^{(n)}$ is a possibly n-dependent parameter set. We say that X is *stochastically dominated* by Y, uniformly in the parameter u, if for any small $\upsilon > 0$ and large $D > 0$, there exists $n_0(\upsilon, D) \in \mathbb{N}$ so that we have $\sup_{u \in \mathsf{U}^{(n)}} \mathbb{P}\Big(\mathsf{X}^{(n)}(u) > n^{\upsilon} \mathsf{Y}^{(n)}(u)\Big) \le n^{-D}$, for a sufficiently large $n \ge n_0(\upsilon, D)$. We interchangeably use the notation $\mathsf{X} = \mathrm{O}_{\prec}(\mathsf{Y})$ or $\mathsf{X} \prec \mathsf{Y}$ if X is stochastically dominated by Y, uniformly in u, when there is no danger of confusion. In addition, we say that an n-dependent event $\Omega \equiv \Omega(n)$ holds *with high probability* if for a $D > 1$, there exists $n_0 = n_0(D) > 0$ so that $\mathbb{P}(\Omega) \ge 1 - n^{-D}$, when $n \ge n_0$. Finally, for a random vector $\mathbf{u}$, we say it is sub-Gaussian if for any deterministic vector $\mathbf{a}$, we have $\mathbb{E}(\exp(\mathbf{a}^\top \mathbf{u})) \le \exp(\|\mathbf{a}\|_2^2/2)$.

2. Mathematical framework and background

2.1. **Mathematical framework.** We focus on the following model for $\mathcal{X}$ and $\mathcal{Y}$. Assume that two sensors collect clean signals as $\{\mathbf{u}_{ix}\}_{i=1}^n \in \mathbb{R}^{p_1}$ and $\{\mathbf{u}_{iy}\}_{i=1}^n \in \mathbb{R}^{p_2}$ from two sub-Gaussian distributions. Since we focus on the distance of pairwise samples, without loss of generality, we assume that

$$\mathbb{E}\mathbf{u}_{ix} = \mathbf{0} \quad \text{and} \quad \mathbb{E}\mathbf{u}_{iy} = \mathbf{0}. \tag{4}$$

In practice, the clean signals $\{\mathbf{u}_{ix}\}$ and $\{\mathbf{u}_{iy}\}$ are contaminated by two sequences of i.i.d. sub-Gaussian noise $\{\mathbf{z}_i\} \in \mathbb{R}^{p_1}$ and $\{\mathbf{w}_i\} \in \mathbb{R}^{p_2}$, respectively, so that the data generating process follows

$$\mathbf{x}_i = \mathbf{u}_{ix} + \mathbf{z}_i,\ \mathbf{y}_i = \mathbf{u}_{iy} + \mathbf{w}_i\,, \tag{5}$$

where

$$\mathbb{E}\mathbf{z}_i = \mathbf{0}_{p_1},\ \mathrm{Cov}(\mathbf{z}_i) = \mathbf{I}_{p_1},\ \mathbb{E}\mathbf{w}_i = \mathbf{0}_{p_2},\ \mathrm{Cov}(\mathbf{w}_i) = \mathbf{I}_{p_2}. \tag{6}$$

We further assume that $\mathbf{z}_i$ and $\mathbf{w}_i$ are independent with each other and also independent of $\{\mathbf{u}_{ix}\}$ and $\{\mathbf{u}_{iy}\}$. We are mainly interested in the *high dimensional* setting; that is, p_1 and p_2 are comparably as large as n. More specifically, we assume that there exists some small constant $0 < \gamma < 1$ such that

$$\gamma \le c_1 := \frac{n}{p_1} \le \gamma^{-1},\ \gamma \le c_2 := \frac{n}{p_2} \le \gamma^{-1}. \tag{7}$$

Denote $S_1 = \mathrm{Cov}(\mathbf{u}_{ix})$ and $S_2 = \mathrm{Cov}(\mathbf{u}_{iy})$, and simplify the discussion, we assume S_1 and S_2 admit the following spectral decomposition

$$S_k = \mathrm{diag}\{\sigma_{k1}^2, \cdots, \sigma_{kd}^2, 0, \cdots, 0\} \in \mathbb{R}^{p_k \times p_k},\ k = 1, 2, \tag{8}$$

where d_1 and d_2 are fixed integers. The SNRs in our setting are defined as $\{\sigma_{1i}^2\}$ and $\{\sigma_{2i}^2\}$, respectively, so that for all $1 \le i \le d_1$ and $1 \le j \le d_2$,

$$\sigma_{1i}^2 \asymp n^{\zeta_{1i}},\ \sigma_{2j}^2 \asymp n^{\zeta_{2j}}\,, \tag{9}$$

for some constants $\zeta_{1i}, \zeta_{2j} \ge 0$. To avoid repetitions, we summarize the assumptions as follows.

Assumption 2.1. *Throughout the paper, we assume that (4)–(9) hold.*

In view of (8), the model (5) for each sensor is related to the spiked covariance matrix models [29]. Note that we do not impose any condition on the dependence of $\mathbf{u}_{ix}$ and $\mathbf{u}_{iy}$. In the usual sensor fusion application, we assume that $\mathbf{u}_{ix}$ and $\mathbf{u}_{iy}$ are *simultaneously* sampled from *one* random variable; that is $\mathbf{u}_{ix} = \mathbf{u}_{iy}$. Moreover, note that while it is out of the scope of usual application consideration, it is possible to consider $\mathbf{u}_{ix}$ and $\mathbf{u}_{iy}$ to be sampled from two independent random vectors. Below we will show some examples and results indicating what we may obtain if $\mathbf{u}_{ix}$ and $\mathbf{u}_{iy}$ are sampled from two independent random vectors. In brief, the results would be counterintuitive.

We now comment that the seemingly simple model includes the commonly considered nonlinear *common manifold model*. In the literature, common manifold model means that two sensors sample simultaneously from *one* low dimensional manifold; that is, $\mathbf{u}_{ix} = \mathbf{u}_{iy} \in M$, where M is a low dimensional smooth and compact manifold. Since we are interested in the kernel matrices depending on pairwise distance, which is invariant to rotation, when combined with the Nash's embedding theory, the common manifold model becomes a special case of the model (5). We refer readers to [7] for a detailed discussion of this relationship. We should emphasize that this relationship does not mean that we could understand the manifold structure by studying the spiked covariance model. The problem we are asking here is the nontrivial relationship between the noisy and clean affinity matrices, while the problem of exploring the manifold structure from the *clean* datasets [12, 19] is a different one. By answering the problem in this paper, when combined with the knowledge of manifold learning with clean datasets, we know how to explore the manifold structure from *noisy* datasets.

Before proceeding, we should mention that in general, the samples from two sensors might contain other structures in addition to the low dimensional manifold M. For example, in [], the principle bundle structure is considered to model the "nuisance", which can be understood as the "deterministic noise". Moreover, it is possible that the datasets captured by two sensors are not exactly on the manifold M, but from different manifolds that are diffeomorphic to M []. Specifically, the first sensor samples points $\{\mathbf{u}_{ix}\}$ from $\phi_1(M)$, while the second sensor simultaneously samples points $\{\mathbf{u}_{iy}\}$ from $\phi_2(M)$, where ϕ_1 and ϕ_2 are both diffeomorphisms and $\mathbf{u}_{iy} = \phi_1(\phi_2^{-1}(\mathbf{u}_{ix}))$. While it is possible to consider the above more complicated model, since we are interested in studying how noise impacts NCCA and AD, in this paper we focus on the vanilla common manifold model model; that is, ϕ_1 and ϕ_2 are identities maps.

2.2. **Some random matrix theory background.** In this subsection, we introduce some random matrix theory background and necessary notations. Let $\mathbf{Z} \in \mathbb{R}^{p_1 \times n}$ be the noise data matrix associated with $\{\mathbf{z}_i\}$; that is, the i-th column $\mathbf{Z}$ is $\mathbf{z}_i$, and consider the scaled noise $s\mathbf{Z}$, where $s > 0$ stands for the standard deviation of the scaled noise. Denote the empirical spectral distribution (ESD) of $\mathbf{Q} = \frac{s^2}{p_1}\mathbf{Z}^\top\mathbf{Z}$ as

$$\mu_{\mathbf{Q}}(x) = \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}_{\{\lambda_i(\mathbf{Q}) \le x\}}, \ x \in \mathbb{R}.$$

It is well-known that in the high dimensional regime (7), $\mu_{\mathbf{Q}}$ has the same asymptotic [31] as the so-called MP law [38], denoted as ν_{c_1,s^2}, satisfying

$$\nu_{c_1,s^2}(I) = (1-c_1)_+\chi_I(0) + \zeta_{c_1,s^2}(I)\,, \tag{10}$$

where $I \subset \mathbb{R}$ is a measurable set, χ_I is the indicator function and $(a)_+ := 0$ when $a \le 0$ and $(a)_+ := a$ when $a > 0$,

$$d\zeta_{c_1,s^2}(x) = \frac{1}{2\pi s^2}\frac{\sqrt{(\lambda_{+,1}-x)(x-\lambda_{-,1})}}{c_n x}\mathrm{d}x\,, \tag{11}$$

$\lambda_{+,1} = (1+s^2\sqrt{c_1})^2$ and $\lambda_{-,1} = (1-s^2\sqrt{c_1})^2$.

Denote

$$\tau_1 \equiv \tau_1(\lambda_1) := 2\left(\frac{\lambda_1}{p_1}+1\right),\ \tau_2 \equiv \tau_2(\lambda_2) := 2\left(\frac{\lambda_2}{p_2}+1\right), \tag{12}$$

and for $k = 1, 2$,

$$\varsigma_k \equiv \varsigma_k(\tau_k) := 1 - 2\upsilon\exp(-\upsilon\tau_k) - \exp(-\upsilon\tau_k)\,. \tag{13}$$

For any constant $\mathsf{a} > 0$, denote T_{a} be the shifting operator that shifts a probability measure ν defined on $\mathbb{R}$ by a; that is

$$\mathrm{T}_{\mathsf{a}}\nu(I) = \nu(I-\mathsf{a}), \tag{14}$$

where $I - \mathsf{a}$ means the shifted set. Using the notation (10), for $k = 1, 2$, denote

$$\nu_k := \mathrm{T}_{\varsigma_k}\nu_{c_k,\eta},\ \text{where}\ \ \eta = 2\upsilon\exp(-2\upsilon). \tag{15}$$

Since the statements make use of the results of free multiplication of random matrices [6], we introduce some notations from free probability theory here. For a brief summary, we refer the readers to Section 2.3. For any two probability measures μ_x and μ_y compactly supported on $\mathbb{R}_+$ (not both delta measures are supported on $\{0\}$), we denote the associated *free multiplicative convolution* as $\mu_x \boxtimes \mu_y$. Using (15), denote the free multiplicative convolution of ν_1 and ν_2 as $\nu_1 \boxtimes \nu_2$. Finally, we introduce the n-dependent quantile of some probability measure. For a given probability measure μ and $n \in \mathbb{N}$, define $\gamma_\mu(j)$ as

$$\int_{\gamma_\mu(j)}^{\infty}\mu(\mathrm{d}x) = \frac{j}{n}. \tag{16}$$

2.3. **A brief summary of free multiplication of random matrices.** In this subsection, we summarize some preliminary results about free multiplication of random matrices from [6, 28]. Given some probability measure μ, its Stieltjes transform and M-transform are defined as

$$m_\mu(z) = \int\frac{1}{x-z}\mathrm{d}\mu(x),\ M_{\mu(z)} = \frac{zm_\mu(z)}{1+zm_\mu(z)},\ z \in \mathbb{C}\backslash\mathbb{R}_+.$$

We next introduce the subordination functions utilizing M-transform [28, 47]. For any two probability measures μ_x and μ_y, there exist some analytic functions $\Omega_x(z), \Omega_y(z)$ satisfying that

$$zM_{\mu_x}(\Omega_y(z)) = zM_{\mu_y}(\Omega_x(z)) = \Omega_x(z)\Omega_y(z),\ \text{for}\ z \in \mathbb{C}\backslash\mathbb{R}_+. \tag{17}$$

Armed with the subordination functions, we now introduce the free multiplicative convolution of μ_x and μ_y, denoted as $\mu_x \boxtimes \mu_y$; see Definition 2.7 of [6].

Definition 2.2. *Denote the analytic function M by*

$$M(z) := M_{\mu_x}(\Omega_y(z)) = M_{\mu_y}(\Omega_x(z)). \tag{18}$$

Then the free multiplicative convolution $\mu_x \boxtimes \mu_y$ is defined as the unique probability measure that (18) holds for all $z \in \mathbb{C}\backslash\mathbb{R}_+$, i.e., $M(z) \equiv M_{\mu_x\boxtimes\mu_y}(z)$ is the M-transform of $\mu_x \boxtimes \mu_y$. Moreover, Ω_x and Ω_y are referred to as the subordination functions.

For ν_1 and ν_2 defined in (15), we have two sequences $a_k = \gamma_{\nu_1}(k-1)$ and $b_k = \gamma_{\nu_2}(k-1)$, where $1 \le k \le n$. Note that we have

$$\int_{a_k}^{E_{+,1}} \mathrm{d}\nu_1(x) = \frac{k-1}{n}, \ \int_{b_k}^{E_{+,2}} \mathrm{d}\nu_2(x) = \frac{k-1}{n}, \tag{19}$$

where $E_{+,1}, E_{+,2}$ are the right edges of ν_1 and ν_2, respectively. Denote two $n \times n$ positive definite matrices Σ_1 and Σ_2 as follows

$$\Sigma_1 = \mathrm{diag}\{a_1, a_2, \cdots, a_n\}, \ \Sigma_2 = \mathrm{diag}\{b_1, b_2, \cdots, b_n\}. \tag{20}$$

Let $\mathbf{U}$ be an $n \times n$ Haar distributed random matrix in $O(n)$ and denote

$$\mathbf{H} = \Sigma_2 \mathbf{U} \Sigma_1 \mathbf{U}^\top.$$

The following lemma summarizes the rigidity of eigenvalues of $\mathbf{H}$.

Lemma 2.3. *Suppose (20) holds. Then we have that*

$$\sup_j |\lambda_j(\mathbf{H}) - \gamma_{\nu_1\boxtimes\nu_2}(j)| \prec n^{-2/3}\widetilde{j}^{-1/3}, \ \widetilde{j} := \min\{p_1 \wedge n + 1 - j, j\}.$$

Proof. The proof follows from Theorems 2.14 and 2.20 of [6] since the Assumptions 2.2, 2.4 and 2.7 of [6] are satisfied. Especially, (iii) of Assumption 2.2 holds due to the square root behavior of the MP laws as indicated by (11). □

3. Main results (I)–classic bandwidth: $h_1 \asymp p_1, h_2 \asymp p_2$

In this section, we state our main results regarding the eigenvalues of $\mathbf{N}$ and $\mathbf{A}$ when $h_k \asymp p_k$, where $k = 1, 2$. For definiteness, we assume that $h_1 = p_1$ and $h_2 = p_2$. In what follows, for the ease of statements, we focus on reporting the results for $d = 1$ and hence omit the subscripts of the indices i, j in (9). For the general setting with $d > 1$, we refer the readers to Remark 3.5 below for more details. Finally, we focus on reporting the results for the NCCA matrix $\mathbf{N}$. The results for the AD matrix $\mathbf{A}$ are similar. For the details of the AD matrix $\mathbf{A}$, we refer the readers to Remark 3.4 below. Moreover, by symmetry, without loss of generality, we always assume that $\zeta_1 \ge \zeta_2$; that is, the first sensor always has larger SNR.

3.1. Noninformative region: $0 \le \min\{\zeta_1, \zeta_2\} < 1$.

In this subsection, we state the results when at least one of the sensors contains strong noise, or equivalently small SNR. In this case, the NCCA and AD will not be able to provide useful information for the underlying common manifold.

3.1.1. *When both sensors have small SNRs,* $0 \le \zeta_2 \le \zeta_1 < 1$. In Theorem 3.1 below, we consider the case when $0 \le \zeta_2 \le \zeta_1 < 1$, i.e., both of the sensors have small SNRs such that the noise dominates the signal in both sensors. For some fixed integers $\mathsf{s}_1, \mathsf{s}_2$ satisfy that

$$4 \le \mathsf{s}_k \le C4^{\mathfrak{d}_k}, \ \mathfrak{d}_k := \left\lceil \frac{1}{1-\zeta_k} \right\rceil + 1, \ k = 1, 2, \tag{21}$$

where $C > 0$ is some constant, denote T as

$$\mathsf{T} := \begin{cases} 8, & 0 \le \zeta_2 \le \zeta_1 < 0.5, \\ \mathsf{s}_1 + 8, & 0 \le \zeta_2 < 0.5 \le \zeta_1 < 1, \\ \mathsf{s}_1 + \mathsf{s}_2 + 8, & 0.5 \le \zeta_2 \le \zeta_1 < 1. \end{cases} \tag{22}$$

Moreover, define $\mathfrak{e}_k < 0, k = 1, 2$, as

$$\mathfrak{e}_k := (\zeta_k - 1)\left(\left\lceil \frac{1}{1-\zeta_k} \right\rceil + 1\right) + 1. \tag{23}$$

Theorem 3.1. *Suppose Assumption 2.1 holds with* $0 \le \zeta_2 \le \zeta_1 < 1$, $h_1 = p_1$, $h_2 = p_2$ *and* $d_1 = d_2 = 1$. *Moreover, we assume that [Do we still need Gaussian assumption?][YES. Gaussian is needed under null setting.]*

$$\mathbf{z}_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \mathbf{I}_{p_1}), \ \mathbf{w}_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \mathbf{I}_{p_2}), \tag{24}$$

and there exists some constant $\tau > 0$ *such that*

$$|c_k - 1| \ge \tau, \ k = 1, 2. \tag{25}$$

Then, when n *is sufficiently large, we have that for* $i > \mathsf{T}$ *in (22),*

$$\left|\lambda_i(n^2\mathbf{N}) - \exp(4\upsilon)\gamma_{\nu_1 \boxtimes \nu_2}(j)\right| = \mathrm{O}_{\prec}\left(\max\left\{n^{\frac{\zeta_1 - 1}{2}}, n^{\mathfrak{e}_1}\right\}\right), \tag{26}$$

where $\mathfrak{e}_1$ *is defined in (23).*

Intuitively, in this region we cannot obtain any information about the signal, since asymptotically the noise dominate the signal. In practice, the datasets might fall in this region when both sensors are corrupted or the environment noise is too strong. This intuitive is confirmed by Theorem 3.1, where we focus on reporting the bulk eigenvalues of $\mathbf{N}$ under the scaling n^2. As discussed in [7, 14], when the noise dominates the signal, the outlier eigenvalues are mainly from the kernel function expansion or the Gram matrix and hence are not useful to study the underlying manifold structure. The number of these outlier eigenvalues depend on the SNR as can be seen in (22), which can be figured out from the kernel function expansion.

We should point out that (24) and (25) are mainly technical assumptions and commonly used in the random matrix theory literature. (25) ensures that the Gram matrices are bounded from below and (24) ensures that the eigenvectors of the Gram matrix are Haar distributed. They guarantee that the bulk eigenvalues of $\mathbf{N}$ can be characterized by the free multiplicative convolution. We believe that (25) can be removed and (24) can be weakened, for example, only assuming sub-Gaussian. Since this is not the focus of the current paper, we will pursue this direction in future works. [HT: please update this accordingly.]

Finally, we mention that Theorem 3.1 holds for more general kernel function beyond the Gaussian kernel. For example, as discussed in [7, Remark 2.4], we can choose a general kernel functions which is decreasing, C^3 and $f(2) > 0$.

3.1.2. *When one sensor has a small SNR,* $0 \le \zeta_2 < 1 \le \zeta_1 < \infty$. In Theorem 3.2 below, we consider that $0 \le \zeta_2 < 1 \le \zeta_1 < \infty$, i.e., one of the sensors has large SNR whereas the other is dominated by the noise. We prepare some notations here. Let $\mathbf{W}_{1,s}$ and $\mathbf{W}_{2,s}$ be the affinity matrices associated with $\{\mathbf{u}_{ix}\}$ and $\{\mathbf{u}_{iy}\}$, respectively, where the subscript s stands for the short-hand notation for the signal. In other words, $\mathbf{W}_{1,s}$ and $\mathbf{W}_{2,s}$ are constructed from the clean signal. In general, since h_1 may be different from h_2, $\mathbf{W}_{1,s}$ and $\mathbf{W}_{2,s}$ might be different. Denote

$$\widetilde{\mathbf{W}}_{1,s} = \exp(-2\upsilon)\mathbf{W}_{1,s} + (1 - \exp(-2\upsilon))\mathbf{I}. \tag{27}$$

Analogously, we denote the associated degree matrix and transition matrix as $\widetilde{\mathbf{D}}_{1,s}$ and $\widetilde{\mathbf{A}}_{1,s}$ respectively, that is,

$$\widetilde{\mathbf{A}}_{1,s} = \widetilde{\mathbf{D}}_{1,s}^{-1}\widetilde{\mathbf{W}}_{1,s}. \tag{28}$$

We can define $\widetilde{\mathbf{W}}_{2,s}$ and $\widetilde{\mathbf{A}}_{2,s}$ similarly. Note that from the random walk perspective, $\widetilde{\mathbf{A}}_{1,s}$ (and $\widetilde{\mathbf{A}}_{2,s}$ as well) describe a lazy random walk on the clean dataset. We further introduce some other $n \times n$ matrices,

$$\mathbf{W}_{1,c}(i,j) = \exp\left(-2\upsilon\frac{(\mathbf{u}_{ix} - \mathbf{u}_{jx})^\top(\mathbf{z}_i - \mathbf{z}_j)}{p_1}\right) \quad \text{and} \quad \widetilde{\mathbf{W}}_{1,c} := \widetilde{\mathbf{W}}_{1,s} \circ \mathbf{W}_{1,c}\,.$$

We then define the associate degree matrix and transition matrix as $\widetilde{\mathbf{D}}_{1,c}$ and $\widetilde{\mathbf{A}}_{1,c}$, respectively, that is

$$\widetilde{\mathbf{A}}_{1,c} = \widetilde{\mathbf{D}}_{1,c}^{-1}\widetilde{\mathbf{W}}_{1,c}\,. \tag{29}$$

$\widetilde{\mathbf{W}}_{1,c}$ and $\widetilde{\mathbf{A}}_{1,c}$ will be used when $\zeta_1 \ge 2$ is too large so that the bandwidth $h_1 \asymp p_1$ is insufficient to capture the relationship between two different samples.

With the above notations and (13), denote

$$\widetilde{\mathbf{N}} := \begin{cases} \exp(2\upsilon)\widetilde{\mathbf{A}}_{1,s}\left(\varsigma_2\mathbf{I} + 2\frac{\upsilon\exp(-\upsilon\tau_2)}{p_2}\mathbf{W}^\top\mathbf{W}\right), & 1 \le \zeta_1 < 2; \\ \exp(2\upsilon)\widetilde{\mathbf{A}}_{1,c}\left(\varsigma_2\mathbf{I} + 2\frac{\upsilon\exp(-\upsilon\tau_2)}{p_2}\mathbf{W}^\top\mathbf{W}\right), & \zeta_1 \ge 2. \end{cases} \tag{30}$$

Using s_2 defined in (21), denote

$$\mathsf{S} := \begin{cases} 4, & 0 \le \zeta_2 < 0.5, \\ \mathsf{s}_2 + 4, & 0.5 \le \zeta_2 < 1. \end{cases} \tag{31}$$

Theorem 3.2. *Suppose Assumption 2.1 holds with* $0 \le \zeta_2 < 1 \le \zeta_1 < \infty$, $h_1 = p_1$, $h_2 = p_2$ *and* $d_1 = d_2 = 1$. *Then we have that for* $i > \mathsf{S}$ *in (31)*

$$\left|\lambda_i(n\mathbf{N}) - \lambda_i(\widetilde{\mathbf{N}})\right| = \mathrm{O}_\prec\left(\max\left\{n^{\mathfrak{e}_2}, n^{\frac{\zeta_2-1}{2}}\right\}\right), \tag{32}$$

where $\mathfrak{e}_2$ *is defined in (23) and* $\widetilde{\mathbf{N}}$ *is defined in (30). Furthermore, when* ζ_1 *is larger in the sense that for any given small constant* $\delta \in (0,1)$,

$$\zeta_1 > \frac{2}{\delta} + 1, \tag{33}$$

then with probability at least $1 - \mathrm{O}\left(n^{1-\delta(\zeta_1-1)/2}\right)$, *for some sufficiently small constant* $\epsilon > 0$ *and some constant* $C > 0$ *and all* $i \ge \mathsf{S}$, *we have*

$$|\lambda_i(n\mathbf{N}) - \exp(2\upsilon)\gamma_{\nu_2}(i)| \le C\max\{n^{-1/2+\epsilon}, n^{\mathfrak{e}_2},\ n^{\frac{\zeta_2-1}{2}}\}. \tag{34}$$

This is a potentially confusing region. In practice, it captures the situation when one sensor is corrupted so that the signal part becomes weak. Since we still have one sensor available with strong SNR, it is expected that we could still obtain something useful. However, it is shown in Theorem 3.2 that the corrupted sensor unfortunately contaminates the overall performance of the sensor fusion algorithm. Unlike the results in Theorem 3.1, the scaling is n^{-1}. Note that since the first sensor has a large SNR, the noisy transition matrix $\mathbf{A}_1$ is close to the transition matrix $\widetilde{\mathbf{A}}_{1,s}$, which only depends on the signal part when $1 \leq \zeta_1 < 2$, and the transition $\widetilde{\mathbf{A}}_{1,c}$ which is a mixture of the signal and noise when $\zeta_1 \geq 2$. This fact has been shown in [7]. However, for the second sensor, due to the strong noise, $\mathbf{A}_2$ will be close to a perturbed Gram matrix which comes from the high dimensional noise impact. Consequently, as illustrated in (32), the NCCA matrix will be close to $\widetilde{\mathbf{N}}$ which is a combination of the signal and noise. In the extreme case when ζ_1 is larger in the sense of (33), the chosen bandwidth $h_1 = p_1$ is too small compared with the signal so that the transition matrix $\mathbf{A}_1$ will be close to the identity matrix. Consequently, as in (34), the NCCA matrix will be mainly characterized by the perturbed Gram matrix whose limiting ESD follows the MP law ν_2 with proper scaling.

We should however emphasize that as has been elaborated in [7], when the SNR is large, particularly when $\zeta_1 > 2$, we should consider a different bandwidth, particularly the bandwidth determined by the percentile of pairwise distance that is commonly considered in practice. It is thus natural to ask if the bandwidth h_1 is chosen "properly", would we obtain useful information eventually. We will answer this question in the later section.

3.2. **Informative region:** $\min\{\zeta_1, \zeta_2\} \geq 1$**.** In this subsection, we state the results when both of the sensors. Recall (28), (29) and denote $\widetilde{\mathbf{A}}_{2,s}, \widetilde{\mathbf{A}}_{2,c}$ analogously for the point cloud $\mathcal{Y}$. For some constant $C > 0$, denote

$$\mathsf{R} := \begin{cases} C \log n, & \zeta_2 = 1 \\ C n^{\zeta_2 - 1}, & 1 < \zeta_2 < 2. \end{cases} \tag{35}$$

Theorem 3.3. *Suppose Assumption 2.1 holds with* $1 \leq \zeta_2 \leq \zeta_1 < \infty$, $h_1 = p_1$, $h_2 = p_2$ *and* $d_1 = d_2 = 1$. *Then we have that:*

(1). *When* $1 \leq \zeta_2 \leq \zeta_1 < 2$, *we have*

$$\left\| \mathbf{N} - \widetilde{\mathbf{A}}_{1,s} \widetilde{\mathbf{A}}_{2,s}^\top \right\| \prec n^{-1/2}. \tag{36}$$

Additionally, for $i > \mathsf{R}$ *in (35)*

$$\lambda_i(\widetilde{\mathbf{A}}_{1,s} \widetilde{\mathbf{A}}_{2,s}^\top) \prec n^{(\zeta_2 - 3)/2}. \tag{37}$$

(2). *When* $1 \leq \zeta_2 < 2 \leq \zeta_1 < \infty$, *we have that (37) holds true and*

$$\|\mathbf{N} - \widetilde{\mathbf{A}}_{1,c} \widetilde{\mathbf{A}}_{2,s}^\top\| \prec n^{-1/2}. \tag{38}$$

Moreover, when ζ_1 *is larger in the sense of (33), we have that with probability at least* $1 - \mathrm{O}\left(n^{1-\delta(\zeta_1-1)/2}\right)$, *for some sufficiently small* $\epsilon > 0$ *and some constant* $C > 0$

$$\left\| \mathbf{N} - \widetilde{\mathbf{A}}_{2,s} \right\| \leq C \left(n^{-1/2+\epsilon} + n \exp(-\upsilon(\sigma_1^2/n)^{1-\delta}) \right). \tag{39}$$

(3). *When $\zeta_1 \geq \zeta_2 \geq 2$, we have that*

$$\left\|\mathbf{N} - \widetilde{\mathbf{A}}_{1,c}\widetilde{\mathbf{A}}_{2,c}^\top\right\| \prec n^{-3/2} + n^{-\zeta_2/2}. \tag{40}$$

Moreover, if ζ_2 is larger in the sense of (33), we have that with probability at least $1 - \mathrm{O}\left(n^{1-\delta(\min\{\zeta_1,\zeta_2\}-1)/2}\right)$, for some constant $C > 0$,

$$\|\mathbf{N} - \mathbf{I}\| \leq Cn\left(\exp(-\upsilon(\sigma_1^2/n)^{1-\delta}) + \exp(-\upsilon(\sigma_2^2/n)^{1-\delta})\right). \tag{41}$$

Theorem 3.3 shows that when $h_k = p_k$, where $k = 1, 2$, and both SNRs are reasonably large, the NCCA matrix from the noisy dataset could be well approximated by that from the clean dataset of the common manifold. Combining (36) and (37), we see that except the first R eigenvalues, the remaining eigenvalues are negligible and not informative. Moreover, (2) and (3) reveal an important information about the bandwidth; that is, if the bandwidth choice is improper, like $h_1 = p_1$ and $h_2 = p_2$, the result could be misleading in general. For instance, when ζ_1 and ζ_2 are large, ideally we should have a "very clean" dataset and we shall expect to obtain useful information about the signal. However, this result says that we cannot obtain any useful information from NCCA; particularly, see (41). This however is intuitively true, since when the bandwidth is too small, the relationship of two distinct points cannot be captured by the kernel; that is, when $i \neq j$, $\exp(-\|\mathbf{x}_i - \mathbf{x}_j\|^2/p_1) \approx 0$ with high probability (see proof below for a precise statement of this argument or [7]). This problem can be fixed if we choose a proper bandwidth. In Section 4, we will state the corresponding results when the bandwidths are selected using other choices, so that this counterintuitive result will be eliminated.

Remark 3.4. Throughout the paper, we focus on reporting the results of the NCCA matrix. However, our results can also be applied to the AD matrix with a minor modification based on their definitions in (3). Specifically, Theorem 3.1 holds for $n^2\mathbf{A}$, Theorem 3.2 holds for $n\mathbf{A}$ and Theorem 3.3 holds for $\mathbf{A}$ by replacing $\widetilde{\mathbf{A}}_{1,s}\widetilde{\mathbf{A}}_{2,s}^\top$ with $\widetilde{\mathbf{A}}_{1,s}\widetilde{\mathbf{A}}_{2,s}$, $\widetilde{\mathbf{A}}_{1,c}\widetilde{\mathbf{A}}_{2,s}^\top$ with $\widetilde{\mathbf{A}}_{1,c}\widetilde{\mathbf{A}}_{2,s}$ and $\widetilde{\mathbf{A}}_{1,c}\widetilde{\mathbf{A}}_{2,c}^\top$ with $\widetilde{\mathbf{A}}_{1,c}\widetilde{\mathbf{A}}_{2,c}$. Since the proof is similar, we omit details.

Remark 3.5. In the above theorems, we focus on reporting the results for the case $d_1 = d_2 = 1$ in (8). In this remark, we discuss how to generalize the results to the setting when $d_1 > 1$ or $d_2 > 1$. First, when $0 \leq \sigma_{1i}^2, \sigma_{2j}^2 < 1, 1 \leq i \leq d_1, 1 \leq j \leq d_2$, Theorem 3.1 still holds after minor modification, for example, $\mathfrak{d}_k$ in (21) should be replaced by $\sum_{i=1}^{d_k}\mathfrak{d}_{ki}, \mathfrak{d}_{ki} := \left\lceil \frac{1}{1-\zeta_{ki}} \right\rceil + 1$ and the error bound in (26) should be replaced by

$$\max\left\{\max_i\{n^{\frac{\zeta_{1i}-1}{2}}\}, \max_j\{n^{\mathfrak{e}_{2j}}\}\right\},$$

where $\mathfrak{e}_{2j}$'s are defined similarly as in (23). Similar arguments apply for Theorem 3.2. Second, when $\zeta_{1i}, \zeta_{1j} \geq 1$, where $1 \leq i \leq d_1, 1 \leq j \leq d_2$, Theorem 3.3 holds by setting $\zeta_k := \max_j\{\zeta_{kj}\}, k = 1, 2$. Finally, suppose that there exist some integers $r_k < d_k, k = 1, 2$, such that

$$\zeta_{k1} \geq \zeta_{k2} \geq \cdots \geq \zeta_{k,r_k} \geq 1 > \zeta_{k,r_k+1} \geq \cdots \zeta_{k,d_k} \geq 0.$$

Then we have that Theorem 3.3 still holds by setting $\zeta_k := \zeta_{k1}, k = 1, 2$, and the affinity and transition matrices in (28) should be defined using signal part with

large SNRs. For example, $\mathbf{W}_{1,s}$ should be defined via

$$\mathbf{W}_{1,s}(i,j) = \exp\left(-\upsilon \frac{\|\widetilde{\mathbf{u}}_{ix} - \widetilde{\mathbf{u}}_{jx}\|_2^2}{h_1}\right), \ \widetilde{\mathbf{u}}_{ix} = (\mathbf{u}_{ix}(1), \cdots, \mathbf{u}_{ix}(r_1), 0, \cdots, 0).$$

The detailed statements and proofs are similar to the setting $d_1 = d_2 = 1$ except for extra notational complicatedness. Since this is not the main focus of the current paper, we omit details here.

4. Main results (II)–adaptive choice of bandwidth

As discussed after Theorem 3.3, when the SNRs are larger, the classic bandwidth choices are too small compared to the signals. The main reason has been elaborated in [7] when we have only one sensor. In the two sensors case, the argument is similar; that is, the distance of two different points is much larger than the bandwidth with high probability so that we may not be able to see the clean signals directly when $\zeta_1, \zeta_2 \geq 2$ as stated in (2) and (3) of Theorem 3.3. One solution to this trouble has been discussed in [7] when we have one sensor. In that paper, we consider the usually applied bandwidth selection approach in practice; that is, the bandwidth is decided by the percentile of all pairwise distances. It is thus natural to hypothesize that the same solution would hold for the kernel sensor fusion approach; that is, by choosing the bandwidth for each sensor adaptively by the same approach, we could resolve this problem. As in Section 3, we focus on the case $d_1 = d_2 = 1$, the discussion for the general setting is similar to that of Remark 3.5. As before, we also assume that $\zeta_1 \geq \zeta_2$.

We first recall the adaptive choice of data-dependent bandwidth. This approach is based on the theoretical understanding of the kernel matrix under the high dimensional setup [7], which is motivated and closely related to the empirical approach commonly used in daily practice. Let $\nu_{\mathrm{dist},1}$ and $\nu_{\mathrm{dist},2}$ be the empirical distributions of pairwise distances $\{\|\mathbf{x}_i - \mathbf{x}_j\|\}_{i\neq j}$ and $\{\|\mathbf{y}_i - \mathbf{y}_j\|\}_{i\neq j}$ respectively. Then we choose the bandwidth $h_1 > 0$ and $h_2 > 0$ by

$$\int_0^{h_1} \mathrm{d}\nu_{\mathrm{dist},1} = \omega_1 \quad \text{and} \quad \int_0^{h_2} \mathrm{d}\nu_{\mathrm{dist},2} = \omega_2, \tag{42}$$

where $0 < \omega_1 < 1$ and $0 < \omega_2 < 1$ are some fixed given constants. Denote $\widetilde{\mathbf{A}}_{1,s}$ in the same fashion as (44), $\widetilde{\mathbf{N}}$ in a way similar to (47), and $\mathbf{A}_{1,s}$ as in (45) using (42). Similarly, we can define the counterparts for the point cloud $\mathcal{Y}$.

Below, we focus on reporting the results of the NCCA matrix. The discussion for the AD matrix is similar to that of Remark 3.4. Recall $\mathbf{W}_{1,s}$ and $\mathbf{W}_{2,s}$ are the affinity matrices associated with $\{\mathbf{u}_{ix}\}$ and $\{\mathbf{u}_{iy}\}$. With a little bit abuse of notation, for $k = 1, 2$, we denote

$$\widetilde{\mathbf{W}}_{k,s} = \exp\left(-\upsilon \frac{2p_k}{h_k}\right) \mathbf{W}_{k,s} + \left(1 - \exp\left(-\upsilon \frac{2p_k}{h_k}\right)\right) \mathbf{I}, \tag{43}$$

where $\mathbf{W}_{k,s}$ are constructed using the adaptively selected bandwidth h_k. Clearly, $\mathbf{W}_{k,s}$ and $\widetilde{\mathbf{W}}_{k,s}$ differ by an isotropic spectral shift, and when $\zeta_k > 1$, asymptotically $\mathbf{W}_{k,s}$ and $\widetilde{\mathbf{W}}_{k,s}$ are the same. Note that compared to (27), the difference is that we use the modified bandwidth in (43). This difference is significant, particularly when ζ_k is large. Indeed, when ζ_k is large, $\mathbf{W}_{k,s}$ defined in (27) is close to an identify matrix; that is, no information about the signal is encoded, while $\mathbf{W}_{k,s}$

defined in (43) encodes information of the signal. Specifically, asymptotically we can show that $\mathbf{W}_{k,s}$ defined in (43) converges to an integral operator defined on the manifold, whose spectral structure is commonly used in manifold learning society to study the signal. See [7] for more discussion. We can then define the transition matrices as

$$\widetilde{\mathbf{A}}_{k,s} = \widetilde{\mathbf{D}}_{k,s}^{-1}\widetilde{\mathbf{W}}_{k,s}, k = 1, 2. \tag{44}$$

Moreover, we denote

$$\mathbf{A}_{k,s} = \mathbf{D}_{k,s}^{-1}\mathbf{W}_{k,s}, \ k = 1, 2. \tag{45}$$

Compared to (44), (45) does not contain the scaling and shift of the signal parts. Recall (12) and (13). Denote

$$\varsigma_{k,h} \equiv \varsigma_{k,h}(\tau_k, h_k) := 1 - \frac{2\upsilon p_k}{h_k}\exp\left(-\upsilon\frac{\tau_k p_k}{h_k}\right) - \exp\left(-\upsilon\frac{\tau_k p_k}{h_k}\right), \ k = 1, 2. \tag{46}$$

Theorem 4.1. *Suppose Assumption 2.1 holds with the adaptively chosen bandwidths h_1, h_2 and $d_1 = d_2 = 1$. Recall (14), (10) and (46). Then we have that:*

(1). *$0 \le \zeta_2 < 1$ (at least one sensor has a low SNR). When $0 \le \zeta_1 < 1$ as well, Theorem 3.1 holds under the assumption of (24) and (25) by replacing $\nu_k, k = 1, 2$, with*

$$\widetilde{\nu}_k := \mathrm{T}_{\varsigma_{k,h}}\nu_{c_k,\eta_k}, \quad \text{where} \quad \eta_k = \frac{2p_k\upsilon\exp(-2p_k\upsilon/h_k)}{h_k}$$

and $\exp(4\upsilon)$ with $\exp(4\upsilon p_1 p_2/(h_1 h_2))$ in (26). When $\zeta_1 \ge 1$, Theorem 3.2 holds by replacing $\mathbf{N}$ by $\widetilde{\mathbf{N}}$ in (32), where

$$\widetilde{\mathbf{N}} := \begin{cases} \exp\left(\upsilon\frac{2p_2}{h_2}\right)\widetilde{\mathbf{A}}_{1,s}\left(\varsigma_{2,h}\mathbf{I} + \frac{2\upsilon\exp(-\upsilon p_2\tau_2/h_2)}{h_2}\mathbf{W}^\top\mathbf{W}\right), & 1 \le \zeta_1 < 2; \\ \exp\left(\upsilon\frac{2p_2}{h_2}\right)\widetilde{\mathbf{A}}_{1,c}\left(\varsigma_{2,h}\mathbf{I} + \frac{2\upsilon\exp(-\upsilon p_2\tau_2/h_2)}{h_2}\mathbf{W}^\top\mathbf{W}\right), & \zeta \ge 2. \end{cases} \tag{47}$$

and when ζ_1 is large in the sense of (33), (34) holds replacing ν_2 with $\widetilde{\nu}_2$ and $\exp(2\upsilon)$ with $\exp(2p_2\upsilon/h_2)$.

(2). *$\zeta_2 \ge 1$ (both sensors have high SNRs). In this case, we have that*

$$\left\|\mathbf{N} - \widetilde{\mathbf{A}}_{1,s}\widetilde{\mathbf{A}}_{2,s}^\top\right\| \prec n^{-1/2}. \tag{48}$$

Moreover, for some constant $C > 0$ and $i \ge C\log n$, we have

$$\lambda_i(\widetilde{\mathbf{A}}_{1,s}\widetilde{\mathbf{A}}_{2,s}^\top) \prec n^{-1}. \tag{49}$$

Finally, when $\zeta_2 > 1$, we have that

$$\left\|\mathbf{N} - \mathbf{A}_{1,s}\mathbf{A}_{2,s}^\top\right\| \prec n^{-1/2} + n^{1-\zeta_2}. \tag{50}$$

Theorem 4.1 (1) states that if both sensors have low SNRs, the NCCA matrix has a similar spectral behavior as that in Theorem 3.1; that is, when the SNRs are small, due to the noise impact, even if there exists a common component, we may not obtain useful result. The reason is that we still have $h_k \asymp p_k, k = 1, 2$, with high probability (see (84)), so the bandwidth choice does not influence the conclusion. Especially, most of the eigenvalues of $\mathbf{A}$ are governed by the free multiplication convolutions of two MP type laws, which are essentially the limiting empirical spectral distributions of Gram matrices only containing white noise.

On the other hand, when the signals are stronger; that is, $\zeta_1, \zeta_2 \ge 1$, we are able to approximate the associated clean NCCA matrix of the underlying clean common component, as is detailed in Theorem 4.1 (2). This result can be interpreted as that

NCCA is robust to the noise. Especially, when $\zeta_1, \zeta_2 > 1$, we see that $\mathbf{A}_{1,s}$ and $\mathbf{A}_{2,s}$ come from the clean dataset directly. Finally, we point out that compared to (35), except the top $O(\log n)$ eigenvalues and eigenfunctions, the remaining eigenvalues are not information. When $\zeta_1, \zeta_2 \geq 2$, the NCCA matrix is always informative compared to (2) and (3) of Theorem 3.3. As a result, when combined with the existing theory about AD [33, 46], the first few eigenpairs of NCCA and AD captures the geometry of the common manifold under the manifold setup.

Theorem 4.1 (1) also describes the behavior of NCCA when one sensor has a high SNR while the other one has a low SNR, which is the most interesting and counterintuitive case. In this case, even if the bandwidths of both sensors are generated according to (42), the NCCA matrix encodes limited information about the signal. Indeed, the NCCA matrix is close to a product matrix which is a mixture of signal and noise, shown in (47). While $\widetilde{\mathbf{A}}_{1,s}$ contains information about the signal, it is contaminated by $\varsigma_{2,h}\mathbf{I} + \frac{2\upsilon \exp(-\upsilon p_2 \tau_2/h_2)}{h_2}\mathbf{W}^\top\mathbf{W}$ via production, which comes from the noise dominant dataset collected from the other sensor. While the spectral behavior of $\varsigma_{2,h}\mathbf{I} + \frac{2\upsilon \exp(-\upsilon p_2 \tau_2/h_2)}{h_2}\mathbf{W}^\top\mathbf{W}$ follows the shifted and scaled MP law, overall we obtain limited information about the signal if we apply the kernel-based sensor fusion algorithm. In this case, it is better to simply consider the dataset with a high SNR. Based on the above discussion and practical experience, we would like to mention a potential danger if we directly apply NCCA (or AD) without confirming the signal quality. This example warns us that if we directly apply AD without any sanity check, it may result in a misleading conclusion, or give us a lower quality information. Therefore, before applying NCCA and AD, it is suggested to carry out the common practice by detecting the existence of signals in each of the sensors.

For the choices of the constants ω_1 and ω_2, we comment that in practice, usually researchers choose $\omega_k = 0.25$ or 0.5 [43]. In [7], we propose an algorithm, Algorithm 1 to adaptively choose the values of them. The main idea behind is that the algorithm seeks for a bandwidth so that the affinity matrix has the most number of outlier eigenvalues. We refer the readers to [7, Section 3.2] for more details.

Last but not the least, we point out that our results can be potentially used to detect the common components. Usually, researchers count on the background knowledge to decide if common information exists. For example, it is not surprising that two electroencephalogram channels share the same brain activity. However, while physiologically the brain and heart share common information [42], it is less clear if the electroencephalogram signal and the electrocardiogram signal share anything in common, and what is the common information. Answering this complicated question may need a lot of scientific works, but the first step toward it is a powerful tool to confirm if two sensors share the same information. In Section **??**, we show some efforts towards this direction by analyzing some edge statistics. Since this is not the focus of the current paper, we will address this issue in the future work.

5. Proof of main theorems

In this section, we provide proofs of the main theoretical results in Sections 3 and 4.

5.1. **Proof of Theorem 3.1.** We need the following notations. Let $\Phi_1 = (\phi_{1,1}, \ldots, \phi_{1,n})$ with $\phi_{1,i} = \frac{1}{p_1}\|\mathbf{x}_i\|_2^2 - (1 + \sigma_1^2/p_1)$, $i = 1, 2, \cdots, n$. Similarly, we define $\Phi_2 = (\phi_{2,1}, \ldots, \phi_{2,n})$ with $\phi_{2,i} = \frac{1}{p_2}\|\mathbf{y}_i\|_2^2 - (1 + \sigma_2^2/p_2)$, $i = 1, 2, \cdots, n$. For $k = 1, 2$, we denote

$$\mathrm{Sh}_{k0}(\tau_k) := f(\tau_k)\mathbf{1}\mathbf{1}^\top, \tag{51}$$

$$\mathrm{Sh}_{k1}(\tau_k) := f'(\tau_k)[\mathbf{1}\Phi_k^\top + \Phi_k\mathbf{1}^\top], \tag{52}$$

$$\mathrm{Sh}_{k2}(\tau_k) := \frac{f^{(2)}(\tau_k)}{2}\left[\mathbf{1}(\Phi_k \circ \Phi_k)^\top + (\Phi_k \circ \Phi_k)\mathbf{1}^\top + 2\Phi_k\Phi_k^\top + 4\frac{(\sigma_k^2+1)^2 + p_k}{p_k^2}\mathbf{1}\mathbf{1}^\top\right]. \tag{53}$$

Proof. Case (1). $0 \le \zeta_2 \le \zeta_1 < 0.5$. By (92), we conclude that

$$\left\|n\mathbf{D}_1^{-1} - \exp(2\upsilon)\mathbf{I}\right\| = \mathrm{O}_\prec(n^{-1/2}),\ \left\|n\mathbf{D}_2^{-1} - \exp(2\upsilon)\mathbf{I}\right\| = \mathrm{O}_\prec(n^{-1/2}). \tag{54}$$

Therefore, it suffices to consider $\mathbf{W}_1\mathbf{W}_2$. For the ease of statement, we denote that for $k = 1, 2$ and $t \in \mathbb{N}$

$$\ell_{kt} := (-\upsilon)^t \exp(-\upsilon\tau_k),\ \mathrm{Sh}_k = \sum_{j=0}^{2}\mathrm{Sh}_{kj}. \tag{55}$$

With the above notations, by Lemma A.4, we have that

$$(\mathbf{W}_1 - \mathrm{Sh}_1)(\mathbf{W}_2 - \mathrm{Sh}_2) = \left(-2\frac{\ell_{11}}{p_1}\mathbf{X}^\top\mathbf{X} + \varsigma_1\mathbf{I} + \mathrm{O}_\prec(n^{-1/2})\right)\left(-2\frac{\ell_{21}}{p_2}\mathbf{Y}^\top\mathbf{Y} + \varsigma_2\mathbf{I} + \mathrm{O}_\prec(n^{-1/2})\right). \tag{56}$$

Let

$$\mathbf{P}_1 := -2\frac{\ell_{11}}{p_1}\mathbf{X}^\top\mathbf{X} + \varsigma_1\mathbf{I},\ \mathbf{P}_2 := -2\frac{\ell_{21}}{p_2}\mathbf{Y}^\top\mathbf{Y} + \varsigma_2\mathbf{I}. \tag{57}$$

Since $d_1 = d_2 = 1$ and $\mathbf{u}_{ix}$ and $\mathbf{u}_{iy}$ contain samples from the common manifold, we can set

$$\boldsymbol{u} = (u_1, \cdots, u_n)^\top \in \mathbb{R}^n.$$

Moreover, we denote

$$\boldsymbol{z} = (\mathbf{z}_{i1}, \cdots, \mathbf{z}_{in})^\top,\ \boldsymbol{w} = (\mathbf{w}_{i1}, \cdots, \mathbf{w}_{in})^\top \in \mathbb{R}^n.$$

Withe above notations, denote

$$\Delta_1 := -2\frac{\ell_{11}}{p_1}\boldsymbol{u}\boldsymbol{u}^\top,\ \Delta_2 := -2\frac{\ell_{21}}{p_2}\boldsymbol{u}\boldsymbol{u}^\top, \tag{58}$$

and

$$\Upsilon_1 := -2\frac{\ell_{11}}{p_1}(\boldsymbol{u}\boldsymbol{z}^\top + \boldsymbol{z}\boldsymbol{u}^\top),\ \Upsilon_2 := -2\frac{\ell_{21}}{p_2}(\boldsymbol{u}\boldsymbol{w}^\top + \boldsymbol{w}\boldsymbol{u}^\top), \tag{59}$$

and

$$\mathbf{T}_1 = -2\frac{\ell_{11}}{p_1}\mathbf{Z}^\top\mathbf{Z} + \varsigma_1\mathbf{I},\ \mathbf{T}_2 = -2\frac{\ell_{21}}{p_2}\mathbf{W}^\top\mathbf{W} + \varsigma_2\mathbf{I}.$$

Note that

$$\mathbf{P}_k = \mathbf{T}_k + \Delta_k + \Upsilon_k,\ k = 1, 2. \tag{60}$$

Moreover, Δ_k are rank-one matrices and by (85),

$$\Delta_k = \mathrm{O}_\prec\left(n^{\zeta_k}\right),\ \Upsilon_k = \mathrm{O}_\prec\left(n^{\frac{\zeta_k - 1}{2}}\right). \tag{61}$$

In light of (60), we can write

$$\mathbf{P}_1\mathbf{P}_2 = \prod_{k=1}^{2}(\mathbf{T}_k + \Delta_k + \Upsilon_k). \tag{62}$$

We can further write

$$\mathbf{P}_1\mathbf{P}_2 = \mathbf{T}_1\mathbf{T}_2 + \mathbf{R}_1 + \mathbf{R}_2, \tag{63}$$

where $\mathbf{R}_1$ is defined as

$$\mathbf{R}_1 := \mathbf{T}_1\Delta_2 + \Delta_1\mathbf{P}_2,$$

and $\mathbf{R}_2$ is defined as

$$\mathbf{R}_2 := \mathbf{T}_1\Upsilon_2 + \Upsilon_1\mathbf{P}_2.$$

On one hand, it is easy to see that $\mathrm{rank}(\mathbf{R}_1) \leq 2$. On the other hand, by (61), (86) and Lemma A.8, using the assumption that $\zeta_1 \geq \zeta_2$, we obtain that

$$\mathbf{R}_2 = \mathrm{O}_{\prec}(n^{\frac{\zeta_1-1}{2}}). \tag{64}$$

Denote the spectral decompositions of $\mathbf{T}_1$ and $\mathbf{T}_2$ as

$$\mathbf{T}_1 = \mathbf{U}_1\mathbf{\Lambda}_1\mathbf{U}_1^\top,\ \mathbf{T}_2 = \mathbf{U}_2\mathbf{\Lambda}_2\mathbf{U}_2^\top. \tag{65}$$

Let $\{a_k\}$ and $\{b_k\}$ be the quantiles of ν_1 and ν_2, respectively as constructed via (19). For some small $\epsilon > 0$, let $\{\lambda_i\}$ be the eigenvalues of $\mathbf{T}_1$, we denote the event as follows

$$\Xi := \left\{\sup_{i\geq 1}|\lambda_i - a_i| \leq \widetilde{i}^{-1/3}n^{-2/3+\epsilon}\right\},\ \widetilde{i} := \min\{(p_1-1)\wedge n + 1 - j, j\}. \tag{66}$$

Since $\bar{\mathbf{W}}$ is a Gaussian random matrix, we have that $\mathbf{U}_2$ is a Haar orthogonal random matrix. Since $\bar{\mathbf{Z}}$ and $\bar{\mathbf{W}}$ are independent, we have that $\mathbf{U} := \mathbf{U}_2^\top\mathbf{U}_1$ is also a Haar orthogonal random matrix when $\mathbf{U}_1$ is fixed. Since Lemma A.8 implies that Ξ is a high probability event, in what follows, we focus our discussion on the high probability event Ξ and $\mathbf{U}_1$ is a deterministic orthonormal matrix.

On one hand, $\mathbf{T}_1\mathbf{T}_2$ have the same eigenvalues as $\mathbf{\Lambda}_2\mathbf{U}\mathbf{\Lambda}_1\mathbf{U}^\top$. On the other hand, by Lemma A.8, we have that for $\mathbf{H} := \Sigma_2\mathbf{U}\Sigma_1\mathbf{U}^\top$

$$\left\|\mathbf{\Lambda}_2\mathbf{U}\mathbf{\Lambda}_1\mathbf{U}^\top - \mathbf{H}\right\| \prec n^{-2/3},$$

where Σ_1 and Σ_2 are diagonal matrices containing $\{a_k\}$ and $\{b_k\}$, respectively. Note that the rigidity of the eigenvalues of $\mathbf{H}$ has been studied in [6] and summarized in Lemma 2.3. Together with Lemma 2.3, we conclude that for $i \geq 1$

$$|\lambda_i(\mathbf{T}_1\mathbf{T}_2) - \gamma_{\nu_1\boxtimes\nu_2}(i)| \prec n^{-2/3}. \tag{67}$$

Note that

$$n^2\mathbf{N} = n^2\mathbf{D}_1^{-1}(\mathbf{W}_1 - \mathrm{Sh}_1)(\mathbf{W}_2 - \mathrm{Sh}_2)\mathbf{D}_2^{-1} + n^2\mathbf{D}_1^{-1}\left(\mathbf{W}_1\mathrm{Sh}_2 + \mathrm{Sh}_1\mathbf{W}_2 - \mathrm{Sh}_1\mathrm{Sh}_2\right)\mathbf{D}_2^{-1}. \tag{68}$$

We then analyze the rank of $\mathbf{W}_1\mathrm{Sh}_2 + \mathrm{Sh}_1\mathbf{W}_2 + \mathrm{Sh}_1\mathrm{Sh}_2$. Recall that for any compatible matrices A and B, we have that

$$\mathrm{rank}(AB) \leq \min\{\mathrm{rank}(A), \mathrm{rank}(B)\}.$$

Since $\mathrm{rank}(\mathrm{Sh}_1) \leq 3$ and $\mathrm{rank}(\mathrm{Sh}_2) \leq 3$, we conclude that

$$\mathrm{rank}(\mathbf{W}_1\mathrm{Sh}_2 + \mathrm{Sh}_1\mathbf{W}_2 + \mathrm{Sh}_1\mathrm{Sh}_2) \leq 6.$$

Consequently, we have that

$$\operatorname{rank}(\mathbf{R}) \le 6,\ \mathbf{R} := n^2\mathbf{D}_1^{-1}\left(\mathbf{W}_1\mathrm{Sh}_2 + \mathrm{Sh}_1\mathbf{W}_2 - \mathrm{Sh}_1\mathrm{Sh}_2\right)\mathbf{D}_2^{-1}. \tag{69}$$

By (62), (63), (68) and (69), utilizing (54), we obtain that

$$n^2\mathbf{N} = \frac{1}{\exp(-4\upsilon)}\mathbf{T}_1\mathbf{T}_2 + n^2\mathbf{D}_1^{-1}\mathbf{R}_1\mathbf{D}_2^{-1} + \mathbf{R} + \mathrm{O}_\prec(n^{\frac{\zeta_1-1}{2}}), \tag{70}$$

where we used $\|\mathbf{T}_1\| \prec 1, \|\mathbf{T}_2\| \prec 1$. Since $\operatorname{rank}(\mathbf{R}) \le 6$, $\operatorname{rank}(\mathbf{R}_1) \le 2$, by (67), we have finished our proof for case (1).

Case (2). $0 \le \zeta_2 < 0.5 \le \zeta_1 < 1$. Recall (60). In this case, according to Lemma A.4, we require a high order expansion up to the degree of $\mathfrak{d}_1$ in (21) for $\mathbf{W}_1$. Recall (23) and (62). By Lemma A.4, we have that

$$\mathbf{W}_1 - \mathrm{Sh}_{\mathsf{d}} - \mathrm{Sh}_1 = \mathbf{P}_1 + \mathrm{O}_\prec\left(n^{\mathfrak{e}_1}\right), \tag{71}$$

where Sh_1 is defined in (55) and Sh_{d} is defined in (87) below satisfying $\operatorname{rank}(\mathrm{Sh}_{\mathsf{d}}) \le \mathsf{s}_1$, $4 \le \mathsf{s}_1 \le C4^{\mathfrak{d}_1}$. Using a decomposition similar to (68) with (71), by (61) and the assumption $\zeta_1 \ge \zeta_2$, we obtain that

$$\begin{aligned}
n^2\mathbf{N} &= n^2\mathbf{D}_1^{-1}\left(\mathbf{T}_1 + \Delta_1 + \Upsilon_1 + \mathrm{O}_\prec(n^{\mathfrak{e}_1})\right)\left(\mathbf{T}_2 + \Delta_2 + \Upsilon_2 + \mathrm{O}_\prec(n^{-1/2})\right)\mathbf{D}_2^{-1} \\
&\quad + n^2\mathbf{D}_1^{-1}\left((\mathbf{W}_1 - \mathrm{Sh}_1 - \mathrm{Sh}_{\mathsf{d}})\mathrm{Sh}_2 + (\mathrm{Sh}_1 + \mathrm{Sh}_{\mathsf{d}})\mathbf{W}_2\right)\mathbf{D}_2^{-1} \\
&= n^2\mathbf{D}_1^{-1}\left(\mathbf{T}_1 + \mathrm{O}_\prec(n^{\mathfrak{e}_1})\right)\left(\mathbf{T}_2 + \mathrm{O}_\prec(n^{-1/2})\right)\mathbf{D}_2^{-1} + n^2\mathbf{D}_1^{-1}\Delta_1(\mathbf{P}_2 + \mathrm{O}_\prec(n^{-1/2}))\mathbf{D}_2^{-1} \\
&\quad + n^2\mathbf{D}_1^{-1}\mathbf{T}_1(\Delta_2 + \mathrm{O}_\prec(n^{-1/2}))\mathbf{D}_2^{-1} \\
&\quad + n^2\mathbf{D}_1^{-1}\left((\mathbf{W}_1 - \mathrm{Sh}_1 - \mathrm{Sh}_{\mathsf{d}})\mathrm{Sh}_2 + (\mathrm{Sh}_1 + \mathrm{Sh}_{\mathsf{d}})\mathbf{W}_2\right)\mathbf{D}_2^{-1} + \mathrm{O}_\prec(n^{\frac{\zeta_1-1}{2}}).
\end{aligned} \tag{72}$$

It is easy to see that the rank of the second to the fourth terms of (72) can be bounded by $\mathsf{s}_1 + 8$. On other hand, by (92), the first inequality of (54) should be replaced by

$$\left\|n(\mathbf{D}_1)^{-1} - \exp(2\upsilon)\mathbf{I}\right\| = \mathrm{O}_\prec(n^{\zeta_1-1}). \tag{73}$$

The rest of the discussion follows from the case (1). This completes our proof for case (2).

Case (3). $0.5 \le \zeta_2 \le \zeta_1 < 1$. The discussion is similar to case (2) except that we also need to conduct a high order expansion for $\mathbf{W}_2$. Similar to (71), by Lemma A.4, we have that

$$\mathbf{W}_2 - \widetilde{\mathrm{Sh}}_{\mathsf{d}} - \mathrm{Sh}_2 = \mathbf{P}_2 + \mathrm{O}_\prec\left(n^{\mathfrak{e}_2}\right). \tag{74}$$

By decomposition similar to (72), with (74), by (61), we have that

$$\begin{aligned}
n^2\mathbf{N} &= n^2\mathbf{D}_1^{-1}\left(\mathbf{T}_1 + \mathrm{O}_\prec(n^{\mathfrak{e}_1})\right)\left(\mathbf{T}_2 + \mathrm{O}_\prec(n^{\mathfrak{e}_2})\right)\mathbf{D}_2^{-1} \\
&\quad + n^2\mathbf{D}_1^{-1}\left((\Delta_1 + \mathrm{Sh}_{\mathsf{d}} + \mathrm{Sh}_1)\mathbf{W}_2 + (\mathbf{W}_1 - \mathbf{P}_1 - \mathrm{Sh}_{\mathsf{d}} - \mathrm{Sh}_1)(\widetilde{\mathrm{Sh}}_{\mathsf{d}} + \Delta_2 + \mathrm{Sh}_2)\right)\mathbf{D}_2^{-1} + \mathrm{O}_\prec(n^{\frac{\zeta_1-1}{2}}).
\end{aligned}$$

On one hand, the rank of the second term of right-hand side of the above equation can be bounded by $\mathsf{s}_1 + \mathsf{s}_2 + 8$. On the other hand, the first term can be again analyzed in the same way as heading from (66) to (67) using Lemma 2.3. Finally, by (92), similar to (73), we have that

$$\left\|n(\mathbf{D}_2)^{-1} - \exp(2\upsilon)\mathbf{I}\right\| = \mathrm{O}_\prec(n^{\zeta_2-1}). \tag{75}$$

The rest of the proof follows from the discussion of case (1). This completes the proof of Case (3) using the fact $\zeta_2 \le \zeta_1$. □

5.2. **Proof of Theorem 3.2.** In this subsection, we prove Theorem 3.2 when $0 \leq \zeta_2 < 1 \leq \zeta_1 < \infty$.

Proof. Case (1). $0 \leq \zeta_2 < 0.5$. We first decompose that

$$n\mathbf{N} = \mathbf{A}_1 n(\mathbf{W}_2 - \mathrm{Sh}_2)\mathbf{D}_2^{-1} + n\mathbf{A}_1\mathrm{Sh}_2\mathbf{D}_2^{-1}. \tag{76}$$

First, we have that $\mathrm{rank}(n\mathbf{A}_1\mathrm{Sh}_2\mathbf{D}_2^{-1}) \leq 3$. Moreover, using the decomposition (60), similar to (70), by (61), we can further write that

$$n\mathbf{N} = n\mathbf{A}_1\mathbf{T}_1\mathbf{D}_2^{-1} + n\mathbf{A}_1\mathrm{Sh}_2\mathbf{D}_2^{-1} + n\mathbf{A}_1\Delta_2\mathbf{D}_2^{-1} + \mathrm{O}_{\prec}(n^{\frac{\zeta_2-1}{2}}). \tag{77}$$

Second, by Lemma A.6, we have that

$$\left\|\mathbf{A}_1 - \widetilde{\mathbf{A}}_{1,s}\right\| \prec n^{-1/2}, \quad \left\|\mathbf{A}_1 - \widetilde{\mathbf{A}}_{1,c}\right\| \prec n^{-\alpha/2} + n^{-3/2}. \tag{78}$$

Together with (54) and the fact that $\|\mathbf{T}_1\| = \mathrm{O}_{\prec}(1)$, using the definition (30), we have that

$$\left\|\mathbf{A}_1 n\mathbf{T}_1\mathbf{D}_2^{-1} - \widetilde{\mathbf{N}}\right\| \prec n^{-1/2}. \tag{79}$$

We can therefore conclude the proof using (77).

Next, when ζ_1 is larger in the sense of (33), by Lemma A.6, we find that with probability at least $1 - \mathrm{O}\left(n^{1-\delta(\zeta_1-1)/2}\right)$, for some constant $C > 0$,

$$\|\mathbf{A}_1 - \mathbf{I}\| \leq Cn\exp\left(-\upsilon(\sigma_1^2/n)^{1-\delta}\right). \tag{80}$$

Consequently, we have

$$\|n\mathbf{N} - n\mathbf{A}_2\| \leq \|\mathbf{A}_1 - \mathbf{I}\|\,\|n\mathbf{A}_2\| \leq n^2\exp\left(-\upsilon(\sigma_1^2/n)^{1-\delta}\right), \tag{81}$$

where in the second inequality we use the fact that $\|\mathbf{A}_2\| \prec 1$ since $\lambda_1(\mathbf{A}_2) = 1$. By a result analogous to (90) for $\mathbf{A}_2$, we have that for $i > 3$,

$$|\lambda_i(n\mathbf{A}_2) - \exp(2\upsilon)\gamma_{\nu_2}(i)| \prec n^{-1/2}. \tag{82}$$

Together with (81), we conclude our proof.

Case (2). $0.5 \leq \zeta_2 < 1$. The discussion is similar to case (1) except that we need to conduct a high order expansion for $\mathbf{A}_2$. Note that

$$n\mathbf{N} = \mathbf{A}_1 n(\mathbf{W}_2 - \mathrm{Sh}_2 - \widetilde{\mathrm{Sh}_{\mathsf{d}}})\mathbf{D}_2^{-1} + n\mathbf{A}_1(\mathrm{Sh}_2 + \widetilde{\mathrm{Sh}_{\mathsf{d}}})\mathbf{D}_2^{-1}.$$

By (61), (74) and (75), we have that

$$n\mathbf{N} = \exp(2\upsilon)\mathbf{A}_1\mathbf{T}_2 + n\mathbf{A}_1\Delta_1\mathbf{D}_2^{-1} + \mathrm{O}_{\prec}(n^{\frac{\zeta_2-1}{2}} + n^{\mathfrak{e}_2}) + n\mathbf{A}_1(\mathrm{Sh}_2 + \widetilde{\mathrm{Sh}_{\mathsf{d}}})\mathbf{D}_2^{-1}.$$

We can therefore conclude our proof by (78) with a discussion similar to (79). Finally, when ζ_1 is larger, we can conclude our proof using a discussion similar to (82) with (81) and Lemma A.6. Together with (75), we conclude the proof. □

5.3. **Proof of Theorem 3.3.** In this subsection, we prove Theorem 3.3 when $\zeta_1 \geq \zeta_2 \geq 1$.

Proof. For part (1), (36) follows from (96) and an analogous result for $\mathbf{A}_2$ that

$$\left\|\mathbf{A}_2 - \widetilde{\mathbf{A}}_{2,s}\right\| \prec n^{-1/2}, \tag{83}$$

as well as the facts that $\|\mathbf{A}_1\| = \mathrm{O}_{\prec}(1), \|\mathbf{A}_2\| = \mathrm{O}_{\prec}(1)$. Second, (37) follows from (97), (2) of Lemma A.2 and the assumption that $\zeta_2 \leq \zeta_1$.

For part (2) and (3), the proof follows from (1) of Lemma A.6 and its counterpart for $\mathbf{A}_2$.

□

5.4. **Proof of Theorem 4.1.** In this subsection, we prove the results of Theorem 4.1. We first study the adaptive bandwidth h_1 and h_2. When $0 \leq \zeta_2 < 1$, by Lemma A.3 about the sub-Gaussian random vector, we have that for $i \neq j$,

$$\|\mathbf{y}_i - \mathbf{y}_j\|^2 = 2(p_2 + \sigma_2^2) + \mathrm{O}_{\prec}(p_2^{\zeta_2} + \sqrt{p_2}).$$

Since $\zeta_2 < 1$, $\|\mathbf{y}_i - \mathbf{y}_j\|_2^2$ are concentrated around $2p_2$. Then for any $\omega \in (0,1)$ and h_2 chosen according to (42), we have that

$$h_2 \asymp p_2. \tag{84}$$

Similarly, when $0 \leq \zeta_1 < 1$, we have that $h_1 \asymp p_1$. Now we can prove part (1) when $0 \leq \zeta_1 < 1$. Denote

$$f\left(\frac{\|\mathbf{x}_i - \mathbf{x}_j\|_2^2}{h_1}\right) = g_1\left(\frac{\|\mathbf{x}_i - \mathbf{x}_j\|_2^2}{p_1}\right), \ f(x) = \exp(-\upsilon x),$$

where $g_k(x) := f(p_k x/h_k)$, $k = 1, 2$. Since $\frac{p_k}{h_k} \asymp 1$, we can apply the proof of Theorem 3.1 to the kernel functions $g_k(x), k = 1, 2$. The only difference is that the constant υ is now replaced by $\upsilon p_k/h_k$. When $\zeta_1 \geq 1$, the modification is similar except that we also need to use (2) of Lemma A.6.

The other two cases can be obtained similarly by recalling the following fact. For any $\zeta_1 \geq 1$, let h_1 be the bandwidth selected using (42), we have that for some constants $C_1, C_2 > 0$, with high probability

$$C_1(\sigma_1^2 \log^{-1} n + p_1) \leq h_1 \leq C_2 \sigma_1^2 \log^2 n.$$

Also, note that (2) of Lemma A.6 holds. See Corollary 3.2 of [7] for the proof. With this fact, for part (2), (48) follows from (99) and its counterpart for $\mathbf{A}_2$ and the fact $\|\mathbf{A}_1\| \prec 1, \|\mathbf{A}_2\| \prec 1$; (49) follows from (100) and its counterpart for $\mathbf{A}_2$ and (2) of Lemma A.2; (50) follows from (101) and its counterpart for $\mathbf{A}_2$ and the assumption $\zeta_1 \geq \zeta_2$.

References

[1] Z. Bao, J. Hu, G. Pan, and W. Zhou. Canonical correlation coefficients of high-dimensional Gaussian vectors: Finite rank case. *The Annals of Statistics*, 47(1):612 – 640, 2019.

[2] M. Belkin and P. Niyogi. Towards a theoretical foundation for Laplacian-based manifold methods. *Journal of Computer and System Sciences*, 74(8):1289–1308, 2008.

[3] C. Bordenave. On Euclidean random matrices in high dimension. *Electron. Commun. Probab.*, 18:no. 25, 8, 2013.

[4] M. L. Braun. Accurate error bounds for the eigenvalues of the kernel matrix. *Journal of Machine Learning Research*, 7(82):2303–2328, 2006.

[5] X. Cheng and A. Singer. The spectrum of random inner-product kernel matrices. *Random Matrices: Theory and Applications*, 02(04):1350010, 2013.

[6] X. Ding and H. C. Ji. Local laws for multiplication of random matrices and spiked invariant model. *arXiv preprint arXiv 2010.16083*, 2020.

[7] X. Ding and H.-T. Wu. Impact of signal-to-noise ratio and bandwidth on graph Laplacian spectrum from high-dimensional noisy point cloud. *arXiv preprint arXiv 2011.10725*, 2020.

[8] X. Ding and H. T. Wu. On the spectral property of kernel-based sensor fusion algorithms of high dimensional data. *IEEE Transactions on Information Theory*, 67(1):640–670, 2021.

[9] Y. Do and V. Vu. The spectrum of random kernel matrices: Universality results for rough and varying kernels. *Random Matrices: Theory and Applications*, 02(03):1350005, 2013.

[10] D. Dov, R. Talmon, and I. Cohen. Kernel-based sensor fusion with application to audio-visual voice activity detection. *IEEE Transactions on Signal Processing*, 64(24):6406–6416, 2016.

[11] D. Dov, R. Talmon, and I. Cohen. Sequential audio-visual correspondence with alternating diffusion kernels. *IEEE Transactions on Signal Processing*, 66(12):3100–3111, 2018.

[12] D. B. Dunson, H.-T. Wu, and N. Wu. Diffusion based Gaussian process regression via heat kernel reconstruction. *Applied and Computional Harmonic Analysis*, 2021.

[13] N. El Karoui. On information plus noise kernel random matrices. *The Annals of Statistics*, 38(5):3191 – 3216, 2010.

[14] N. El Karoui. The spectrum of kernel random matrices. *The Annals of Statistics*, 38(1):1 – 50, 2010.

[15] N. El Karoui and H.-T. Wu. Graph connection Laplacian and random matrices with random blocks. *Information and Inference: A Journal of the IMA*, 4(1):1–44, 2015.

[16] N. El Karoui and H.-T. Wu. Graph connection Laplacian methods can be made robust to noise. *The Annals of Statistics*, 44(1):346 – 372, 2016.

[17] L. Erdős and H. Yau. *A Dynamical Approach to Random Matrix Theory*. Courant Lecture Notes. Courant Institute of Mathematical Sciences, New York University, 2017.

[18] Z. Fan and A. Montanari. The spectral norm of random inner-product kernel matrices. *Probab. Theory Related Fields*, 173(1-2):27–85, 2019.

[19] N. García Trillos, M. Gerlach, M. Hein, and D. Slepcev. Error estimates for spectral convergence of the graph Laplacian on random geometric graphs toward the Laplace-Beltrami operator. *Found. Comput. Math.*, 20(4):827–887, 2020.

[20] F. Gustafsson. *Statistical Sensor Fusion*. Professional Publishing House, 2012.

[21] D. R. Hardoon, S. Szedmak, and J. Shawe-Taylor. Canonical correlation analysis: An overview with application to learning methods. *Neural Computation*, 16(12):2639–2664, 2004.

[22] M. Hein, J.-Y. Audibert, and U. von Luxburg. From graphs to manifolds – weak and strong pointwise consistency of graph Laplacians. In P. Auer and R. Meir, editors, *Learning Theory*, pages 470–485, 2005.

[23] M. Hein, J.-Y. Audibert, and U. von Luxburg. Graph Laplacians and their convergence on random neighborhood graphs. *J. Mach. Learn. Res.*, 8:1325–1368, 2007.

[24] R. Horn and C. Johnson. *Matrix Analysis*. Cambridge University Press, 2nd edition, 2012.

[25] P. Horst. Relations among m sets of measures. *Psychometrika*, 26(2):129–149, 1961.
[26] H. Hotelling. Relations between two sets of variates. *Biometrika*, 28:321–377, 1936.
[27] H. Hwang, K. Jung, Y. Takane, and T. S. Woodward. A unified approach to multiple-set canonical correlation analysis and principal components analysis. *British Journal of Mathematical and Statistical Psychology*, 66(2):308–321, 2013.
[28] H. C. Ji. Regularity Properties of Free Multiplicative Convolution on the Positive Line. *International Mathematics Research Notices*, 07 2020. rnaa152.
[29] I. M. Johnstone. On the distribution of the largest eigenvalue in principal components analysis. *Ann. Statist.*, 29(2):295–327, 2001.
[30] S. P. Kasiviswanathan and M. Rudelson. Spectral norm of random kernel matrices with applications to privacy. In *Approximation, Randomization, and Combinatorial Optimization. Algorithms and Techniques, APPROX/RANDOM 2015, August 24-26, 2015, Princeton, NJ, USA*, volume 40 of *LIPIcs*, pages 898–914. Schloss Dagstuhl - Leibniz-Zentrum für Informatik, 2015.
[31] A. Knowles and J. Yin. Anisotropic local laws for random matrices. *Probab. Theory Related Fields*, 169(1-2):257–352, 2017.
[32] D. Lahat, T. Adali, and C. Jutten. Multimodal data fusion: An overview of methods, challenges, and prospects. *Proceedings of the IEEE*, 103(9):1449–1477, 2015.
[33] R. R. Lederman and R. Talmon. Learning the geometry of common latent variables using alternating-diffusion. *Applied and Computational Harmonic Analysis*, 44(3):509–536, 2018.
[34] O. Lindenbaum, Y. Bregman, N. Rabin, and A. Averbuch. Multiview kernels for low-dimensional modeling of seismic events. *IEEE Transactions on Geoscience and Remote Sensing*, 56(6):3300–3310, 2018.
[35] O. Lindenbaum, A. Yeredor, M. Salhov, and A. Averbuch. Multi-view diffusion maps. *Information Fusion*, 55:127–149, 2020.
[36] G.-R. Liu, Y.-L. Lo, J. Malik, Y.-C. Sheu, and H.-T. Wu. Diffuse to fuse eeg spectra–intrinsic geometry of sleep dynamics for classification. *Biomedical Signal Processing and Control*, 55:101576, 2020.
[37] Z. Ma and F. Yang. Sample canonical correlation coefficients of high-dimensional random vectors with finite rank correlations. *arXiv preprint arXiv 2102.03297*, 2021.
[38] V. A. Marchenko and L. A. Pastur. Distribution of eigenvalues for some sets of random matrices. *Mathematics of the USSR-Sbornik*, 1(4):457–483, 1967.
[39] N. F. Marshall and M. J. Hirn. Time coupled diffusion maps. *Applied and Computational Harmonic Analysis*, 45(3):709–728, 2018.
[40] T. Michaeli, W. Wang, and K. Livescu. Nonparametric canonical correlation analysis. In *Proceedings of the 33rd International Conference on International Conference on Machine Learning - Volume 48*, ICML'16, page 1967–1976, 2016.
[41] N. S. Pillai and J. Yin. Universality of covariance matrices. *The Annals of Applied Probability*, 24(3):935 – 1001, 2014.
[42] M. A. Samuels. The brain–heart connection. *Circulation*, 116(1):77–84, 2007.

[43] T. Shnitzer, M. Ben-Chen, L. Guibas, R. Talmon, and H.-T. Wu. Recovering hidden components in multimodal data with composite diffusion operators. *SIAM J. Math. Data Sci.*, 1(3):588–616, 2019.
[44] A. Singer. From graph to manifold laplacian: The convergence rate. *Applied and Computational Harmonic Analysis*, 21(1):128–134, 2006.
[45] A. Singer and H.-T. Wu. Vector diffusion maps and the connection Laplacian. *Comm. Pure Appl. Math.*, 65(8):1067–1144, 2012.
[46] R. Talmon and H.-T. Wu. Latent common manifold learning with alternating diffusion: Analysis and applications. *Applied and Computational Harmonic Analysis*, 47(3):848–892, 2019.
[47] D. Voiculescu. Multiplication of certain non-commuting random variables. *Journal of Operator Theory*, 18(2):223–235, 1987.
[48] L. Xiao, J. M. Stephen, T. W. Wilson, V. D. Calhoun, and Y.-P. Wang. A manifold regularized multi-task learning model for iq prediction from two fmri paradigms. *IEEE Transactions on Biomedical Engineering*, 67(3):796–806, 2019.
[49] J. Zhao, X. Xie, X. Xu, and S. Sun. Multi-view learning overview: Recent progress and new challenges. *Information Fusion*, 38:43–54, 2017.

Appendix A. Additional technical lemmas

In this section, we provide some auxiliary lemmas. The following lemma is commonly referred to as the *Gershgorin circle theorem*.

Lemma A.1. *Let $A = (a_{ij})$ be a real $n \times n$ matrix. For $1 \leq i \leq n$, let $R_i = \sum_{j \neq i} |a_{ij}|$ be the sum of the absolute values of the non-diagonal entries in the i-th row. Let $D(a_{ii}, R_i) \subseteq \mathbb{R}$ be a closed disc centered at a_{ii} with radius R_i. Such a disc is called a* Gershgorin disc. *Every eigenvalue of $A = (a_{ij})$ lies within at least one of the Gershgorin discs $D(a_{ii}, R_i)$, where $R_i = \sum_{j \neq i} |a_{ij}|$.*

Proof. See [24, Section 6.1]. □

The following lemma provides some deterministic inequalities for the products of matrices.

Lemma A.2. *(1). Suppose that $\mathbf{L}$ is a real symmetric matrix with nonnegative entries and $\mathbf{E}$ is another real symmetric matrix. Then we have that*

$$\sigma_1(\mathbf{L} \circ \mathbf{E}) \leq \max_{i,j} |\mathbf{E}(i,j)| \sigma_1(\mathbf{L}),$$

where $\sigma_1(\mathbf{L})$ stands for the largest singular value of $\mathbf{L}$.
(2). Suppose A and B are two $n \times n$ positive definite matrices. Then for all $1 \leq k \leq n$, we have that

$$\lambda_k(A)\lambda_n(B) \leq \lambda_k(AB) \leq \lambda_k(A)\lambda_1(B).$$

Proof. For (1), see [14, Lemma A.5]; (2) follows from Courant–Fischer–Weyl's min-max principle using

$$\lambda_k(AB) = \lambda_k(\sqrt{B}A\sqrt{B}) = \min_{\substack{F\subset\mathbb{R}^n\\ \dim(F)=k}} \left(\max_{x\in F\setminus\{0\}} \frac{(\sqrt{B}A\sqrt{B}x, x)}{(x,x)} \right)$$
$$= \min_{\substack{F\subset\mathbb{R}^n\\ \dim(F)=k}} \left(\max_{x\in F\setminus\{0\}} \frac{(A\sqrt{B}x, \sqrt{B}x)}{(\sqrt{B}x, \sqrt{B}x)} \frac{(Bx,x)}{(x,x)} \right).$$

□

The following Lemma A.3 collects some concentration inequalities.

Lemma A.3. *Suppose Assumption 2.1 holds with $d_1 = d_2 = 1$. Moreover, we assume that $0 \le \zeta_1, \zeta_2 < 1$ in (9), and $\{\mathbf{z}_i\}$ and $\{\mathbf{w}_i\}$ are sub-Gaussian random vectors. Then we have that*

$$\frac{1}{p_1}\left|\mathbf{x}_i^\top \mathbf{x}_j\right| \prec \frac{\sigma_1^2}{n} + \frac{1}{\sqrt{n}}, \ \frac{1}{p_2}\left|\mathbf{y}_i^\top \mathbf{y}_j\right| \prec \frac{\sigma_2^2}{n} + \frac{1}{\sqrt{n}}, \tag{85}$$

and

$$\left|\frac{1}{p_1}\|\mathbf{x}_i\|_2^2 - \left(1+\frac{\sigma_1^2}{p_1}\right)\right| \prec \frac{\sigma_1^2}{n} + \frac{1}{\sqrt{n}}, \ \left|\frac{1}{p_2}\|\mathbf{y}_i\|_2^2 - \left(1+\frac{\sigma_2^2}{p_2}\right)\right| \prec \frac{\sigma_2^2}{n} + \frac{1}{\sqrt{n}}. \tag{86}$$

Proof. See Lemma A.2 of [7]. □

In the following lemma, we prove some results regarding the concentration of the affinity matrices when $0 \le \zeta_1 < 1$ and $0 \le \zeta_2 < 1$.

Lemma A.4. *Follow the notations (51)–(53). Recall (21) for $d_1 = d_2 = 1$. For $f(x) = \exp(-vx)$, we denote Sh_d and $\widetilde{\mathrm{Sh}}_d$ such that*

$$\mathrm{Sh}_\mathsf{d}(i,j) := \sum_{k=3}^{\mathfrak{d}_1-1} \frac{f^{(k)}(\tau_1)\mathbf{L}_x(i,j)^k}{k!}, \ \widetilde{\mathrm{Sh}}_\mathsf{d}(i,j) := \sum_{k=3}^{\mathfrak{d}_2-1} \frac{f^{(k)}(\tau_2)\mathbf{L}_y(i,j)^k}{k!}, \tag{87}$$

where $\mathbf{L}_x = \mathbf{O}_x - \mathbf{P}_x$, and $\mathbf{O}_x$ and $\mathbf{P}_x$ are defined as follows

$$\mathbf{O}_x(i,j) = (1-\delta_{ij})(\phi_i + \phi_j) \ , \mathbf{P}_x(i,j) = (1-\delta_{ij})\frac{\mathbf{x}_i^\top \mathbf{x}_j}{p_1}.$$

Moreover, $\mathbf{L}_y$ can be defined similarly. Denote

$$\mathbf{K}_1 = \begin{cases} -2f'(\tau_1)p_1^{-1}\mathbf{X}^\top\mathbf{X} + \varsigma_1\mathbf{I}_n + \mathrm{Sh}_{10}(\tau) + \mathrm{Sh}_{11}(\tau) + \mathrm{Sh}_{12}(\tau), & 0 \le \alpha_1 < 0.5 \\ -2f'(\tau_1)p_1^{-1}\mathbf{X}^\top\mathbf{X} + \varsigma_1\mathbf{I}_n + \mathrm{Sh}_{10}(\tau) + \mathrm{Sh}_{11}(\tau) + \mathrm{Sh}_{12}(\tau) + \mathrm{Sh}_\mathsf{d}, & 0.5 \le \alpha_1 < 1. \end{cases} \tag{88}$$

Recall (23). Suppose $\{\mathbf{z}_i\}$ and $\{\mathbf{w}_i\}$ are sub-Gaussian random vectors, when $0 \le \zeta_1 < 1$, and $h_1 = p_1$, we have that:

$$\mathbf{W}_1 = \mathbf{K}_1 + \mathrm{O}_\prec(n^{\mathfrak{e}_1} + n^{-1/2}). \tag{89}$$

Moreover, we have that

$$\mathbf{A}_1 = \frac{1}{nf(\tau_1)}\mathbf{K}_1 + \mathrm{O}_\prec(n^{\mathfrak{e}_1} + n^{-1/2}). \tag{90}$$

Finally, for s_1 in (21), we have that

$$\mathrm{rank}(\mathrm{Sh}_\mathsf{d}) \le \mathsf{s}_1. \tag{91}$$

Similar results hold for $\mathbf{A}_2, \mathbf{W}_2$ *using* $\widetilde{\phi}_k, 1 \le k \le n$, $\widetilde{\mathrm{Sh}}_{\mathsf{d}}$ *and* $\mathrm{Sh}_{2i}, i = 0, 1, 2$.

Proof. First, (89) has been proved in [7] using entry-wise Taylor expansion and Lemma A.1; see the proof of Theorems 2.3 and 2.5 of [7]. Second, (91) has been proved in the proof of Theorem 2.5 of [7]. Third, we prove (90). By Lemma A.3 and a discussion similar to [8, Lemma IV.5], when $0 \le \zeta_1 < 1$ and $0 \le \zeta_2 < 1$

$$\left\|(n\mathbf{D}_1)^{-1} - \frac{1}{f(\tau_1)}\mathbf{I}\right\| \prec n^{\mathfrak{e}_1} + n^{-1/2}, \ \left\|(n\mathbf{D}_2)^{-1} - \frac{1}{f(\tau_2)}\mathbf{I}\right\| \prec n^{\mathfrak{e}_2} + n^{-1/2}. \quad (92)$$

Consequently,

$$\begin{aligned}\left\|n\mathbf{A}_1 - \frac{1}{f(\tau_1)}\mathbf{K}_1\right\| &\le \left\|(n\mathbf{D}_1)^{-1}\mathbf{W}_1 - \frac{1}{f(\tau_1)}\mathbf{W}_1\right\| + \frac{1}{f(\tau_1)}\|\mathbf{W}_1 - \mathbf{K}_1\| \\ &\prec (n^{\mathfrak{e}_1} + n^{-1/2})(\|\mathbf{W}_1\| + 1),\end{aligned}$$

where we used the fact that $\tau_1 < \infty$. Since $\max_{i,j}|\mathbf{W}_1(i,j)| \prec 1$, by Lemma A.1, we conclude that $\|\mathbf{W}_1\| \prec n$. This concludes our proof. □

In the following lemma, we record a result regarding the matrix norm of the signal affinity matrix $\mathbf{W}_{1,s}$. [HT: is this lemma used?]

Lemma A.5. *For* $\mathbf{W}_{1,s}$ *defined using* $\{\mathbf{u}_{ix}\}$ *with* $d_1 = 1$ *and* $h_1 = p_1$, *when* $\zeta_1 \ge 1$ *and* $\{\mathbf{z}_i\}$ *are sub-Gaussian random vectors, we have that with high probability*

$$\|\mathbf{W}_{1,s}\| = \mathrm{O}\left(n^{\max\{0, \frac{3-\zeta_1}{2}\}}\right). \quad (93)$$

Moreover, with high probability, we have that

$$\sup_i |\mathbf{D}_{1,s}(i,i)| = \mathrm{O}\left(n^{\max\{0, \frac{3-\zeta_1}{2}\}}\right). \quad (94)$$

Finally, for each fixed $1 \le i \le n$ *and some small* $\epsilon > 0$, *with high probability, there exists* $\mathrm{O}(n^{\max\{0, \frac{3-\zeta_1}{2}\}})$ j *terms such that* $|\mathbf{u}_{ix}(1) - \mathbf{u}_{jx}(1)| \le n^{-\epsilon}$ *and the rest* j *terms satisfy that* $|\mathbf{u}_{ix}(1) - \mathbf{u}_{jx}(1)| > n^{-\epsilon}$.

Proof. First, (93) has been proved in equation (4.32) of [7]. Second, (94) follows from a discussion similar to (93); see the proof of Lemma 3.6 and Corollary 2.10 of [7] for more details. Finally, the third statement has been justified in the proof of Lemma 3.6 of [7]. □

In the following lemma, we collect the results regarding the affinity matrices when $\zeta_1 \ge 1$ and $\zeta_2 \ge 1$. Recall $\widetilde{\mathbf{A}}_{1,s}$ defined via (28).

Lemma A.6. *Suppose Assumption 2.1 holds with* $d_1 = d_2 = 1$, $\zeta_1, \zeta_2 \ge 1$ *and* $\{\mathbf{z}_i\}$ *and* $\{\mathbf{w}_i\}$ *are sub-Gaussian random vectors. For some constant* $C > 0$, *denote*

$$\mathsf{T}_1 := \begin{cases} C\log n, & \zeta_1 = 1; \\ Cn^{\zeta_1 - 1}, & 1 < \zeta_2 < 2. \end{cases} \quad (95)$$

Then we have:
(1). When $h_1 = p_1$, *if* $1 \le \zeta_1 < 2$,

$$\left\|\mathbf{A}_1 - \widetilde{\mathbf{A}}_{1,s}\right\| \prec n^{-1/2}. \quad (96)$$

Moreover, moreover, we have that for $i > \mathsf{T}_1$ *in (95),*

$$\lambda_i(\widetilde{\mathbf{A}}_{1,s}) \prec n^{(\zeta_1 - 3)/2}. \quad (97)$$

On the other hand, when $\zeta_1 \geq 2$, *we have that*

$$\|\mathbf{A}_1 - \widetilde{\mathbf{A}}_{1,c}\| \prec n^{-\zeta_1/2} + n^{-3/2}.$$

Finally, when α_1 *is the larger in the sense that (33) holds, we have that with probability at least* $1 - \mathrm{O}(n^{1-\delta(\zeta_1-1)/2})$, *for some constant* $C > 0$,

$$\|\mathbf{A}_1 - \mathbf{I}\| \leq Cn \exp\left(-\upsilon n^{(\zeta_1-1)(1-\delta)}\right). \tag{98}$$

(2). When $h_1 = p_1 + \sigma_1^2$, *we have that*

$$\left\|\mathbf{A}_1 - \widetilde{\mathbf{A}}_{1,s}\right\| \prec n^{-1/2}. \tag{99}$$

Moreover, we have that for $i > C \log n$

$$\lambda_i(\widetilde{\mathbf{A}}_{1,s}) \prec n^{-1}. \tag{100}$$

Recall (45). Finally, when $\zeta_1 > 1$, *we have that*

$$\|\mathbf{A}_1 - \mathbf{A}_{1,s}\| \prec n^{-1/2} + n^{1-\zeta_1}. \tag{101}$$

Similar results hold for $\mathbf{A}_2$.

Proof. See Corollary 2.11 and Theorem 3.1 of [7]. □

In the following lemma, we prove a rough result regarding on the spectrum of the normalized affinity matrix under the setup (5) based on the discussion of [13]. [HT: is this lemma used in the current proof?]

Lemma A.7. *Consider* $\mathbf{x}_i$ *in (5) with* $\sigma_1 = \sqrt{p_1}$ *in (8) with* $d_1 = 1$. *Consider the random matrix* $\mathbf{K}$ *whose* (i,j)*th entry satisfies that*

$$\mathbf{K}(i,j) = \frac{1}{n} \exp\left(-\upsilon \frac{\|\mathbf{x}_i - \mathbf{x}_j\|^2}{p_1}\right).$$

Moreover, let $\widetilde{\mathbf{K}}$ *be a deterministic matrix whose entry satisfies that*

$$\widetilde{\mathbf{K}}(i,j) = \begin{cases} \frac{1}{n} \exp\left(-\upsilon\left(\|\mathbf{u}_{ix} - \mathbf{u}_{jx}\|^2 + 2\right)\right) & i \neq j, \\ \frac{1}{n} & i = j. \end{cases}$$

Then for any small $0 < \delta < 1$, *we have that*

$$\left\|\mathbf{K} - \widetilde{\mathbf{K}}\right\| \prec n^{-1/2}.$$

Proof. See Theorem 2.1 of [13]. In fact, [13] proves the results under the Frobenius norm. We can follow their discussion and prove the results for operator norm using Lemma A.1. □

Finally, we record the results for the rigidity of eigenvalues of non-spiked Gram matrix. Denote the non-spiked Gram matrix as $\mathbf{S}$, where

$$\mathbf{S} := \frac{1}{p_1}\mathbf{Z}^\top \mathbf{Z},$$

and its eigenvalues as $\lambda_1 \geq \cdots \geq \lambda_n$. Recall (10) and (16).

Lemma A.8. *Suppose* $\{\mathbf{z}_i\}$ *are sub-Gaussian random vectors satisfying (7), (6) and (25). Then we have*

$$\sup_j \left|\lambda_j - \gamma_{\nu_{c_1,1}}(j)\right| \prec n^{-2/3}\widetilde{j}^{-1/3}, \ \widetilde{j} := \min\left\{p_1 \wedge n + 1 - j, j\right\}.$$

Proof. See [41, Theorem 3.3]. □

Department of Statistics, University of California, Davis, CA, USA
Email address: xcading@ucdavis.edu

Department of Mathematics and Department of Statistical Science, Duke University, Durham, NC, USA
Email address: hauwu@math.duke.edu