Vector diffusion maps and random matrices with random blocks

Noureddine El Karoui ∗ and Hau-tieng Wu †

May 22, 2014

Abstract

Vector diffusion maps (VDM) is a modern data analysis technique that is starting to be applied for the analysis of high dimensional and massive datasets. Motivated by this technique, we study matrices that are akin to the ones appearing in the null case of VDM, i.e the case where there is no structure in the dataset under investigation. Developing this understanding is important in making sense of the output of the VDM algorithm - whether there is signal or not.

We hence develop a theory explaining the behavior of the spectral distribution of a large class of random matrices, in particular random matrices with random block entries. Numerical work shows that the agreement between our theoretical predictions and numerical simulations is generally very good.

1 Introduction

Vector diffusion maps (henceforth VDM) [51] is a new and promising data analysis technique for high dimensional and massive datasets. VDM and its variants are currently being used for the analysis of the cryo-Electron-microscope (cryoEM) problem [33, 51, 54], dynamical systems analysis [55], sensor network localization [19], multi-view reconstruction [36, 59], etc. VDM is a conceptual and practical generalization of the spectral techniques in statistical and machine learning, for example, Laplacian Eigenmap [3, 4, 5, 34], Diffusion Maps (DM) [17, 40, 41, 50], etc. The idea underlying these methods is that the data to be analyzed – though high-dimensional in the form given to us (think of a high-resolution picture/image as a point in the high-dimensional Euclidean space) has in fact a relatively low-dimensional structure. An idealized model is that the data points actually live on a low-dimensional geometric object, for example, a manifold, embedded in a high-dimensional Euclidean space. This model can be understood as a generalization of the model considered in principal components analysis, where the data points are assumed to - approximately - live on a low-dimensional affine space embedded in a high-dimensional Euclidean space.

Under this low dimensional assumption, DM works by doing variants of (kernel) principal components analysis on data points. Indeed, when the low-dimensional geometric object is a manifold, it gives a way to approximate the local geodesic distance on the underlying manifold and then estimate spectral properties of the heat kernel of its Laplace-Beltrami operator. It can be shown that DM is theoretically capable [6, 7, 17] to recover the geometrical and topological structure of this manifold. VDM works on more complicated objects compared with DM – an extra group relationship between the data points is assumed in addition to the low dimensional geometric structure. Take the image data for example. Depending on the problem, in that setting, two rotated versions of the same image may be considered to be different or the same objects. In fact, while they appear very different in data-analytic methods simply operating on data point, such as DM, we might view them as a single object by taking the rotation into account. In other words, we “group” the dataset into subsets so that images in each subset are the same up to rotation. In VDM, the subsets are viewed as a new point cloud and the group relationship among the images are included in the

∗UC Berkeley, Department of Statistics; Contact: nkaroui@berkeley.edu ; Support from NSF grant DMS-0847647 (CA-REER) is gratefully acknowledged.
†Stanford University, Department of Mathematics; Contact: hauwu@stanford.edu ; Support from AFOSR grant FA9550-09-1-0643 is gratefully acknowledged.

Keywords: Stieltjes transform, concentration of measure, random matrices, vector diffusion maps, semi-circle law, spectral geometry;
AMS MSC 2010 Classification: 60F99, 53A99
analysis. One direct benefit of taking the rotation into account is dimension reduction of the dataset since
the rotation information is removed. Moreover, even in this much more complicated setting, when the
subsets can be parametrized by a manifold, VDM provides tools to understand the local geodesic distance
on the manifold and the spectral properties of the heat kernel of the connection Laplacian (the natural
and relevant differential-geometric object - see [8, Chapter 1]) on it. We give more details on VDM later
in the introduction.

Though the motivation of VDM is linked to particular datasets [33, 51, 54]. VDM is naturally also
interesting as a tool addressing problems arising in many modern aspects of statistical learning, applied
mathematics and what is increasingly called data-science. A common target of analysis for these fields
is big datasets, which are more and more prevalent. In addition to its size, the explosively increasing
dimensionality of the data and the inevitable noise inside the data are two important features of modern
datasets. To handle the high dimensionality of the dataset, it is common to assume the existence of low
dimensional structure or sparsity inside the data, and design the analysis based on these assumptions. To
deal with the noise in this large $p$ (i.e many measurements per observation), large $n$ (i.e many observations)
setup, we have to take its peculiar and sometime counterintuitive behavior into account. It is therefore
natural to seek to understand the impact of “noise” - broadly defined - on the behavior of our algorithms.
As readers familiar with random matrix theory will know, the impact of noise in high-dimension can be
dramatic (see e.g [26] and [25]). This important issue is the focus of the current paper. As we will see,
VDM gives rise to a specific kind of random matrices. We study generalizations of this kind of matrices
and show that they have sometimes surprising properties.

1.1 On vector diffusion maps

Estimating the intrinsic parameters from an observation dataset $\mathcal{X}$ is a main mission in data analysis.
We call the set of intrinsic parameters the parameter space $\mathcal{P}$. As discussed above, it is commonly believed
that the parameter space has a lower dimensional structure. In many cases, this low dimensional structure
attenuates various difficulties, for example reducing the impact of what is sometime called “the curse of
dimensionality”. The model space associated with the parameter space, denoted as $\mathcal{B}$, is commonly assumed
to be the range of a transformation from $\mathcal{P}$. In some cases, $\mathcal{X}$ is the same as $\mathcal{B}$, and we may have interest
to infer $\mathcal{P}$ from $\mathcal{X}$ for the sake of extracting more understanding about the system. However, in some other
cases there might be a gap between $\mathcal{X}$ and $\mathcal{B}$. Indeed, in addition to noise, $\mathcal{B}$ might be different from $\mathcal{X}$
due to the deformation introduced by the way we observe the system or other natural processes. We call
the deformation the nuisance parameter and denote it as $\mathcal{N}$.

Consider the following toy example. Take a density function describing the finger and wrist of a child
containing growth plates. We assume that the growth plates is parameterized by a set of parameters $v \in \mathcal{P} \subset \mathbb{R}^d$, where $d \geq 1$, and $v$ varies from time to time. We denote the density function as $\psi_v : \mathbb{R}^3 \to \mathbb{R}^+$
emphasizing the dependence on $v$. At different time stamps we take X-ray images of $\psi_v$ from a fixed
rotational position $R_0 \in SO(3)$ by the X-ray transformation, denoted as

$$T_{\psi_v}(R_0)(x, y) := \int_{-\infty}^{\infty} \psi_v(xR_0^1 + yR_0^2 + tR_0^3)dt, \quad \text{where} \quad R_0 = \begin{bmatrix} R_0^1 & R_0^2 & R_0^3 \end{bmatrix},$$

(1)

$(x, y) \in \mathbb{R}^2$ and we call the unit vector $R_0^3$ the projection direction. We would like to study how the growth
plate is parametrized by $\mathcal{P}$. In this problem, the model space is $\mathcal{B} = \{T_{\psi_v}(R_0); v \in \mathcal{P}\}$. However, the
observation dataset $\mathcal{X}$ might be different from $\mathcal{B}$ since the child’s hand might vary from time to time, that is,$\mathcal{X} = \{T_{\psi_v}(R(v)R_0); v \in \mathcal{P}, R(v) \in SO(3)\}$, where $R(v)$ is a random sample of $SO(3)$. In other words,
the model space depends only on $\mathcal{P}$, while the observation dataset depend on not only $\mathcal{P}$ but also $SO(3)$,
where the extra parameters are the nuisance parameters describing how the patient rotates his head, that is,$\mathcal{N} = SO(3)$.

In general, we may formulate the above framework by the group action. Consider a metric space $Y$
equipped with a metric $d$, and a group $G$ with the identity element $e$. We call $Y$ the total space and $G$ the
structure group. The left group action of $G$ on $Y$ is a map from $G \times Y$ onto $Y$

$$G \times Y \to Y, \quad (g, x) \mapsto g \circ x$$

(2)
so that \((gh) \circ x = g \circ (h \circ x)\) is satisfied for all \(g, h \in G\) and \(x \in Y\) and \(e \circ x = x\) for all \(x \in Y\). The right group action can be defined in the same way and can be constructed by composing the left group action with the inverse group operation, so it is sufficient to discuss left actions. Take the parameter space \(\mathcal{P}\) and the model space \(\mathcal{B}\). Suppose the observation dataset \(\mathcal{X}\) is located in \(Y\), the nuisance parameter is \(G\) which acts on \(Y\), and \(\mathcal{B} = \mathcal{X}/G\). In other words, \(\mathcal{X}\) is not only parameterized by \(\mathcal{P}\), but also by \(G\). Note that \(\mathcal{X} = G \circ \mathcal{B}\) is a special case. In general, by the nature of the setup, the group action may be non-isometric, which corresponds to non-rigid deformations in the image registration literature. From the data analysis viewpoint, removing these nuisance parameters is generally helpful, for example for dimension reduction and so on.

Besides the nuisance parameter, the underlying structure of \(\mathcal{P}\) is important. In fact, even if \(\mathcal{N} = \emptyset\), or if we managed to remove \(N\) from \(\mathcal{X}\), the underlying structure of \(\mathcal{P}\) might be informative. For example, in the X-ray transform \([1]\), the projection directions of all possible projection images are parametrized by the 2-dimensional sphere \(S^2\), which contains rich geometric and topological structures. To take these non-trivial structures into account, spectral methods are commonly applied, for instance Laplacian Eigenmap, DM, etc., and lots of successes have been reported. Please see, for example, \([3, 4, 5, 17]\) and the references inside. An additional benefit of spectral methods is that they are generally based on computationally efficient algorithms.

The importance of \(\mathcal{P}\) and \(\mathcal{N}\) were discussed separately above. In some situations, the combination of \(\mathcal{P}\) and \(\mathcal{N}\) might lead to further structural information about \(\mathcal{P}\). One particular example is the class averaging algorithm aiming to improve the signal to noise ratio of the images collected from the cryoEM \([33, 51, 54]\), which is a spectral method generalizing DM.

We summarize the VDM algorithm considered in \([51, 53, 60]\) under the above framework, which motivates the block random matrix theory in this study. Suppose \(\mathcal{N} = O(m)\), where \(m \in \mathbb{N}\). Take a set of \(n > 0\) random samples from \(\mathcal{X}\), denoted as \(\mathcal{X}_n\), which corresponds to the finite random samples of the model space, denoted as \(\mathcal{B}_n = \mathcal{X}_n/O(m)\). Construct a graph \(G = (V, E)\), where \(V\) represents \(\mathcal{B}_n\), and build up a weight function \(w : E \rightarrow \mathbb{R}^+\) from the distance between pairs in \(\mathcal{B}_n\). In addition to the weight function \(w\), build up a group-valued function \(g : E \rightarrow O(m)\) quantifying the nuisance parameters among data on the vertex. Then, build up the \(n \times n\) block matrix \(S\) with \(m \times m\) block as the weighted matrix, where the \((i, j)\)-th entry of \(S\) is:

\[
S_{ij} = \begin{cases} 
    w(i,j)g(i,j) & \text{when } (i,j) \in E \\
    0 & \text{otherwise} 
\end{cases} 
\] 

\(3\)

and a \(n \times n\) block diagonal matrix \(D\) with the \(i\)-th diagonal block

\[
D_{ii} = \sum_{j \neq i} w(i,j)I_m. 
\] 

\(4\)

VDM is based on analyzing the eigenvectors and eigenvalues of the connection graph Laplacian \(C := D^{-1}S\) which leads to statistics describing \(\mathcal{P}\), for example, the vector diffusion distance. We refer the reader to Appendix A for details, in particular the VDM algorithm for the class averaging algorithm in the cryoEM problem. We mention that when \(\mathcal{B}\) is a \(d\)-dim smooth closed Riemannian manifold with metric \(g\), the edges are determined by the metric \(g\) and the nuisance group is parameterized by the frame group \(O(d)\), from the spectral geometry theorem \([6, 60]\), the eigenstructure of \(C\) contains geometric and topological information of the tangent bundle of \(\mathcal{B}\), which is complimentary to those provided by DM.

Up to now, the discussion is based on the assumption that the observation dataset is noise free. When noise exists, we ask how the noise influences the spectral method, in particular in the large \(p\) large \(n\) setup.
The influence on the kernel random matrix, for example DM, has been studied in [15, 25, 26]. We are now interested in how the noise influences the VDM(-like) algorithm. Motivated by the class averaging algorithm based on the VDM in the cryoEM problem, we consider how much confidence we have on the result by studying the following null hypothesis:

\[ H_0 : \psi = 0, \tag{5} \]

that is, the molecule of interest \( \psi : \mathbb{R}^3 \to \mathbb{R}^+ \) does not exist but the noise exists. In Section C-1, we will carefully derive the connection graph Laplacian in the class averaging algorithm based on the VDM under \( H_0 \), where all the projection images are purely independent noise. We mention that the connection graph Laplacian built up in this way is a block random matrix with additional dependent structure among the blocks introduced by the underlying low dimensional structure assumption and the way we prepare the data, and hence the more general random matrix theory is needed.

The above motivating problem opens the following general question – when a random matrix follows additional structures, does its empirical spectral distribution still converge to the semi-circle law? We are particularly interested in the case when the random matrix is a block matrix, so that each block is randomly sampled from a matrix group and there are some relationship among blocks.

1.2 Random matrices with random blocks

In light of the structure of the matrix \( S \) appearing in VDM, it is natural to study random matrices whose entries are random blocks. In the VDM case, those blocks are quite dependent. We give some more details on the “null” case of VDM when it is applied to the class averaging algorithm in the Appendix (see C-1), where we show for instance that the marginal distribution of \( g_{i,j} \) is the Haar distribution on \( SO(2) \).

A natural question is to understand what the limiting spectrum of \( S \) should look like when our dataset is basically “pure noise”.

As a first approximation of this problem, we develop a theory that characterizes the limiting distribution of random matrices with independent random blocks. Some of our results also apply to the case where the blocks are not independent, for instance when the random block matrix we consider has a “kernel”-like structure. As we will see, under quite mild assumptions on the block matrices of interest, we show that their limiting distribution is a scaled Wigner semi-circle law.

Tackling the spectral distribution of the matrix \( D \) appearing in VDM for the simulations we considered requires a number of specialized computations. Because the paper is already quite long, they are not done here and we study instead a broad class of related models. From our numerical work, it is clear that the results we get are relevant to the issues encountered in VDM and inform our thinking about this class of algorithms.

The paper is organized as follows. In Section 2 we develop a theory to explain the behavior of the limiting spectral distribution of random matrices with random block entries - a fair amount of dependence being allowed. We also present a number of situations that are intuitively close to the null case of VDM case where our theory applies. In Section 3 we present numerical work to investigate the agreement between our theoretical results and the results of numerical simulations. The appendix contains a number of needed reminders and results.

2 Theory

We consider \( N \times N \) matrices that are Hermitian with above diagonal “block-rows” (or “strips”) of height bounded by a constant \( d \). An example are matrices with i.i.d block entries but the theory we develop here applies more generally. We apply later this general theory to block matrices. There has been work on block matrices with various patterns, applying mostly to situations where the blocks are large, i.e their size is going to infinity asymptotically ([11], [43], [46]), or to specific patterns ([21] for a tridiagonal example). Our work extends and generalizes to much weaker dependence structures some results of Girko [30].
Our analysis is based on Stieltjes transforms. Throughout, we call
\[ m_n(z) = \frac{1}{N} \text{trace} \left( (M - z\text{Id}_N)^{-1} \right), \]
the Stieltjes transform of \( M \), the random matrix of interest. In much of our analysis, \( d \) is held fixed. We will let \( N \) grow to infinity.

We first show that the limiting distribution of these matrices is asymptotically deterministic (the central result in this direction is Theorem 2.1). We then show that to understand the limit, it is enough to understand a Gaussian counterpart to the matrix we are studying - this is the content of Theorem 2.5. As an application of this result, we get Theorem 2.6, p.17, which shows convergence to the Wigner semi-circle law for a very broad class of random block matrices. We ([27],[24]) and other researchers ([45], [32], [12], [13]) have used similar ideas (at a high-level at least) or rather subset of these ideas in the past. Of course, a large amount of work is still needed to tackle the particular problems we care about here.

2.1 Asymptotically deterministic character of limiting spectral distributions

We have the following “master” theorem.

**Theorem 2.1.** Suppose that the \( N \times N \) Hermitian matrix \( M \) is such that, for independent random variables \( \{Z_i\}_{i=1}^n \) and a matrix valued function \( f \),
\[ M = f(Z_1, \ldots, Z_n). \]
Suppose further that for all \( 1 \leq i \leq n \), there exists a matrix \( N_i \) such that
\[ N_i = f_i(Z_1, \ldots, Z_{i-1}, Z_{i+1}, Z_n) \]
and \( \text{rank}(M - N_i) \leq d_i \). (When \( i = 1 \), \( N_i = f_1(Z_2, \ldots, Z_n) \) and when \( i = n \), \( N_n = f_n(Z_1, \ldots, Z_{n-1}) \). \( f_i \)'s are simply matrix-valued functions.) Let \( z \in \mathbb{C}^+ \) and \( \text{Im}[z] = v > 0 \). Call
\[ m_n(z) = \frac{1}{N} \text{trace} \left( (M - z\text{Id})^{-1} \right). \]
Then, for any \( t > 0 \),
\[ P \left( |m_n(z) - \mathbb{E}(m_n(z))| > t \right) \leq C \exp \left( -c \frac{N^2 v^2 t^2}{\sum_{i=1}^n d_i^2} \right), \tag{6} \]
where \( C \) and \( c \) are two constants that do not depend on \( n \) nor \( d_i \)'s.

The previous theorem is a McDiarmid-style result for Stieltjes transforms - based on rank approximations. The conceptual approach is similar to the one we used in [24].

**Proof.** Let us call \( F_i = \sigma \{ Z_k \}_{1 \leq k \leq i} \) (i.e the \( \sigma \)-field generated by the random variables \( Z_k \)'s for \( k \leq j \) and \( F_0 = \{ \emptyset \} \). Of course,
\[ m_n(z) - \mathbb{E}(m_n(z)) = \sum_{i=1}^n \mathbb{E}(m_n(z)|F_{n-i+1}) - \mathbb{E}(m_n(z)|F_{n-i}). \]
This is clearly a sum of martingale differences, by construction. Let us \( m_n^{(i)}(z) = \frac{1}{N} \text{trace} \left( (N_i - z\text{Id})^{-1} \right). \)
Note that under our assumptions, \( N_i \) is independent of \( Z_i \), since it involves only \( \{Z_k\}_{k \neq i} \). Therefore,
\[ \mathbb{E}(m_n^{(i)}(z)|F_i) = \mathbb{E}(m_n^{(i)}(z)|F_{i-1}). \]
Hence,
\[ \mathbb{E}(m_n(z)|F_{n-i+1}) - \mathbb{E}(m_n(z)|F_{n-i}) = \mathbb{E}(m_n(z) - m_n^{(n-i+1)}(z)|F_{n-i+1}) \]
\[ - \mathbb{E}(m_n(z) - m_n^{(n-i+1)}(z)|F_{n-i}). \]
Our assumptions also guarantee that $M_{ij} = M - N_i$ is of rank at most $d_i$. Lemma B-1 in the Appendix gives

$$\left| \mathbb{E} \left( m_n(z) - m_n^{(n-i+1)}(z)|\mathcal{F}_{n-i+1} \right) \right| \leq \frac{d_i}{Nv}, \text{ and } \left| \mathbb{E} \left( m_n(z) - m_n^{(n-i+1)}(z)|\mathcal{F}_{n-i} \right) \right| \leq \frac{d_i}{Nv}.$$  Therefore,

$$\left| \mathbb{E} \left( m_n(z)|\mathcal{F}_{n-i+1} \right) - \mathbb{E} \left( m_n(z)|\mathcal{F}_{n-i} \right) \right| \leq 2 \frac{d_i}{Nv}.$$  Hence, $m_n(z) - \mathbb{E}(m_n(z))$ is a sum of bounded martingale differences. Applying the Azuma-Hoeffding inequality (as in [24] to deal with the details we have to handle here), we get that, for any $t > 0$

$$P \left( |m_n(z) - \mathbb{E}(m_n(z))| > t \right) \leq C \exp \left( -c \frac{N^2v^2t^2}{\sum_{i=1}^{n} d_i^2} \right),$$

as announced in the Theorem. □

Theorem 2.1 yields a simple proof of the following result that plays a central role in our work.

**Theorem 2.2.** Suppose the $N \times N$ Hermitian matrix $M$ can be written

$$M = \sum_{1 \leq i, j \leq n} \Theta_{i,j},$$

where $\Theta_{i,j} = f_{i,j}(Z_i, Z_j)$ is a $N \times N$ matrix and the random variables $\{Z_i\}_{i=1}^n$ are independent. ($f_{i,j}$'s are simply matrix valued functions of our random variables.)

Let $M_i$ be the Hermitian matrix

$$M_i = \Theta_{i,i} + \sum_{j \neq i} \left( \Theta_{i,j} + \Theta_{j,i} \right).$$

Assume that rank($M_i$) $\leq d_i$. Let $z \in \mathbb{C}^+$ and $\text{Im}[z] = v > 0$. Call

$$m_n(z) = \frac{1}{N} \text{trace} \left( (M - \text{Id})^{-1} \right).$$

Then, for any $t > 0$,

$$P \left( |m_n(z) - \mathbb{E}(m_n(z))| > t \right) \leq C \exp \left( -c \frac{N^2v^2t^2}{\sum_{i=1}^{n} d_i^2} \right),$$

where $C$ and $c$ are two constants that do not depend on $N$, $n$ nor $d_i$'s.

**Proof.** Let us call

$$N_i = M - M_i.$$

It is clear that $N_i = f_i(Z_1, \ldots, Z_{i-1}, Z_{i+1}, \ldots, Z_n)$. In other words, $N_i$ does not depend on $Z_i$. By our assumption on rank($M_i$) = rank($M - N_i$), we see that the hypotheses made in Theorem 2.1 are satisfied in the context of Theorem 2.2. Therefore, the conclusions of Theorem 2.1 apply here, too, and Theorem 2.2 is shown. □

### 2.1.1 Consequences of Theorem 2.2

The following consequences of Theorem 2.2 are tailored towards our applications to “random-strip” matrices and VDM-like matrices.

**Corollary 2.1.** Suppose the $N \times N$ Hermitian matrix $M$ can be written

$$M = \sum_{i=1}^{n} M_i,$$
where \( M_i \) are independent with \( \text{rank}(M_i) \leq d_i \). Let \( z \in \mathbb{C}^+ \) and \( \text{Im}[z] = v > 0 \). Call

\[
m_n(z) = \frac{1}{N} \text{trace} \left( (M - z \text{Id})^{-1} \right).
\]

Then, for any \( t > 0 \),

\[
P \left( |m_n(z) - \mathbb{E}(m_n(z))| > t \right) \leq C \exp \left( -c \frac{N^2 v^2 t^2}{\sum_{i=1}^n d_i^2} \right),
\]

where \( C \) and \( c \) are two constants that do not depend on \( n \) nor \( d_i \)'s.

Proof. The corollary is a simple consequence of Theorem 2.2. Indeed, we can apply Theorem 2.2 with \( M_i = \Theta_{i,i} \) and \( \Theta_{i,j} = \Theta_{j,i} = 0 \) if \( i \neq j \) to get Corollary 2.1. The "latent variable" \( Z_i \) is simply in this case the vector of elements of \( \Theta_{i,i} \).

As a simple consequence of the previous corollary, we have the following result which is important for the rest of the paper.

**Theorem 2.3.** Suppose the matrix \( M \) has a "strip" structure, i.e. it is composed of \( n \) strips of size \( d \times N \), where \( N = nd \), and the portions of strips that are above the \((d \times d)\) diagonal are independent. Let \( z \in \mathbb{C}^+ \) with \( \text{Im}[z] = v > 0 \). Call

\[
m_n(z) = \frac{1}{N} \text{trace} \left( (M - z \text{Id})^{-1} \right).
\]

Then, for constants \( C \) and \( c \) that do not depend on \( n \), \( d \) or our model, we have

\[
\forall t > 0, \ P \left( |m_n(z) - \mathbb{E}(m_n(z))| > t \right) \leq C \exp \left( -cnv^2 t^2 \right).
\]

Hence, the Stieltjes transform \( m_n \) of the random matrices \( M \) of interest is asymptotically deterministic. So is their spectral distribution. In the case where \( \mathbb{E}(m_n(z)) \) has a limit, the convergence of \( m_n(z) \) (and hence the spectral distribution of \( M \)) is the sense of a.s convergence.

**Proof.** This theorem is a simple consequence of Corollary 2.1 where the matrix \( M_i \) correspond to the element of the \( i \)-th strip that is above the diagonal and to the corresponding Hermitian transpose. Here \( d_i \leq 2d \) for all \( i \).

In some situations that are more complicated (for the instance the analysis of vector diffusion maps in certain cases), we will need to be able to handle more dependent structures within the matrix \( M \). The following corollary is relevant to those cases. It is targeted towards kernel-like structures.

**Corollary 2.2.** Suppose the Hermitian matrix \( M \) is a block matrix whose \((i,j)\) block can be written

\[
M[i,j] = f_{ij}(Z_i, Z_j),
\]

where \( Z_i \)'s are independent random variables. \( f_{ij} \) are deterministic functions, but could depend on \( i \) and \( j \). Suppose the matrices \( M[i,j] \) are square \( \tilde{d}_i \times \tilde{d}_j \) block matrices.

Then Theorem 2.2 applies with \( d_i = 2\tilde{d}_i \).

This type of kernel-like matrices is of particular interest to us - as VDM and its building blocks naturally give rise to such matrices.

**Proof.** In this situation, the \( N \times N \) matrix \( \Theta_{i,j} \) of Theorem 2.2 is simply the matrix consisting of 0’s except in its \((i,j)\) block (corresponding to the matrix \( M \)'s \((i,j)\) block) where it is equal to \( M[i,j] \).

Define \( M_i \) as in Theorem 2.2. All we have to do to show the validity of the corollary is therefore to verify that \( \text{rank}(M - M_i) \leq d_i = 2\tilde{d}_i \).

It is clear that \( M - M_i \) is a matrix that contains only 0’s except on its \( i \)-th block row and column. Of course, if \( A \) is a \( N \times N \) Hermitian matrix of the form

\[
A = \begin{pmatrix}
    A_{11} & A_{12} \\
    A_{21} & 0_{(N-d)\times(N-d)}
\end{pmatrix},
\]

\( \text{rank}(A) \leq 2d \). Indeed, \( A \) can be written by using at most \( 2d \) vectors (and their transposed version). So any vector \( v \) orthogonal to these \( 2d \) vectors is such that \( Av = 0_N \).

Hence, \( \text{rank}(M - M_i) \leq 2\tilde{d}_i = d_i \). We can therefore apply Theorem 2.2 under the hypotheses stated in our corollary.
2.2 Characterizing the limit: dependence on low-order moments

Let us call $M[i,i]$ the blocks on the diagonal of the block-diagonal of the matrix $M$. Let us call $M^{(0)}$ the matrix obtained by replacing the block diagonal entries of $M$ by $0_{d\times d}$ and leaving the other elements of $M$ intact. We first note that when $M$ is such that $\sup_{1 \leq i \leq n} ||M[i,i]||_2 = o_P(1)$, Weyl’s inequality gives $|||M - M^{(0)}|||_2 = o_P(1)$. The spectral distributions of $M$ and $M^{(0)}$ are therefore the same in the large $N$ limit. So we will often assume that the block-diagonal of $M$ is 0 - and effectively work with $M^{(0)}$ - keeping in mind that this assumption can be removed at very low cost provided the block diagonal entries of $M$ do not grow too fast. ($M$ will eventually take the form $M = M/\sqrt{N}$, where $M$ has independent strips with distributions independent of $n$ (except for the size of the strips). So assuming that $|||M[i,i]|||_2 = o_P(1)$ will turn out to be rather minimal.)

2.2.1 Preliminaries

Lemma 2.1. Let us call $A_{12}$ be a $d \times (n-1)d$ real random matrix and $A_{21} = A'_{12}$. Let us assume that $\exists R \in R^+$ such that for any symmetric, real, deterministic matrix $\Gamma$,

$$E \left( ||A_{12}(\Gamma-zI)^{-1}A'_{12} - E \left( A_{12}(\Gamma-zI)^{-1}A'_{12} \right) ||_2 \right) \leq \frac{R}{v},$$

$$E \left( ||A_{12}(\Gamma-zI)^{-2}A'_{12} - E \left( A_{12}(\Gamma-zI)^{-2}A'_{12} \right) ||_2 \right) \leq \frac{R}{v^2},$$

and

$$|||E \left( A_{12}(\Gamma-zI)^{-2}A_{21} \right) |||_2 \leq K(z),$$

for a given function $K$.

For $N = nd$, let $T_n$ be the $N \times N$ matrix

$$T_n = \begin{pmatrix} 0_{d \times d} & A_{12} \\ A_{21} & Z_{22} \end{pmatrix}$$

where $Z_{22}$ is a real, symmetric and deterministic matrix.

Let

$$\mathcal{W}_n^{11}(z) = [-zI_d - E \left( A_{12}(Z_{22} - zI)^{-1}A_{21} \right)]^{-1},$$

and

$$\mathcal{L}(z) = \text{trace} \left( (Z_{22} - zI)^{-1} \right) - \text{trace} \left( [zI_d + E \left( A_{12}(Z_{22} - zI)^{-1}A_{21} \right)]^{-1} E \left( A_{12}(Z_{22} - zI)^{-2}A_{21} \right) \right)$$

Then, under our assumptions,

$$|E \left( \text{trace} \left( (T_n - zI)^{-1} \right) - \text{trace} \left( \mathcal{W}_n^{11}(z) \right) - \mathcal{L}(z) \right) | \leq d(2 + K(z)) \frac{R}{v^3}.$$

Proof. We call

$$(T_n - zI)^{-1} = \begin{pmatrix} T_n^{11}(z) & T_n^{12}(z) \\ T_n^{21}(z) & T_n^{22}(z) \end{pmatrix},$$

where $T_n^{11}(z)$ is $d \times d$ and $T_n^{22}(z)$ is $d(n-1) \times d(n-1)$.

Using the standard block inversion formula (see [35], p.18), we see that the top-left $d \times d$ block of $(T_n - zI)^{-1}$ is

$$T_n^{11}(z) = [-zI_d - A_{12}(Z_{22} - zI)^{-1}A_{21}]^{-1}.$$

We note that $(Z_{22} - zI)^{-1} = S_1 + iS_2$ where $S_1$ and $S_2$ are real symmetric matrices. Furthermore, after diagonalizing $Z_{22}$ it is clear that $S_2$ is positive semi-definite. So we have

$$S_3 = zI_d + A_{12}(Z_{22} - zI)^{-1}A_{21} = (A_{12}S_1A'_{12} + \text{Re} [z I]) + i(A_{12}S_2A'_{12} + vI).$$

Of course, the eigenvalues of $A_{12}S_2A'_{12} + vI$ are greater than $v$: the matrix $A_{12}S_2A'_{12}$ is positive semi-definite. Therefore, by applying the Fan-Hoffman Theorem (Proposition III.5.1 in [3]) to $-iS_3$, we see that
the singular values of $S_3$ are all greater than $v$, so that $|||S_3^{-1}|||_2 \leq \frac{1}{v}$. This shows that $|||T^{11}_n(z)|||_2 \leq 1/v$.

The same argument also yields $|||W^{11}_n(z)|||_2 \leq 1/v$.

Since $C^{-1} - D^{-1} = C^{-1}(D - C)D^{-1}$,

$$T^{11}_n(z) - W^{11}_n(z) = T^{11}_n(z) \left[ A_{12}(Z_{22} - z\text{Id})^{-1}A_{21} - E \left( A_{12}(Z_{22} - z\text{Id})^{-1}A_{21} \right) \right] W^{11}_n(z).$$

So it is clear that $E\left(|||T^{11}_n(z) - W^{11}_n(z)|||_2\right) \leq \frac{1}{v^2} \frac{R}{v}$. By Weyl’s majorant theorem ([9], Theorem II.3.6), we conclude that

$$E\left(|||\text{trace} \left( T^{11}_n(z) - W^{11}_n(z) \right) \right) \leq \frac{d}{v^2} \frac{R}{v}.$$ We clearly also have

$$E \left( |\text{trace} \left( T^{11}_n(z) - W^{11}_n(z) \right) | \right) \leq \frac{d}{v^2} \frac{R}{v}.$$ Let us now work on

$$T^{22}_n(z) = (Z_{22} - z\text{Id} + \frac{1}{z}A_{21}A_{12})^{-1},$$

the bottom-right $(n - 1)d \times (n - 1)d$ diagonal block of $(T_n - z\text{Id})^{-1}$.

Since $A_{21} = A'_{12}$ is $(n - 1)d \times d$, we see that $A_{21}A_{12}$ is a rank-at-most-$d$ matrix of size $(n - 1)d \times (n - 1)d$. The Sherman-Morrison-Woodbury formula ([35], p.19) gives, if $B = (C + \frac{1}{z}XX')$, with $C$ a $(n - 1)d \times (n - 1)d$ matrix and $X$ a $(n - 1)d \times d$ matrix,

$$B^{-1} = C^{-1} - C^{-1}X(z\text{Id}_d + X'C^{-1}X)^{-1}X'C^{-1}.$$ Therefore,

$$\text{trace} \left( B^{-1} \right) - \text{trace} \left( C^{-1} \right) = -\text{trace} \left( (z\text{Id}_d + X'C^{-1}X)^{-1}X'C^{-2}X \right).$$ Note that inside the trace on the right-hand side we have two $d \times d$ matrix. For us $C = Z_{22} - z\text{Id}$ and $X = A_{21}$. We have seen above that

$$|||(z\text{Id}_d + A_{12}C^{-1}A_{21})^{-1}|||_2 \leq \frac{1}{v}.$$ Hence, using our assumption on $E \left( |||A_{12}(\Gamma - z\text{Id})^{-2}A_{12}' - E \left( A_{12}(\Gamma - z\text{Id})^{-2}A_{12}' \right) |||_2 \right)$ as well as Weyl’s majorant theorem, we have

$$E \left( |\text{trace} \left( (z\text{Id}_d + A_{12}C^{-1}A_{21}^{-1} \left[ A_{21}C^{-2}A_{21} - E \left( A_{12}C^{-2}A_{21} \right) \right] \right) \right) \right) \leq \frac{d}{v^2} \frac{R}{v^2}.$$ Let us call $\Xi = E \left( A_{12}C^{-2}A_{21} \right)$. Of course,

$$\left[(z\text{Id}_d + A_{12}C^{-1}A_{21})^{-1} - (z\text{Id}_d + E \left( A_{12}C^{-1}A_{21} \right))^{-1} \right]|\Xi| \leq K(z) \frac{\text{Id}_d + E \left( A_{12}C^{-1}A_{21} \right)}{v^2}.$$ Therefore, since we have assumed that $|||\Xi|||_2 \leq K(z)$, we have

$$||| \left[(z\text{Id}_d + A_{12}C^{-1}A_{21})^{-1} - (z\text{Id}_d + E \left( A_{12}C^{-1}A_{21} \right))^{-1} \right]|\Xi|||_2 \leq \frac{K(z)}{v^2} |||E \left( A_{12}C^{-1}A_{21} \right) - A_{12}C^{-1}A_{21}|||_2.$$ We have established that

$$E \left( |\text{trace} \left( \left[(z\text{Id}_d + A_{12}C^{-1}A_{21})^{-1} - (z\text{Id}_d + E \left( A_{12}C^{-1}A_{21} \right))^{-1} \right]|\Xi| \right) \right) \leq \frac{d}{v^2} \frac{K(z)}{v}.$$ So we finally conclude that, if

$$\Delta = (z\text{Id}_d + A_{12}C^{-1}A_{21})^{-1}A_{12}C^{-2}A_{21} - (z\text{Id}_d + E \left( A_{12}C^{-1}A_{21} \right))^{-1}E \left( A_{12}C^{-2}A_{21} \right),$$

we have

$$E \left( |\text{trace} \left( \Delta \right) | \right) \leq \left( \frac{d}{v^2} + \frac{dK(z)}{v^2} \right) \frac{R}{v}.$$
We are now in position to state our “strip-replacement” theorem.

**Theorem 2.4.** Let us consider the $N \times N$ real symmetric matrices

$$ T_n(A) = \begin{pmatrix} 0_{d \times d} & A_{12} \\ A_{21} & Z_{22} \end{pmatrix}, \text{ and } T_n(B) = \begin{pmatrix} 0_{d \times d} & B_{12} \\ B_{21} & Z_{22} \end{pmatrix}, $$

where, of course, $A_{21} = A_{12}^\prime$ and $B_{21} = B_{12}^\prime$, and $Z_{22}$ is a deterministic symmetric matrix. Suppose $A_{12}$ and $B_{12}$ satisfy the assumptions of Lemma 2.1. Suppose further that for any deterministic vector $u$,

$$ E \left( (A_{12}u)(A_{12}u)^\prime \right) = E \left( (B_{12}u)(B_{12}u)^\prime \right). $$

Then

$$ \left| E \left( \text{trace} \left( (T_n(A) - z \text{Id})^{-1} \right) - \text{trace} \left( (T_n(B) - z \text{Id})^{-1} \right) \right) \right| \leq d(2 + K(z)) \frac{2R}{v^3}. \quad (9) $$

The same is true when $Z_{22}$ is assumed to be random but independent of $A_{12}$ and $B_{12}$.

**Proof.** The proof is essentially immediate once we realize that the assumption

$$ E \left( (A_{12}u)(A_{12}u)^\prime \right) = E \left( (B_{12}u)(B_{12}u)^\prime \right) $$

implies that, in the notation of the previous Lemma, the deterministic approximating quantity

$$ \text{trace} \left( W_{11}^n(z) \right) + \mathcal{L}(z) $$

takes the same value whether the expectations involve $A_{12}$ or $B_{12}$.

The case of random $Z_{22}$ is treated by conditioning on $Z_{22}$ and getting an upper bound on the quantities we care about that does not depend on $Z_{22}$. \hfill \Box

### 2.2.2 Main result

Let $M_n$ be a $N \times N$ matrix, where $N = nd$. We write

$$ M_n = \begin{pmatrix} M_n(1) \\ M_n(2) \\ \vdots \\ M_n(n) \end{pmatrix}, \text{ where } M_n(i) \in \mathbb{R}^{d \times N}. $$

We refer to the matrices $M_n(i)'s$ as block rows or strips. We further write the $d \times N$ matrix $M_n(i)$ as

$$ M_n(i) = \begin{pmatrix} \mathcal{R}_n(i) \\ \mathcal{M}_n(i) \end{pmatrix}_{i \times d \times (n-i) \times d}, \text{ where } \mathcal{R}_n(i) \in \mathbb{R}^{d \times (id)}. $$

$\mathcal{M}_n(i)$ is the $d \times (n-i)d$ strip that is on the $i$-th block row above the block diagonal of $M_n$. We call $\tilde{M}_n(i) = [0_{d \times (id)}M_n(i)]$. $\tilde{M}_n(i)$ is clearly a $d \times N$ matrix.

**Assumption B1** Let $z \in \mathbb{C}^+$ a complex number with positive imaginary part, $0 < v = \text{Im}[z]$. If $\Gamma$ is a real, symmetric, deterministic matrix, we assume that

$$ \frac{1}{nd} E \left( \|M_n(i)(\Gamma - z \text{Id})^{-1}M_n(i)^\prime - E \left( \tilde{M}_n(i)(\Gamma - z \text{Id})^{-1}\tilde{M}_n(i)^\prime \right) \|_2 \right) \leq \frac{R_i}{v}, \quad (\text{Assumption-B1.1}) $$

where $R_i \in \mathbb{R}_+$ is independent of $\Gamma$.

We also assume that $\tilde{M}_n(i)$ is such that there exists a function $K_i$ of $z$ such that

$$ \| \frac{1}{nd} E \left( \tilde{M}_n(i)(\Gamma - z \text{Id})^{-2}\tilde{M}_n(i)^\prime \right) \|_2 \leq K_i(z). \quad (\text{Assumption-B1.2}) $$

We assume that $K_i$ is bounded in $i$.

We finally assume that

$$ \frac{1}{nd} E \left( \|\tilde{M}_n(i)(\Gamma - z \text{Id})^{-2}\tilde{M}_n(i)^\prime - E \left( \tilde{M}_n(i)(\Gamma - z \text{Id})^{-2}\tilde{M}_n(i)^\prime \right) \|_2 \right) \leq \frac{R_i}{v^2}, \quad (\text{Assumption-B1.3}) $$

where $R_i \in \mathbb{R}_+$ is independent of $\Gamma$.  

10
About Assumption-B1.1 Note that the matrix \( \tilde{M}_n(i)(\Gamma - z\text{Id})^{-1}\tilde{M}_n(i)' \) is \( d \times d \) and \( d \) is assumed to be fixed in our analysis. So if we call \( v_{k,j}(i) \) the \((k,j)\) entry of
\[
\frac{1}{nd} \left( \tilde{M}_n(i)(\Gamma - z\text{Id})^{-1}\tilde{M}_n(i)' - \mathbf{E} \left( \tilde{M}_n(i)(\Gamma - z\text{Id})^{-1}\tilde{M}_n(i)' \right) \right),
\]
a simple way to check that (Assumption-B1.1) holds for models under consideration is to verify that \( \sup_{1 \leq k,j \leq d} \mathbf{E} (|v_{k,j}(i)|) \leq R_i/v \); in this case (Assumption-B1.1) holds with \( R_i \) replaced by \( dR_i \), since the operator norm of a symmetric matrix is smaller than the maximum \( l_1 \) norm of its rows [B5, p.313].

About Assumption-B1.2 We note that the assumption about \( K_i \) is easily satisfied: for instance, if we assume that there exists a constant \( C_i \) such that for any deterministic unit vector \( u \), \( ||\mathbf{E} \left( \tilde{M}_n(i)uu'\tilde{M}_n(i)' \right)||_2 \leq C_i \), then after diagonalizing \( \Gamma \), we see that, if \( \text{Im}[z] = v \), \( K_i(z) = C_i/v^2 \) is a valid choice.

We are now in position of stating our main theorem.

**Theorem 2.5.** Let \( M_n \) be an \( N \times N \) block matrix with random block-rows satisfying Assumption B1. Assume furthermore that \( \mathbf{E} (M_n) = 0 \) and that its block diagonal is 0. Call \( m_n(z) \) the Stieltjes transform of \( M_n/\sqrt{N} \), i.e.
\[
m_n(z) = \frac{1}{N} \text{trace} \left( \frac{M_n}{\sqrt{N}} - z\text{Id} \right)^{-1}.
\]
Assume that the block rows above the diagonal (i.e. the \( M_n(i) \) in our notation) are independent.

Let \( GM_n \) be a block matrix with Gaussian random blocks, with mean 0. Call \( gm_n(z) \) the Stieltjes transform of \( GM_n/\sqrt{N} \).

Call \( M_n(i) \) the \( d \times (n-i)d \) random matrix corresponding to the \( i \)-th block row of \( M_n \) above the diagonal. Call \( \tilde{M}_n(i) = [0_{d \times (id)}]M_n(i) \). Call \( G\tilde{M}_n(i) \) the corresponding \( d \times nd \) matrix for the matrix \( GM_n \) and suppose that Assumption B1 is satisfied for it, too.

Assume that for all \( 1 \leq i \leq n \) and for any deterministic (unit) vector \( u \),
\[
\mathbf{E} \left( \tilde{M}_n(i)uu'\tilde{M}_n(i)' \right) = \mathbf{E} \left( G\tilde{M}_n(i)uu'G\tilde{M}_n(i)' \right).
\]
Then
\[
|\mathbf{E} (m_n(z) - gm_n(z))| \leq \frac{1}{nd} \sum_{i=1}^{n} R_i g(z, K_i),
\]
where \( g(z, K_i) = 2d(2+K(z))d^{1/2} \). In particular, if \( \sum R_i/n \to 0 \) as \( n \to \infty \), the limiting spectral distribution of \( M_n \) is the same as that of \( GM_n \).

**Proof.** We use the Lindeberg method, where we replace the \( i \)-th block row and column by a Gaussian version satisfying Equation (10).

We call \( E_i \) the \( d \times N \) matrix with \( E_i(j, k) = \delta_{k,(i-1)d+j} \). Recall that the block diagonal of \( M_n \) is 0. We note that
\[
M_n = \sum_{i=1}^{n} E_i' \tilde{M}_n(i) + \tilde{M}_n(i)'E_i.
\]

We call, for \( 1 \leq k \leq n-1 \),
\[
I_n(k) = \sum_{i=1}^{k} \left[ E_i' \tilde{M}_n(i) + \tilde{M}_n(i)'E_i \right] + \sum_{i=k+1}^{n} \left[ E_i' G\tilde{M}_n(i) + G\tilde{M}_n(i)'E_i \right],
\]
and extend the definition for \( k = 0 \) and \( k = n \) with \( I_n(0) = GM_n \) and \( I_n(n) = M_n \).

Clearly,
\[
\text{trace} \left( \frac{G\tilde{M}_n}{\sqrt{N}} - z\text{Id} \right)^{-1} - \text{trace} \left( \frac{M_n}{\sqrt{N}} - z\text{Id} \right)^{-1} = \sum_{k=0}^{n-1} \text{trace} \left( \frac{I_n(k)}{\sqrt{N}} - z\text{Id} \right)^{-1} - \text{trace} \left( \frac{I_n(k+1)}{\sqrt{N}} - z\text{Id} \right)^{-1}.
\]

11
Therefore,

\[ |\mathbf{E}(\mathbf{g}\mathbf{m}_n(z)) - \mathbf{E}(\mathbf{m}_n(z))| \leq \frac{1}{nd} \sum_{k=0}^{n-1} \mathbf{E} \left( \text{trace} \left( \left( \frac{I_n(k)}{\sqrt{N}} - z\mathbf{I} \right)^{-1} \right) - \text{trace} \left( \left( \frac{I_n(k+1)}{\sqrt{N}} - z\mathbf{I} \right)^{-1} \right) \right). \]

Now the conditions of Theorem 2.4 are satisfied, so Equation (9) applies to

\[ \mathbf{E} \left( \text{trace} \left( \left( \frac{I_n(k)}{\sqrt{N}} - z\mathbf{I} \right)^{-1} \right) - \text{trace} \left( \left( \frac{I_n(k+1)}{\sqrt{N}} - z\mathbf{I} \right)^{-1} \right) \right), \]

so we have established Equation (11).

\[ \square \]

### 2.3 Application to block random matrices with independent block entries

To show that our theory applies, we just need to verify that Assumptions B1 is satisfied. Let us translate it, in the context of block matrices, to easier-to-verify assumptions about the block matrices constituting the block entries.

We remind the reader that we assume that \( \mathbf{E}(\mathbf{M}_n) = 0 \) and hence the same is true for the random block matrices we are dealing with.

#### 2.3.1 On Assumption-B1.2

In the case of block random matrices with independent block entries, we write

\[ \forall i, \mathbf{M}_n(i) = \left( \frac{\mathbf{M}_n[i, 1]}{d} \frac{\mathbf{M}_n[i, 2]}{d} \ldots \frac{\mathbf{M}_n[i, n]}{d} \right), \]

where \( \mathbf{M}_n[i, k] \) is the \( k \)-th \( d \times d \) block matrix on the \( i \)-th strip of \( \mathbf{M}_n \).

We now present an easy-to-verify condition to make sure that Assumption-B1.2 is satisfied in certain models of interest.

**Lemma 2.2.** Suppose the matrix \( \mathbf{M}_n \) is constituted of \( d \times d \) independent blocks and \( \mathbf{E}(\mathbf{M}_n) = 0 \). Call \( \mathbf{S}_{mj}^{i}[k, j] \) the (cross-) covariance matrix of the \( j \)-th row and \( k \)-th row of the \( m \)-th block matrix on the \( i \)-th strip of \( \mathbf{M}_n \) (i.e. \( \mathbf{M}_n[i, m] \)). If there exists \( C \) such that \( |||\mathbf{S}_m[i, j, k]||| \leq C \), then for any real, symmetric, deterministic matrix \( \mathbf{\Gamma} \),

\[ \left| \frac{1}{nd} \mathbf{E} \left( \left( \tilde{\mathbf{M}}_n(i)(\mathbf{\Gamma} - z\mathbf{I})^{-2}\tilde{\mathbf{M}}_n(i)' \right) \right) \right|_2 \leq \frac{Cd}{v^2}. \]

In other words, Assumption-B1.2 is satisfied with \( K_1(z) = Cd/v^2 \).

**Proof.** The matrix \( \tilde{\mathbf{M}}_n(i) \) is constituted of \( n \) independent \( d \times d \) matrices, \( i \) of them being \( 0_{d\times d} \). Let us call \( \mathbf{A} \) a generic \( d \times N \) matrix, constituted of \( d \times d \) independent blocks with \( \mathbf{E}(\mathbf{A}) = 0 \). We call \( \mathbf{A}[i] \) its \( i \)-th \( d \times d \) block i.e

\[ \mathbf{A} = \left( \frac{\mathbf{A}[1]}{d} \frac{\mathbf{A}[2]}{d} \ldots \frac{\mathbf{A}[n]}{d} \right). \]

If \( \mathbf{T} \) is a deterministic \( N \times N \) matrix, we call \( \mathbf{T}[i, j] \) its \( (i, j) \)-th \( d \times d \) block. We have

\[ \mathbf{A}\mathbf{A}' = \sum_{1 \leq i, j \leq n} \mathbf{A}[i]\mathbf{T}[i, j]\mathbf{A}[j]' . \]

By independence of \( \mathbf{A}[i] \) and \( \mathbf{A}[j] \) when \( i \neq j \), \( \mathbf{E}(\mathbf{A}[i]\mathbf{T}[i, j]\mathbf{A}[j]') = 0_{d\times d} \) when \( i \neq j \). Hence,

\[ \mathbf{E}(\mathbf{A}\mathbf{A}') = \sum_{1 \leq i \leq n} \mathbf{E}(\mathbf{A}[i]\mathbf{T}[i, i]\mathbf{A}[i]') . \]
Note that if we can bound uniformly $\mathbf{E} (|||A[i]T[i, i]A[i]'|||_2)$ by $K(z)$, then we have
\[
\frac{1}{nd} |||\mathbf{E} (ATA')|||_2 \leq \frac{K(z)}{d} .
\]

So let us focus on the $d \times d$ matrix $Q[i] = \mathbf{E} (A[i]T[i, i]A[i]')$. Let us call $r_k$ the $k$-th row of $A[i]$. The $k, j$ entry of $Q[i]$ is just
\[
Q[i](k, j) = \mathbf{E} (r_k T[i, j] r'_k) = \text{trace} (T[i, i] \mathbf{E} (r'_k r_k)) = \text{trace} (T[i, i] \mathcal{S}_i [k, j]) ,
\]
where $\mathcal{S}_i [k, j] = \mathbf{E} (r'_k r_k)$ is the $d \times d$ cross-covariance matrix between the $j$-th and the $k$-th rows of $A[i]$. Assume further that, for some $\epsilon > 0$, $\mathcal{S}_i [k, j] = \mathbf{E} (A[i]'A[i])$. Let us call $r_k$ the $k$-th row of $A[i]$. The $k, j$ entry of $Q[i]$ is just
\[
Q[i](k, j) = \mathbf{E} (r_k T[i, j] r'_k) = \text{trace} (T[i, i] \mathbf{E} (r'_k r_k)) = \text{trace} (T[i, i] \mathcal{S}_i [k, j]) ,
\]
where $\mathcal{S}_i [k, j] = \mathbf{E} (r'_k r_k)$ is the $d \times d$ cross-covariance matrix between the $j$-th and the $k$-th rows of $A[i]$. Suppose that $|||T[i, i]|||_2 \leq \frac{1}{n}$ and $|||\mathcal{S}_i [k, j]|||_2 \leq C$, where $C$ is a constant independent of $i, j, k$. Then,
\[
|Q[i](j, k)| \leq \frac{Cd}{v}, \forall (j, k) .
\]
Therefore,
\[
|||Q[i]|||_2 \leq \frac{Cd^2}{v} .
\]
We note that if $T = (\Gamma - z \text{Id})^{-2}$, where $\Gamma$ is real symmetric, then $|||T|||_2 \leq 1/v^2$, if $v = \text{Im} [z]$.

The Lemma is shown.

\[
\tag{2.3.2}
\text{Concentration of quadratic forms in block rows}
\]

We give sufficient conditions for [Assumption-B1.1] and [Assumption-B1.3] to be satisfied.

**Lemma 2.3.** Let us call $Q = ATA'$, where $A$ is a $d \times (nd)$ real random matrix composed of independent $d \times d$ blocks, denoted by $A[i]$. We assume that $\mathbf{E} (A) = 0$. $T$ is a $nd \times nd$ symmetric matrix with complex entries with $|||T|||_2 \leq \frac{1}{n}$.

Denote by $\mathcal{S}_i [k, k]$ the covariance matrix of the $k$-th row of the matrix $A[i]$. Assume that there exists $C > 0$ such that
\[
\sup_{1 \leq i \leq n} \sup_{1 \leq k \leq d} |||\mathcal{S}_i [k, k]|||_2 \leq C .
\]
Assume further that, for some $\epsilon > 0$, the rows of $A[i]$ have uniformly bounded $2 + 2\epsilon$-th moments, for all $1 \leq i \leq n$. Then, when $d$ is fixed, we have
\[
|||\mathbf{E} (Q - \mathbf{E} (Q))|||_2 = O(\frac{n^{1/(1+\epsilon)} \land n^{1/2}}{v}) .
\]

**Proof.** Let us call $Q(j, k)$ the $(j, k)$ entry of $Q$. Since $d$ is held fixed, to show the result, it is enough to show that
\[
\forall 1 \leq j, k \leq d, \mathbf{E} (||Q(j, k) - \mathbf{E} (Q(j, k))||) = O(\frac{n^{1/(1+\epsilon)} \land n^{1/2}}{v}) .
\]
Let us call $r_j$ the $j$-th row of $A$. Clearly,
\[
Q(j, k) = r_j T r'_k = \frac{1}{4} ((r_j + r_k) T (r_j + r_k)' - (r_j - r_k) T (r_j - r_k)) .
\]
Hence, to understand $Q$, we simply need to understand forms of the type
\[
f(r) = r' Tr
\]
where $r \in \mathbb{R}^{nd}$ is a random vector composed of independent blocks of size $d$. Indeed, given the structure we have assumed for $A$, it is clear that both $r_j + r_k$ and $r_j - r_k$ are vectors composed of independent blocks
of length \( d \). (Our assumptions about \( S_i[k,k] \) implies that the same assumptions is true for all the \( r \)'s we will be looking at, with a upper bound less than \( 2C \).) In other words,

\[
\begin{pmatrix}
  r_1 \\
r_2 \\
\vdots \\
r_n
\end{pmatrix},
\]

where \( r_i \in \mathbb{R}^d \) are independent of each other. We call \( \Sigma[i,i] \) the covariance matrix of \( r_i \).

We have of course, if \( T[i,j] \) denotes the \((i,j)\)-th \( d \times d \) block of \( T \),

\[
f(r) = \sum_{i,j} r'_i T[i,j] r_j \triangleq \sum_i r'_i T[i,i] r_i + \mathcal{R}.
\]

- **On \( \text{var} (\mathcal{R}) \)** By definition,

\[
\mathcal{R} = \sum_{i \neq j} r'_i T[i,j] r_j.
\]

Since \( r_i \) and \( r_j \) are independent when \( i \neq j \), we see that \( \mathbb{E} (\mathcal{R}) = 0 \). So

\[
\text{var} (\mathcal{R}) = \mathbb{E} (\mathcal{R} \mathcal{R}^*) = \sum_{(i \neq j), (k \neq l)} \mathbb{E} (r'_i T[i,j] r_j r'_k T^*[k,l] r_l).
\]

If one of the indices \((i,j,k,l)\) appears exactly once, \( \mathbb{E} (r'_i T[i,j] r_j r'_k T^*[k,l] r_l) = 0 \), by independence of the \( r_j \)'s and the fact that they all have mean 0. Now since each index appears at most once in each pair, we see that each index appear at most twice among the four indices. It is therefore clear that

\[
\text{var} (\mathcal{R}) = \sum_{i \neq j} \mathbb{E} (r'_i T[i,j] r_j r'_j T^*[j,i] r_j) + \mathbb{E} (r'_i T[i,j] r_j r'_j T^*[j,i] r_j) \geq 0.
\]

Of course, by independence,

\[
\mathbb{E} (r'_i T[i,j] r_j r'_j T^*[j,i] r_j) = \mathbb{E} (r'_i T[i,j] \Sigma[j,j] T^*[j,i] r_i) = \text{trace} (T[i,j] \Sigma[j,j] T^*[j,i] \Sigma[i,i]) \geq 0.
\]

Consider the matrix \( D_{\Sigma} \) which is block-diagonal with \( i \)-th diagonal block \( \Sigma[i,i] \). We note that

\[
\text{trace} (T D_{\Sigma} T^* D_{\Sigma}) = \sum_{i,j} \text{trace} (T[i,j] \Sigma[j,j] T^*[j,i] \Sigma[i,i]) \geq 0.
\]

Note further that \( \text{trace} (T[i,i] \Sigma[i,i] T^*[i,i] \Sigma[i,i]) \geq 0 \). So we conclude that

\[
\sum_{i \neq j} \mathbb{E} (r'_i T[i,j] r_j r'_j T^*[j,i] r_j) \leq \text{trace} (T D_{\Sigma} T^* D_{\Sigma}) \leq 4N \frac{C^2}{u^2}.
\]

The same argument works for \( \sum_{i \neq j} \mathbb{E} (r'_i T[i,j] r_j r'_j T^*[i,j] r_j) \) and we conclude that

\[
\text{var} (\mathcal{R}) \leq 8N \frac{C^2}{u^2}.
\]

This naturally implies that

\[
\mathbb{E} (|\mathcal{R} - \mathbb{E} (\mathcal{R})|) \leq 2\sqrt{2} N C / u.
\]

Note that this bound works under the assumption that \( r_i \)'s have uniformly bounded covariances, i.e only 2 moments.

- **On the convergence of \( D_1(r) = \sum_i r'_i T[i,i] r_i \)** Note that

\[
D_1(r) - \mathbb{E} (D_1(r)) = \sum_i X_i,
\]
where $X_i$ are independent and mean 0 random variables in $L_{1+\epsilon}$. Using the Marcinkiewicz-Zygmund inequality ([IO, p. 386]), we see that for any $\epsilon > 0$ there exists $B_{1+\epsilon}$ such that

$$E \left( |D_1(r) - E (D_1(r))|^{1+\epsilon} \right) \leq B_{1+\epsilon} E \left( \sum_i X_i^2 \right)^{(1+\epsilon)/2}.$$ 

We have $(\sum_i X_i^2)^{p/2} = (\sum_i (|X_i|^p)^{2/p})^{p/2} = \|Y\|_{2/p}$, where $Y_i = |X_i|^p$. For $p \in [1, 2]$, we have $2/p \geq 1$, so $\|Y\|_{2/p} \leq \|Y\|_1$. Therefore, when $p \in [1, 2]$,

$$\left( \sum_i X_i^2 \right)^{p/2} \leq \sum_i |X_i|^p.$$ 

Hence,

$$E \left( |D_1(r) - E (D_1(r))|^{1+\epsilon} \right) \leq B_{1+\epsilon} E \left( \sum_i |X_i|^{1+\epsilon} \right).$$

We conclude that when $r_i$’s have $2+2\epsilon$ moments with $0 < \epsilon \leq 1$, we have

$$E \left( |D_1(r) - E (D_1(r))| \right) \leq \left[ E \left( |D_1(r) - E (D_1(r))|^{1+\epsilon} \right) \right]^{1/(1+\epsilon)} \leq C_n n^{1/(1+\epsilon)}\frac{1}{v}.$$

**Conclusion**

We can finally conclude that

$$E \left( \mathcal{Q} - E (\mathcal{Q}) \right) = O(n^{1/(1+\epsilon)} \wedge n^{1/2}).$$

\[\square\]

**Corollary 2.3.** Suppose the symmetric matrix $M_n$ is such that its $i$-th (d-high) block row, $M_n(i)$, is composed of independent $d \times d$ matrices. Denote by $S_{n}[k,k]$ the covariance matrix of the $k$-th row of $M_n[i,m]$. Assume that there exists $C > 0$ such that

$$\sup_{1 \leq i \leq n} \sup_{i \leq m \leq n} \sup_{1 \leq k \leq d} |||S_{n}[k,k]|||_2 \leq C.$$ 

Assume further that the rows of all the $d \times d$ block matrices above the block diagonal of $M_n$ have uniformly bounded ($2+2\epsilon$)-th moments ($\epsilon > 0$) and that $E (M_n) = 0$.

Then Assumption-B1.1 and Assumption-B1.3 hold with $R_i = O(n^{-\epsilon/(1+\epsilon)} \wedge n^{-1/2})$.

The proof is an immediate application of Lemma 2.3. Note that “padding” a block row with 0 block matrices does not change anything to our analysis: just consider the 0 block as a random variables with 0-covariance.

**Corollary 2.4.** Suppose the symmetric matrix $M_n$ is such that $M_n(i)$ is composed of independent $d \times d$ matrices. Suppose the entries of $M_n$ are either bounded or Gaussian. Denote by $S_{n}[k,k]$ the covariance matrix of the $k$-th row of $M_n[i,m]$. Assume that there exists $C > 0$ such that

$$\sup_{1 \leq i \leq n} \sup_{i \leq m \leq n} \sup_{1 \leq k \leq d} |||S_{n}[k,k]|||_2 \leq C.$$ 

Then Assumption-B1.1 and Assumption-B1.3 hold with $R_i = O(n^{-1/2})$.

When the entries of the matrix are Gaussian or bounded, the $2+2\epsilon$-th moment condition is automatically satisfied when our condition on covariance matrices is satisfied.
2.3.3 On $\mathbb{E} \left( \widetilde{M}_n(i)uu'\widetilde{M}_n(i)' \right)$

The following fact will be helpful in establishing equivalence between models from a limiting spectral distribution point of view.

**Fact 2.1.** Let $\widetilde{M}_n^{(1)}$ and $\widetilde{M}_n^{(2)}$ be two random $d \times N$ strips with mean 0. Let us call $C^{(1)}[j,k]$ the cross-covariance between the $j$-th and the $k$-th row of $\widetilde{M}_n^{(1)}$ and $C^{(2)}[j,k]$ the cross-covariance between the $j$-th and the $k$-th row of $\widetilde{M}_n^{(2)}$. Suppose that

$$\forall (j,k), C^{(1)}[j,k] + (C^{(1)}[j,k])' = C^{(2)}[j,k] + (C^{(2)}[j,k])'.$$

Then, if $u$ is any deterministic vector,

$$\mathbb{E} \left( \widetilde{M}_n^{(1)}uu'\widetilde{M}_n^{(1)'} \right) = \mathbb{E} \left( \widetilde{M}_n^{(2)}uu'\widetilde{M}_n^{(2)'} \right).$$

**Proof.** If $r_j$ denotes the $j$-th row of $\widetilde{M}_n$, we have, for the $(k,j)$-th entry of the matrix $\mathbb{E}Q = \mathbb{E} \left( \widetilde{M}_n uu' \widetilde{M}_n' \right)$,

$$\mathbb{E}Q(k,j) = \mathbb{E} \left( r_k uu' r_j' \right) = \text{trace} \left( uu' \mathbb{E} \left( r_j' r_k \right) \right).$$

Let us call $C[j,k]$ the cross-covariance matrix $C[j,k] = \mathbb{E} \left( r_j' r_k \right)$. Since trace$(AB) = $ trace$(BA)$ and trace$(A) = $ trace$(A')$, we have

$$\mathbb{E}Q(k,j) = \text{trace} \left( uu' C[j,k] uu' \right) = \text{trace} \left( uu' C[j,k] \right) = \text{trace} \left( uu' \frac{C[j,k] + C'[j,k]}{2} \right).$$

The result is established. \(\square\)

**A remark on the case of anti-symmetric cross-covariances** We now assume that if $j \neq k$, the cross-covariance matrix $C[j,k] = \mathbb{E} \left( r_j' r_k \right)$ is anti-symmetric. In this case,

$$C[j,k] + C'[j,k] = 0.$$  

This means in particular that if $\widetilde{M}_n^{(1)}$ is such that its rows have anti-symmetric cross-covariance, we can create a “good” $\widetilde{M}_n^{(2)}$ by picking independent vectors matching the covariance of each row of $\widetilde{M}_n^{(1)}$. This way $\widetilde{M}_n^{(2)}$ clearly has anti-symmetric cross-covariance between its rows (indeed the cross-covariance is 0 for all pairs of distinct rows); but each row of $\widetilde{M}_n^{(2)}$ has by construction the same covariance as the corresponding row of $\widetilde{M}_n^{(1)}$. So we have

$$\mathbb{E} \left( \widetilde{M}_n^{(1)}uu'\widetilde{M}_n^{(1)'} \right) = \mathbb{E} \left( \widetilde{M}_n^{(2)}uu'\widetilde{M}_n^{(2)'} \right).$$

**The case of block matrices with mean 0** We now assume the $d \times N$ matrix $\widetilde{M}_n$ is made of $n$ $d \times d$ independent blocks. In that case, $r_j$ and $r_k$ are composed of independent blocks, so $\mathbb{E} \left( r_j' r_k \right)$ is block diagonal. The $l$-th $d \times d$ block on the diagonal is just the cross covariance between the $j$-th and $k$-th row of the $l$-th $d \times d$ block matrix in $\widetilde{M}_n$, since $\mathbb{E} \left( \widetilde{M}_n \right) = 0$. So our assumptions about the cross-covariance of the rows of $\widetilde{M}_n$ in Fact 2.1 can be replaced by assumptions concerning the cross-covariance of the rows of the block matrices making up $\widetilde{M}_n$ and the same result holds.

2.4 Applications

**Definition 1 (σ-Simple Structure).** Let $B$ be a $d \times d$ random matrix. Call $\{r_i\}_{i=1}^d$ its rows. We say that the random matrix $B$ has σ-simple structure if and only if

1. the entries of $B$ have $2 + \epsilon$ moments for some $\epsilon > 0$. 

16
2. if \( j \neq k \), \( \mathbf{E} \left( r_j^* r_k \right) \) is anti-symmetric with \( \| \mathbf{E} \left( r_j^* r_k \right) \|_2 \leq C, \ C > 0. \)

3. for all \( j \), \( \mathbf{E} \left( r_j^* r_j \right) = \sigma^2 \text{Id}_d. \)

The following theorem explains the spectral distributions we saw in a number of our numerical investigations.

**Theorem 2.6.** Suppose \( M_n \) is a Hermitian \( N \times N \) matrix with independent \( d \times d \) block entries above the block diagonal. Let \( DM_n \) be the block-diagonal of \( M_n \). Suppose that \( \| DM_n \|_2 / \sqrt{N} = o_p(1) \) and that \( \text{rank} \left( \mathbf{E} \left( M_n - DM_n \right) \right) = o(N) \).

Suppose the off-diagonal blocks of \( M_n \) have \( \sigma \)-simple structure with \( 2 + \epsilon \) moments and \( \sigma = 1 \). Suppose further that the cross-covariance between the rows of these off-diagonal blocks is uniformly bounded (i.e. independently of \( n \)).

Then the limiting spectral distribution of \( M_n / \sqrt{N} \) is the Wigner semi-circle law. The convergence happens a.s.

Since we are talking about sequences of random variables, it is important to specify how they are built. For our theorem to hold, we assume that the sequence of matrices \( M_n \) is constructed by bordering the matrix \( M_{n-1} \) with an independent block matrix satisfying our assumptions.

**Proof.** A.s convergence of the spectral distribution is an immediate consequence of Theorem 2.3 and the Borel-Cantelli lemma. See [24] for details.

The fact that \( \| DM_n \|_2 / \sqrt{N} = o_p(1) \) guarantees that spectrally, \( M_n / \sqrt{N} \) and \( M_n(0) / \sqrt{N} = (M_n - DM_n) / \sqrt{N} \) are asymptotically equivalent as we discussed earlier.

Since \( \text{rank} \left( \mathbf{E} \left( M_n(0) \right) \right) = o(N) \), we see by Lemma B-1 that \( M_n(0) / \sqrt{N} \) and \( (M_n(0) - \mathbf{E} \left( M_n(0) \right) ) / \sqrt{N} \) are asymptotically spectrally equivalent. Indeed, the modulus of the difference of their Stieltjes transform at \( z \) is less than \( \text{rank} \left( \mathbf{E} \left( M_n(0) \right) \right) / (Nv) = o(1/v) \).

So the theorem holds for \( M_n \) provided we can prove it for \( (M_n(0) - \mathbf{E} \left( M_n(0) \right)) \); this latter matrix is still a matrix of independent blocks. However it has mean 0 and its block diagonal is zero. Our preliminary results have been obtained for matrices of this type.

Our assumptions guarantee that Assumption B1 is met for \( (M_n(0) - \mathbf{E} \left( M_n(0) \right)) \). Therefore we can apply Theorem 2.5.

Since the rows of the block matrices composing \( M_n \) have anti-symmetric cross-covariance, Fact 2.1 and the discussion that follows it show that in the “matching step” of Theorem 2.5, we can use Gaussian matrices with independent rows.

We note that a \( d \times d \) Gaussian matrix with i.i.d. \( \mathcal{N}(0, 1) \) entries has the same covariance for its rows as our initial model, under our assumptions, does. Assumption B1 is trivially met for a random block matrix with this Gaussian distribution on the blocks. Let us call the corresponding \( N \times N \) matrix \( GM_n \).

Theorem 2.5 guarantees that this Gaussian equivalent model has the same limiting spectral distribution as \( (M_n(0) - \mathbf{E} \left( M_n(0) \right)) \) and therefore the same is true for our initial sequence of matrices \( M_n \).

Note that \( GM_n / \sqrt{nd} \) is simply a scaled \( nd \times nd \) GOE (Gaussian orthogonal ensemble) matrix with \( d \times d \) block matrices on the diagonal removed. Call \( BD_n \) the block diagonal matrix of a random matrix drawn according to \( N \times N \) GOE. The norm of this block diagonal matrix \( BD_n \) is simply the maximum of the norms of the \( d \times d \) matrices on the block diagonal. For each such matrix, the operator norm is bounded by the largest row norm and hence by \( d \) times the largest absolute value of the elements of the matrix. Hence, the operator norm of the block diagonal matrix is less than \( d \) times the largest entry (in absolute value) of all these matrices. There are \( nd^2 \) such elements (corresponding to \( nd(d+1)/2 \) independent elements), with variance at most \( 2/(nd) \). Hence, using well-known properties of the maximum of independent Gaussian random variables (see e.g. [20] or [57], p.9), we have

\[
\frac{\| BD \|_2}{\sqrt{nd}} \leq \sqrt{2/(nd)} \sqrt{2 \log(nd^2)} \text{ a.s.}
\]
Clearly, the upper bound goes to 0. Therefore, \( GM_n / \sqrt{nd} \) has the same limiting spectral distribution as \( GOE_{nd} / \sqrt{nd} \), where \( GOE_{nd} \) is a \( N \times N \) random matrix sampled from GOE. Since the limiting spectral distribution of a GOE matrix is the Wigner semi-circle law, we have established the result. \( \Box \)

2.4.1 Case of \( O_d \) and \( SO_d \) sub-blocks

We consider in this paragraph the case of random matrices with independent random sub-blocks drawn at random uniformly (i.e, according to Haar measure) from \( O_d \) and \( SO_d \).

\( O_d \) case

**Fact 2.2.** Matrices drawn according to Haar measure on \( O_d \) have \( \sigma \)-simple structure with \( \sigma = \frac{1}{\sqrt{d}} \).

Therefore, if \( M_n \) is a Hermitian \( N \times N \) block random matrix, with \( N = nd \) and the blocks above the diagonal are drawn i.i.d according to Haar measure on \( O_d \), the limiting spectral distribution of \( M_n / \sqrt{N} \) is the Wigner semi-circle law, scaled by \( d^{-1/2} \) - provided the operator norm of the block diagonal of \( M_n \) is \( o_P(N^{1/2}) \).

**Proof.** We denote by \( O \) a random matrix drawn from \( O_d \) and by \( r_k \) its \( k \)-th row. Since for all \( k \), \( \| r_k \| = 1 \), the entries of \( O \) have infinitely many moments. When \( O \) is drawn according to Haar measure, we have by definition, for any given orthogonal \( O \),

\[
OO \overset{d}{=} O \overset{d}{=} O.
\]

Taking \( O \) to be a permutation matrix, we see that the columns and rows of \( O \) are exchangeable. Taking \( O_j = I_d - 2e_j e_j' \) (where \( e_j \) is the \( j \)-th canonical basis vector), we see that,

\[
\forall k \neq j, (r_k, r_j) \overset{d}{=} (r_k, -r_j).
\]

By Lemma 2.2, this naturally implies that

\[
\text{if } k \neq j \Rightarrow \mathbb{E} (r_j r_k') = 0.
\]

Since \( \| r_j \|^2 = 1 \) and the columns of \( O \) - and hence the entries of \( r_j \) - are exchangeable, we see that, if \( r_j(l) \) is the \( l \)-th entry of \( r_j \),

\[
\forall 1 \leq l \leq d, \quad \mathbb{E} (r_j(l)^2) = \frac{1}{d} \mathbb{E} (\| r_j \|^2) = \frac{1}{d}.
\]

Therefore, by Lemma 2.2,

\[
\forall j, \text{cov} (r_j) = \frac{1}{d} I_d.
\]

\( SO_d \) case

**Fact 2.3.** Matrices drawn according to Haar measure on \( SO_d \) have \( \sigma \)-simple structure with \( \sigma = \frac{1}{\sqrt{d}} \).

Therefore, if \( M_n \) is a Hermitian \( N \times N \) block random matrix, with \( N = nd \) and the blocks above the diagonal are drawn i.i.d according to Haar measure on \( SO_d \), the limiting spectral distribution of \( M_n / \sqrt{N} \) is the Wigner semi-circle law, scaled by \( d^{-1/2} \) - provided the operator norm of the block diagonal of \( M_n \) is \( o_P(N^{1/2}) \).

**Proof.** When \( O \) is drawn according to Haar measure on \( SO_d \), we have by definition, for any given \( O \in SO_d \),

\[
OO \overset{d}{=} O \overset{d}{=} O.
\]

• **Case \( d \geq 3 \)** Take \( O_{j,k,l} \) to encode the permutation \( (j, k, l) \rightarrow (l, j, k) \). Clearly \( O_{j,k,l} \) is in \( SO_d \). This shows that the columns and the rows of \( O \) are exchangeable. Let \( D_{j,k} \) be a diagonal matrix such that

\[
D_{j,k}(i, i) = \begin{cases} -1 & \text{if } i = j \text{ or } k \\ 1 & \text{otherwise.} \end{cases}
\]
Clearly $D_{j,k} \in SO_d$. Since $D_{j,k}O \overset{\mathcal{E}}{=} O$, we have
\[
\text{if } l \neq j \text{ and } l \neq k, (r_j, r_l) \overset{\mathcal{E}}{=} (-r_j, r_l).
\]
This shows that
\[
\mathbf{E} (r_j r'_l) = 0 \text{ if } l \neq j.
\]
The fact that $\text{cov} (r_j) = \text{Id}_d/d$ is proven as in the $O_d$ case.

- **Case $d = 2$**

It is clear geometrically that a matrix from $SO_2$, which is simply a planar rotation, can be written
\[
O_\theta = \begin{pmatrix} \cos(\theta) & -\sin(\theta) \\ \sin(\theta) & \cos(\theta) \end{pmatrix}.
\]
When drawn according to Haar measure, $\theta$ is uniform on $[0, 2\pi]$. (Geometrically, $\theta$ simply represents the angle by which the first canonical basis vector is rotated.) Hence,
\[
r'_1 r_2 = \begin{pmatrix} \cos(\theta) \sin(\theta) & \cos^2(\theta) \\ -\sin^2(\theta) & -\cos(\theta) \sin(\theta) \end{pmatrix}.
\]
So it is clear that when $O$ is drawn according to Haar measure, $\mathbf{E} (r'_1 r_2)$ is anti-symmetric.

The fact that $\text{cov} (r_i) = \frac{1}{2} \text{Id}_2$, $i = 1, 2$, is proven similarly.

\[\square\]

### 2.4.2 Measures on $GL_d$ and $SL_d$

We call $\mathcal{D}_k$ the set of diagonal matrices with $D_{ii} = 1$ except for exactly $k$ indices for which $D_{ii} = -1$.

**Fact 2.4.** Suppose that the $d \times d$ random matrix $B$ is such that it has the singular value decomposition
\[
B = UDV',
\]
where $U$, $D$ and $V$ are independent. Suppose further that $U$ and $V$ (which are of course orthonormal) have laws that are invariant under the action of any permutations and any diagonal matrix in $\mathcal{D}_m$, $1 \leq m \leq 2$.

Then, if the entries of $D$ have $2 + \epsilon$ moments, $B$ has $\sigma$-simple structure with
\[
\sigma^2 = \mathbf{E} \left( \text{trace} \left( B'B \right) \right) /d^2 = \mathbf{E} \left( \text{trace} \left( D^2 \right) \right) /d^2,
\]
if $m = 1$. If $m = 2$, the same statement is true provided $d \geq 3$.

**Proof.** Our assumptions on the entries of $D$ guarantee that the entries of $B$ have four moments.

Let $P$ be a permutation. Since $PU \overset{\mathcal{E}}{=} U$, it is clear that
\[
P B \overset{\mathcal{E}}{=} B.
\]
Hence the rows of $B$ are exchangeable. By a similar argument applied to $BP$, we see that the columns of $B$ are exchangeable.

Suppose $m = 1$. Let $D_j$ be in $\mathcal{D}_1$ with $D_j(j,j) = -1$. Since $D_j B \overset{\mathcal{E}}{=} B$, we see that for any $k \neq j$, $(r_j, r_k) \overset{\mathcal{E}}{=} (-r_j, r_k)$. When $m = 2$ and $d = 3$, we arrive at the same conclusion by using a matrix $D$ in $\mathcal{D}_2$ such that $D(j,j) = -1$, $D(k,k) = 1$ and $D(l,l) = -1$ for $l \neq k$ nor $j$. Such a matrix exists by assumption.

This implies that under our assumptions (see Lemma $\text{B-2}$ for details)
\[
\mathbf{E} \left( r_j r'_k \right) = 0, \text{ when } j \neq k, \text{ and } \mathbf{E} \left( r_j \right) = 0, \text{ for all } j.
\]

Now by exchangeability of the columns of $B$, we see that the diagonal of $\text{cov} (r_j)$ is proportional to $\text{Id}_p$, with proportionality constant $\sigma^2 = \mathbf{E} \left( \| r_j \|^2 \right) /d = \mathbf{E} \left( \text{trace} \left( B'B' \right) \right) /d^2$, the latter equality coming from exchangeability of the rows of $B$.

Suppose $m = 1$. Let $D_j$ be in $\mathcal{D}_1$ with $D_j(j,j) = -1$. Since $B D_j \overset{\mathcal{E}}{=} B$, we see that for any $k \neq j$, if $c_j$ denotes a generic column of $B$, $(c_j, c_k) \overset{\mathcal{E}}{=} (-c_j, c_k)$. This implies that the off-diagonal elements of $\text{cov} (r_k)$ are equal to 0. The case of $m = 2$ is treated as above and we have shown the lemma.

\[\square\]
We have the following corollaries.

**Corollary 2.5.** Suppose that $B$ is $d \times d$ with i.i.d $N(0,1)$ entries. Then $B \in GL_d$ with probability 1. Furthermore, it satisfies the assumptions of Fact 2.4 with $k = 1$.

The proof of the corollary is immediate - owing to elementary facts about Wishart matrices for instance (see [23] or [1]).

**The case of $SL_d$**

**Corollary 2.6.** Let $G$ be a $d \times d$ matrix with i.i.d $N(0,1)$ entries. Suppose

$$B = \frac{\tilde{G}}{|\det(G)|^{1/d}},$$

where $\tilde{G} = G / |\det(G)|$ if $d$ is odd and $\tilde{G} = G$ except that one column of $G$ - picked uniformly at random - is replaced by its opposite when $d$ is even.

Then $B \in SL_d$ for any $d \geq 1$ almost surely. Furthermore, for $d \geq 3$, $B$ satisfies the assumptions of Fact 2.4.

Therefore, if $M_n$ is a Hermitian $N \times N$ block random matrix, with $N = nd$ and the blocks above the diagonal are drawn i.i.d with the same law as $B$ and $d \geq 3$, the limiting spectral distribution of $M_n/\sqrt{N}$ is a scaled Wigner semi-circle law - provided the operator norm of the block diagonal of $M_n$ is $o_P(N^{1/2})$.

To get a finer understanding of this problem -especially for $d = 2$ - we compute the law of the squared singular values of $B$ in the Appendix of the ArXiv version of this paper. It turns out that in the case of $SL_2$, the largest eigenvalue of $B' B$ has Cauchy-like tail. Therefore, the matrix $D$ of Fact 2.4 does not have 2 moments. The situation is therefore structurally different from the other problems we have investigated in this paper. This is of course what we saw in the numerical part of this paper.

**Proof.** Let us write the singular value decompositions of $G$ and $B$ as $G = U(G) D(G) V(G)'$ and $B = U(B) D(B) V(G)'$. Since $d \geq 3$, we have seen that $G$ satisfies the assumptions on $U(G)$ and $V(G)$ we made in Fact 2.4. However, the $U(B)$ and $V(B)$ - though very closely related to $U(G)$ and $V(G)$ - are not independent anymore, since $B \in SL_d$ implies that $\det(U(B)V(B)) = 1$. Because $d \geq 2$, our arguments involving matrices in $D_2$ are still valid (matrices in $D_2$ have determinant 1). Our argument involving permutation now require permutation matrices containing a cycle - as we did in the case of $SO_d$. So our exchangeability arguments actually apply here and the only question we have to grapple with is that of the number of moments of the entries of $B$.

We recall that by using Bartlett’s decomposition, we see that, for independent $\chi_i^2$ random variables,

$$(\det(G))^2 = \prod_{i=1}^{d} \chi_i^2. $$

Recall that the density $f_1$ of $\chi_i^2$ is such that $f_1(x) \sim x^{-1/2}$ at 0 and $f_p$ the density of $\chi_p^2$ is such that $f_p(x) \sim x^{p/2-1}$ at 0. So we see that

$$E\left(\frac{1}{|\det(G)|^{p/d}}\right) < \infty$$

provided $E\left((\chi_1^2)^{-p/(2d)}\right) < \infty$ i.e 1/2 + $p/2d < 1$ or $d > p$. By Holder’s inequality, if $p \geq 1$ and $q = p/(p-1)$,

$$E\left(|B_{i,j}|^k\right) \leq E\left(|G_{i,j}|^{kq}\right)^{1/q} E\left(|\det(G)|^{-kp/d}\right)^{1/p}. $$

So the entries of $B$ have $k$ moments provided $d > kp$ for some $p > 1$. In other words, if $k < d$, the entries of $B$ have $k$ moments.

We conclude that when $d \geq 3$, the assumptions of Fact 2.4 are satisfied.
Figure 1: Histogram of the eigenvalues of $R_{SO(d),n,Haar}$ with entries sampled from $SO(d)$ when $(n,d) = (1000,2)$ (left) and $(1000,3)$ (left middle) and QQplot of the eigenvalues of $R_{SO(d),n,Haar}$ versus the eigenvalues of symmetric Gaussian random matrix of size $dn \times dn$ when $(n,d) = (1000,2)$ (right middle) and $(1000,3)$ (right).

Figure 2: Histogram of the eigenvalues of $R_{O(d),n,Haar}$ with entries sampled from $O(d)$ when $(n,d) = (1000,2)$ (left) and $(1000,3)$ (left middle) and QQplot of the eigenvalues of $R_{O(d),n,Haar}$ versus the eigenvalues of symmetric Gaussian random matrix of size $dn \times dn$ when $(n,d) = (1000,2)$ (right middle) and $(1000,3)$ (right).

3 Numerical work

We now present some numerical work to investigate the agreement between our theoretical results and simulations in “reasonable” dimensions.

3.1 Simulation of group-entry random matrix

We show three simulations for the block random matrix and one simulation of the class averaging algorithm based on VDM under the null hypothesis $H_0$ [5].

[Case 1: Random orthogonal group $SO(d)$ and $O(d)$ with Haar measure] Consider a $n \times n$ symmetric block matrix $R_{SO(d),n,Haar}$ with $d \times d$ entries so that its $(i,j)$-th entry, $i < j$, for all $i,j = 1,\ldots,n$, is uniformly sampled according to the Haar measure on $SO(d)$. We mention that the QR decomposition, at least as implemented in Matlab, leads to random matrices which are not distributed according to Haar measure, and needs to be corrected in order to numerically obtain the uniform samples on $SO(d)$ [8]. The histogram of $R_{SO(d),n,Haar}$’s spectrum is shown in Figure 1 when $n = 1000$ and $d = 2,3$. We also show the QQplot of the eigenvalues of $R_{SO(d),n,Haar}$ versus the eigenvalues of symmetric Gaussian random matrix of size $dn \times dn$, a good approximation to the Wigner semi-circle law. It is clear that the spectral distribution of $R_{SO(d),n,Haar}$ is a scaled semi-circle law, as predicted by our theory.

Next, consider a $n \times n$ symmetric block matrix $R_{O(d),n,Haar}$ with $d \times d$ entries so that its $(i,j)$-th entry, $i < j$, for all $i,j = 1,\ldots,n$, is uniformly sampled according to the Haar measure on $O(d)$ [8]. The histogram of $R_{O(d),n,Haar}$’s spectrum is shown in Figure 2 when $n = 1000$ and $d = 2,3$. We also show the QQplot of the eigenvalues of $R_{O(d),n,Haar}$ versus the eigenvalues of symmetric Gaussian random matrix of size $dn \times dn$. Again, it is clear that we obtain the semi-circle law.

[Case 2: Random orthogonal group $O(d)$ with non-Haar measure] Consider a $n \times n$ symmet-
Figure 3: Histogram of the eigenvalues of $R_{O(d),n,\text{nonHaar}}$ with entries sampled from $O(d)$ but not following the Haar measure when $(n,d) = (1000,2)$ (left) and $(1000,3)$ (left middle). It is clear to see an outlier at 14 in the left subfigure and two outliers in the left middle subfigure. To show that the bulk of the eigenvalues are close to the semi-circle, the QQplot of all the eigenvalues of $R_{O(d),n,\text{nonHaar}}$, which are less than 3, versus the eigenvalues of symmetric Gaussian random matrix of size $dn \times dn$ is plotted in the right middle (resp. right) subplot when $(n,d) = (1000,2)$ (resp. $(n,d) = (1000,3)$). Note the existence of the outliers which might be interpreted as “information”.

Figure 4: Histogram of the eigenvalues of $R_{O(d),n,\text{nonHaar}}$ with $d \times d$ entries so that its $(i,j)$-th entry is the orthogonal matrix in the QR decomposition of a random $d \times d$ matrix. It has been studied in [38] that this sampling scheme on $O(d)$ is non-uniform.

The histogram of $R_{O(d),n,\text{nonHaar}}$’s spectrum is shown in Figure 3 when $n = 1000$ and $d = 2,3$. We also show the QQplot of all the eigenvalues of $R_{O(d),n,\text{nonHaar}}$, which are less than 3, versus the eigenvalues of symmetric Gaussian random matrix of size $dn \times dn$. According to the QQplot, we may infer that the bulk of the empirical spectral distribution follows the semi-circle law. Note that the outliers might be interpreted as “information”, so we should be careful about the sampling scheme on the group matrix. In other words, there exist structures inside the block random matrix that might be misleading.

[Case 3: Special linear group $SL(d,\mathbb{R})$] Consider a $n \times n$ symmetric block matrix $R_{SL(d,\mathbb{R}),n}$ with $d \times d$ entries so that its $(i,j)$-th entry is sampled from $SL(d,\mathbb{R})$ by the following steps. For each $(i,j)$, $i < j$, get a random $d \times d$ matrix with i.i.d. Gaussian entries, and denote it as $g$. If $|\det(g)| = 0$, we resample another matrix until we get $g$ with $|\det(g)| > 0$. Then define $R_{SL(d,\mathbb{R}),n}(i,j)$ to be $|\det(g)|^{-1/d}g$. Then, if $\det(R_{SL(d,\mathbb{R}),n}(i,j)) = 1$, we get a component in $SL(d,\mathbb{R})$; otherwise, flip the sign of the first column to ensure we get a component in $SL(d,\mathbb{R})$.

The histogram of spectra of $R_{SL(d,\mathbb{R}),n}$ with $d = 2,3,4,5$ are shown in Figure 4 when $n = 1000$. We also show the QQplot of the eigenvalues of $R_{SL(d,\mathbb{R}),n}$ versus the eigenvalues of symmetric Gaussian random matrix of size $dn \times dn$ in Figure 5. Note that starting from $SL(4,\mathbb{R})$, the histogram are semi-circle-like.

On the other hand, the histogram of $SL(2,\mathbb{R})$ spreads broadly. The distribution and moments calculation of the sampling scheme on $SL(d,\mathbb{R})$ are detailed on Section C-2 in the ArXiv version of this paper, where we see the Cauchy-like behavior of the square of the largest singular value of $R_{SL(2,\mathbb{R}),n}$ (see Corollary C-1 in the ArXiv version of this paper). So our theory simply does not apply to this case and the fact that we do not get a semi-circle limit in this case is not surprising.

On the other hand, the case of $SL(3,\mathbb{R})$ falls under the umbrella of our theory (see Corollary 2.6). The QQplot of the distribution of $R_{SL(3,\mathbb{R}),n}$ against the semi-circle law shows a few outliers. This of course is not in contradiction with our theoretical results: we have shown convergence of the limiting spectral distribution of $R_{SL(3,\mathbb{R}),n}$ but this naturally does not imply that the extreme eigenvalues of this random matrix convergence to the endpoint of the limiting spectral distribution.

[Case 4: the class averaging algorithm based on VDM] We now show the connection graph Laplacian in the class averaging algorithm based on VDM. Recall that this was the motivation for our study. Under the null hypothesis $H_0$ (5), the discretized images are simulated as

\[ \mathcal{X}_p := \{Z_i^p\}_{i=1}^n \subset \mathbb{R}^p, \]

where $Z_i^p$, $i = 1,\ldots,n$ are prepared in the following way. Take a Gaussian random vector $Z \sim \mathcal{N}(0, I_{(2L+1)^2})$, where $L \in \mathbb{N}$, and obtain $n$ samples $Z_i$, $i = 1,\ldots,n$, i.i.d. from $Z$. Next, index the pixels of $Z_i$ by
Figure 4: Histogram of the eigenvalues of $R_{SL(d,\mathbb{R}),n}$ with $d = 2, 3, 4, 5$ (from left to right) and $n = 1000$.

Figure 5: QQplot of the eigenvalues of $R_{SL(d,\mathbb{R}),n}$ versus the eigenvalues of symmetric Gaussian random matrix of size $dn \times dn$ with $d = 2, 3, 4, 5$ (from left to right) and $n = 1000$.

$\{-L, -L+1, \ldots, L-1, L\} \times \{-L, -L+1, \ldots, L-1, L\}$. The image $Z^p_i$ is obtained by setting all the pixels with norm greater than $L$ to $0$, where $p$ is the number of pixels with indices of norm less than and equal to $L$. Please see Figure 3 for one of the realization. With $X$, for all $i, j = 1, \ldots, n$, evaluate

$$g_{ij} := \arg\min_{g \in SO(2)} \| Z^p_i - g \circ Z^p_j \|_{L^2},$$

where $g \circ Z^p_j$ means the numerical rotation of $Z^p_j$ by $g$. If there are more than one minimizer, we choose the first one as $g_{ij}$. Then, find the rotational invariant distance (RID) by

$$d_{RID,ij} := \min_{g \in SO(2)} \| Z^p_i - g \circ Z^p_j \|_{L^2}.$$

The $SO(2)$ is discretized to $N_r$ equally spaced degrees for the numerical minimization. See Figure 6 for the distribution of the optimal rotation $g_{ij}$ and the distribution of RID.

With $g_{ij}$ and $d_{RID,ij}$ and a chosen $\epsilon > 0$, we build up the $n \times n$ block matrix $S$ so that the $(i,j)$-th block is

$$S_{ij} = e^{-d_{RID,ij}^2/\epsilon} g_{ij},$$

where $i \neq j$, and the $n \times n$ diagonal block matrix $D$ so that the $i$-th diagonal block is

$$D_{ii} = \sum_{j=1, j \neq i}^n e^{-d_{RID,ij}^2/\epsilon} I_2,$$

where $i = 1, \ldots, n$. The histogram of the spectra of $C = D^{-1}S$ with $n = 700$, $L = 31$ and $N_r = 240$ is shown in Figure 6 which follows the semi-circle law. Note that in this case, $p = 3001$.

APPENDIX
quantities of the underlying manifold vary according to exist inside the high dimensional space. To be more precise, depending on the problem, sometimes some A A particular example of the VDM with a metric induced from the canonical metric of L \psi centered at 0. Without loss of generality, we assume that supp \psi \subset B_1(0), where B_1(0) is the Euclidean ball of radius 1 centered at 0. Let R be the canonical metric of SO(3). Recall that the X-ray transform T_\psi : SO(3) \rightarrow L^2(\mathbb{R}^2) is defined in [1] as

T_\psi(R)(x, y) = \int \psi(xR^1 + yR^2 + tR^3)dt = \int \psi(Ru)du^3, \tag{1}

where we denote R = [R^1 R^2 R^3] \in SO(3) in matrix form and u = (u^1, u^2, u^3)^T \in \mathbb{R}^3. Here T_\psi(R) : \mathbb{R}^2 \rightarrow \mathbb{R}^+ is the projection image and has a compact support inside B_1(0). We call R the rotational position of the image T_\psi(R) and R^3 the projection direction. We assume that T_\psi : SO(3) \rightarrow T_\psi(SO(3)) \subset L^2(\mathbb{R}^2) diffeomorphically maps SO(3) to T_\psi(SO(3)). In other words, the X-ray transform diffeomorphically embeds (SO(3), B) into L^2(\mathbb{R}^2) in a non-isometric way. Indeed, as an embedded manifold, T_\psi(SO(3)) is endowed with a metric induced from the canonical metric of L^2(\mathbb{R}^2). With this metric on T_\psi(SO(3)), we define an induced metric on SO(3), denoted as g, by

\begin{equation}
g_R(X_i, X_j) := \langle dT_\psi|_R(X_i), dT_\psi|_R(X_j) \rangle, \tag{2}
\end{equation}

where dT_\psi|_R is the total differential of T_\psi evaluated at R, X_i and X_j, i, j = 1, \ldots, 3, are left invariant vector fields on SO(3) and \langle \cdot, \cdot \rangle is the canonical inner product on L^2(\mathbb{R}^2), that is,

\begin{equation}
\langle f, h \rangle := \iint f(x, y)h(x, y)dxdy, \text{ where } f, h \in L^2(\mathbb{R}^2). \tag{3}
\end{equation}

Therefore, the set of all projection images generated from the X-ray transform, can also be modeled as a Riemannian manifold (SO(3), g) embedded in L^2(\mathbb{R}^2) via an isometric embedding \iota. Note that each point on the embedded manifold T_\psi(SO(3)) \subset L^2(\mathbb{R}^2) is a projection image of \psi. Given \theta \in [0, 2\pi], denote

\begin{equation}
\gamma(\theta) = \begin{bmatrix}
\cos(\theta) \\
\sin(\theta)
\end{bmatrix}, \quad \alpha(\theta) = \begin{bmatrix}
\gamma(\theta) & 0 \\
0 & 1
\end{bmatrix} \in SO(3). \tag{4}
\end{equation}

A A particular example of the VDM

We start from commenting that the assumption “the dataset is distributed on a given low dimensional manifold” needs to be clarified before proceeding with the analysis, in particular, how does the manifold exist inside the high dimensional space. To be more precise, depending on the problem, sometimes some quantities of the underlying manifold vary according to n and p, and some quantities are assumed to be fixed, for example, the dimension. These discrepancies can be clarified by the parameter space, the model space and the nuisance space discussed in the Introduction section. Here we clarify the assumption as well as demonstrate the framework by discussing a particular example – the class averaging algorithm in the cryoEM problem. We refer the reader to [33, 51] and the references therein for more practical and numerical details.

Take a compactly supported L^2 function \psi : \mathbb{R}^3 \rightarrow \mathbb{R}^+ to be the potential of a molecule which is centered at 0. Without loss of generality, we assume that supp\psi \subset B_1^3(0), where B_1^3(0) is the Euclidean ball of radius 1 centered at 0. Figure 6: The class averaging algorithm based on VDM applied to A_3001 with n = 700, L = 31 and N_r = 240. Left: the realization Z_1^{001}; middle left: the distribution of the optimal rotation g_{ij}, where the x-axis is the degree of the rotation; middle right: the distribution of the RID d_{RID,ij}; right: the histogram of the spectra of C.

\begin{align*}
\int_{-\infty}^{\infty} f(x)dx &= \int_{-\infty}^{\infty} f(x)\exp(-i2\pi x)dx, \\
\langle f, h \rangle &= \int_{-\infty}^{\infty} f(x)h(x)dx.
\end{align*}
Fix \( R = [R^1 \ R^2 \ R^3] \in SO(3) \). Denote \( R_1 := R_{R^3} \alpha(\theta) R_{R^3}^{-1} \), where \( R_{R^3} \in SO(3) \) rotates the \( z \) axis to \( R^3 \). Note that \( R_1 \) keeps the unit vector \( R^3 \) fixed and rotates the plane perpendicular to \( R^3 \) by \( \gamma(\theta) \). It is clear that the following holds:

\[
[R_1 \circ T_\psi(R)](x,y) := T_\psi(R_1 R)(x,y) = \int \psi(x) R_1^1 + y R_1^2 + t R_3^2) \, dt = \gamma(\theta) \circ T_\psi(R)(x,y),
\]

where \( \circ \) indicates the group action and the last equality means rotation of the projection image \( T_\psi(R) \) by the angle \( \theta \). This shows that X-ray transform preserves the \( SO(2) \) invariance of the \( SO(3) \) group. From the differential geometry viewpoint, we might view \( T_\psi(SO(3)) \) as the frame bundle of \( T_\psi(SO(3))/SO(2) \approx S^2 \) with the \( SO(2) \) group as the fiber. This viewpoint helps to understand the meaning of the asymptotically behavior of the VDM algorithm. We refer the reader to the standard textbook [10] or the Appendix of [53] for the notion of the frame bundle.

Up to now, the model we consider is continuous and noise free. There are two levels of discretizations we consider in order to model the practical problem. The first level is that we sample uniformly on \( SO(3) \) the angle \( \theta \) we consider in order to model the practical problem. The first level is that we sample the behavior of the VDM algorithm. We refer the reader to the standard textbook [10] or the Appendix of [53] for the notation of the frame bundle.

Note that we focus on these discretized points since \( \text{supp} \ T_\psi(R) \subset B_1^{R^2}(0) \). We denote the digitalization process as a projection operator \( d_p : L^2(\mathbb{R}^2) \rightarrow \mathbb{R}^p \) and denote the final discretized image as \( I^p := d_p(I^\infty) \).

The continuity of the projection operator and the fact that \( T_\psi(SO(3)) \) is a closed and smooth sub-manifold in \( L^2(\mathbb{R}^2) \) imply that

\[
T_{p,\psi}(SO(3)) := d_p(T_\psi(SO(3)))
\]

is compact and continuous in \( \mathbb{R}^p \). Note that in general \( T_{p,\psi}(SO(3)) \) might not have the original manifold structure, even in the topology sense. For example, in the extremal case like \( p = 1 \), we lose all the information and the original manifold structure is destroyed. We assume that there exists \( p_0 > 0 \) so that when \( p > p_0 \), \( T_{p,\psi}(SO(3)) \) is diffeomorphic to \( SO(3) \). Thus, we obtain a set of discretized projection images, which form our clean observation dataset:

\[
X_0 := \{I^p_{i_1}\}^{n}_{i_1=1} \subset T_{p,\psi}(SO(3)).
\]

Note that up to now, the randomness only comes from sampling the \( SO(3) \). However, in practice there exists noise in the observed projection images. We model the noise as a random vector \( Z_i \) of length \( p \) satisfying some conditions, like the sphere-like condition [25]: the noisy observation is then modeled as

\[
X := \{J^p_{i_1}\}^{n}_{i_1=1}, \text{ where } J^p_{i} = I^p_{i} + Z_i,
\]

\( Z_i \) is an i.i.d. \( p \)-dim random vector with zero mean and the covariance matrix is of finite norm. Since the signal to noise ratio (SNR) in the cryoEM problem is usually low, a possible approach is to increase the SNR by the class averaging algorithm [33, 54]. The idea beyond the class averaging algorithm is finding a group of projection images with the same projection direction, and then apply the law of large number to suppress the noise. Thus, the algorithm, together with the above X-ray transform model, can be understood in the following way. Since our interest now is the projection direction, the parameter space is \( S^2 \), which is the quotient space of \( SO(3) \) by the isotopic group at the identity, \( SO(2) \). The observation space is thus the discretized X-ray transform of the molecular, \( T_{p,\psi}(SO(3)) \), the nuisance space is \( SO(2) \), and the model space is \( T_\psi(SO(3))/SO(2) \) which is diffeomorphic to \( S^2 \). Note that in this particular problem, the observation space is equipped with a special structure – the frame bundle of \( T_\psi(SO(3))/SO(2) \), and this structure leads to the vector diffusion map (VDM). Note that now the data is a finite sample of the embedded \( SO(3) \) into \( \mathbb{R}^p \) via \( \iota_p := T_{p,\psi} \), which depends on the discretization \( p \).

In order to apply the VDM algorithm [51], we define the following quantities.
Definition 2 (Rotationally invariant distance (RID)). We define the distance between two images $J_i^p$ and $J_j^p$ as

$$d_{ij} := \min_{g \in SO(2)} \| J_i^p - g \circ J_j^p \|_2,$$

where $g \circ J_j^p$ means the numerical rotation of the discretized image $J_j^p$ by $g$.

Here we assume that the numerical rotation is error free in the sense that $d_p(g \circ T_p(R)) = g \circ d_p T_p(R)$ and $g \circ Z_i$ is again an i.i.d. $p$-dim random vector with zero mean and the covariance matrix is of finite norm. Although this assumption is reasonable in practice, further study is needed to precisely quantify the numerical error.

We also need to define the rotational relationship between $J_i^p$ and $J_j^p$.

Definition 3 (Rotational relationship). The rotational relationship between $J_i^p$ and $J_j^p$ is defined as

$$g_{ij} := \arg \min_{g \in SO(2)} \| J_i^p - g \circ J_j^p \|_2.$$

Then, take a graph $G = (V, E)$, where $V$ represents the finite observations $\mathcal{X}$ and $E$ represents the pairs of two projection images. Define the weight function $w : E \rightarrow \mathbb{R}^+$ by, for example, $w(i, j) = e^{-d_{ij}/\epsilon}$ for some $\epsilon > 0$ and $g : E \rightarrow O(2)$ is defined by $g(i, j) = g_{ij}$. With functions $w$ and $g$, we run VDM as the following.

Build up $S$ and $D$ from the graph $G$ by [3] and [4]. The matrix $C := D^{-1}S$ is referred to as Graph Connection Laplacian [2 51]. It is an operator acting on $v \in \mathbb{R}^{2n}$ by

$$(Cv)[i] = \frac{\sum_{j: (i,j) \in E} e^{-d_{ij}/\epsilon} g(i,j)v[j]}{\sum_{j: (i,j) \in E} e^{-d_{ij}/\epsilon}}, \quad (A-1)$$

where $v[i]$ is a 2-dim vector containing the $(2(i-1) + 1)$-th entry to the $2i$-th entry of $v$. We consider the following symmetric matrix which is similar to $C$

$$\widetilde{S} = D^{-1/2}SD^{-1/2}.$$

Since $\widetilde{S}$ is symmetric, it has a complete set of eigenvectors $v_{VDM,1}, v_{VDM,2}, \ldots, v_{VDM,2n}$ and eigenvalues $\mu_{VDM,1}, \mu_{VDM,2}, \ldots, \mu_{VDM,2n}$, where the eigenvalues are ordered by $|\mu_{VDM,1}| \geq |\mu_{VDM,2}| \geq \ldots \geq |\mu_{VDM,2n}|$.

Then, given the time $t > 0$, we define the vector diffusion mapping (VDM) $V_{t,n}$, which is a map from $\mathcal{X}$ to $\mathbb{R}^{(2n)^2}$ by

$$V_{t,n} : J_i^p \mapsto ((\mu_{VDM,t,i} \mu_{VDM,t,r})^{t^2} (v_{VDM,t,i}[i], v_{VDM,t,r}[i]))^{2n}_{i,r=1}.$$

With this map, the Hilbert-Schmidt norm of the $(i, j)$-th block of $\widetilde{S}$ satisfies

$$\| \widetilde{S}^{2t}(i, j) \|^2_{HS} = \langle V_{t,n}(J_i^p), V_{t,n}(J_j^p) \rangle,$$

that is, $\| \widetilde{S}^{2t}(i, j) \|^2_{HS}$ becomes an inner product for the finite dimensional Hilbert space. The reason we need to raise $\widetilde{S}$ to the power $2t$ - and not simply $t$ - is because all eigenvalues $\mu_{VDM,t,i}$ of $\widetilde{S}$ reside in the interval $[-1, 1]$, and we can not guarantee the positivity of $\mu_{VDM,t}$ when $n$ is finite. We can then define the vector diffusion distance (VDD) to quantify the affinity between nodes $i$ and $j$:

$$d_{VDM,t,n}(J_i^p, J_j^p) := \| V_{t,n}(J_i^p) - V_{t,n}(J_j^p) \|^2.$$

As is shown in [51], VDD is a surrogate of the geodesic distance estimator. Based on the estimated geodesic distance between projection images, we are able to group the images into clusters so that all images in the same cluster share similar projection directions, and hence the class averaging algorithm can be carried out [33 54].

Notice that when there exists noise, there are two sources of randomness in the dataset $\mathcal{X}$. One is from the random samples on the parameter space, and one is from the noise. To extract the information about the parameter space, we thus have to deal with both randomness simultaneously.
B  Technical results

B-1  Bound on the norm of a subblock of a matrix

We recall and prove the following simple fact.

**Fact B.1.** Suppose

\[ T = \begin{pmatrix} T_{11} & T_{12} \\ T_{21} & T_{22} \end{pmatrix}. \]

Then

\[ |||T_{ij}|||_2 \leq |||T|||_2. \]

**Proof.** Recall that if \( W \) is a \( n \times p \) matrix,

\[ |||W|||_2 = \sup_{u \in \mathbb{C}^n, ||u||=1} \sup_{v \in \mathbb{C}^p, ||v||=1} |u^*Wv|. \]

Let us show that \( |||T_{12}|||_2 \leq |||T|||_2 \). Suppose that \( T \) is \( n \times n \) and \( T_{12} \) is \( d \times m \), where \( d + m = n \). Note that

\[ u^*T_{12}v = (u^*0_{n-d})T\begin{pmatrix} 0_d \\ v \end{pmatrix} = \tilde{u}^*Tv. \]

Of course \( \|\begin{pmatrix} 0_d \\ v \end{pmatrix}\| = \|v\| \) and similarly for \( \begin{pmatrix} u \\ 0_{n-d} \end{pmatrix} \). By definition, for any \( u \) and \( v \) with unit norm,

\[ |\tilde{u}^*Tv| \leq \sup_{\alpha \in \mathbb{C}^n, ||\alpha||=1} \sup_{\beta \in \mathbb{C}^n, ||\beta||=1} |\alpha^*T\beta| = |||T|||_2. \]

So we have shown that

\[ |||T_{12}|||_2 \leq |||T|||_2. \]

The same reasoning applies to the other sub-blocks of \( T \).

B-2  On finite rank perturbations and Stieltjes transforms

The following lemma is used repeatedly in our proofs.

**Lemma B-1.** Suppose \( A \) and \( B \) are Hermitian \( n \times n \) matrices. Let \( z \in \mathbb{C}^+ \) and call \( \text{Im}[z] = v > 0 \). Then

\[ \left| \text{trace} \left( (A - z\text{Id})^{-1} \right) - \text{trace} \left( (B - z\text{Id})^{-1} \right) \right| \leq \frac{\text{rank}(A - B)}{v}. \]  \hspace{1cm} (B-2)

**Proof.** Since \( \Delta = A - B \) is Hermitian it is diagonalizable. Therefore, \( \Delta = \sum_{k=1}^{r} \tau_k q_k q_k^* \), where \( \tau_k \in \mathbb{R} \) and \( q_k \in \mathbb{C}^n \). \( r \) is, of course, the rank of \( \Delta \). Let us call, if \( 1 \leq j \leq r \) \( P\Delta_j = \sum_{k=1}^{j} \tau_k q_k q_k^* \) and \( P\Delta_0 = 0_{n \times n} \). We have

\[ (A - z\text{Id})^{-1} - (B - z\text{Id})^{-1} = \sum_{j=0}^{r-1} (B + P\Delta_{r-j} - z\text{Id})^{-1} - (B + P\Delta_{r-j-1} - z\text{Id})^{-1}. \]

Of course, \( B + P\Delta_{r-j-1} \) is a rank-1 perturbation of \( B + P\Delta_{r-j} \). Using Lemma 2.6 of [49], we therefore have

\[ \left| \text{trace} \left( (B + P\Delta_{r-j} - z\text{Id})^{-1} - (B + P\Delta_{r-j-1} - z\text{Id})^{-1} \right) \right| \leq \frac{1}{v}. \]

Therefore, since

\[ \left| \text{trace} \left( (A - z\text{Id})^{-1} - (B - z\text{Id})^{-1} \right) \right| \leq \sum_{j=0}^{r-1} \left| \text{trace} \left( (B + P\Delta_{r-j} - z\text{Id})^{-1} - (B + P\Delta_{r-j-1} - z\text{Id})^{-1} \right) \right|, \]

and the sum on the right hand side contains \( r \) (i.e rank \( (A - B) \) terms), the result stated in the lemma follows.
B-3 Invariance and moments

Lemma B-2. Let $B$ be a random $d \times d$ matrix. Let us call $r_j$, $1 \leq j \leq d$ the rows of $B$, and $c_j$’s the columns of $B$. Suppose that

1. the rows of $B$ are exchangeable and for any $j \neq k$ $(r_j, r_k) \overset{\mathcal{L}}{=} (r_j, -r_k)$.
2. the columns of $B$ are exchangeable and for any $j \neq k$ $(c_j, c_k) \overset{\mathcal{L}}{=} (c_j, -c_k)$.
3. for any $j$, $\text{cov}(r_j)$ exists.

Then we have

1. $\mathbf{E}(r_j) = 0$, for all $j$.
2. $\mathbf{E}(r_j r_k') = 0_{d \times d}$ when $j \neq k$.
3. $\mathbf{E}(r_j r_j') = \gamma \mathbf{I}_d$ for some $\gamma$.

Proof. Since we assume that $\text{cov}(r_j)$ exists for any $j$, it is clear that $\mathbf{E}(r_j r_k')$ exists by the Cauchy-Schwarz inequality. Since $(r_j, r_k) \overset{\mathcal{L}}{=} (r_j, -r_k)$ for $j \neq k$, we have

$$r_j r_k' \overset{\mathcal{L}}{=} -r_j r_k' \text{ and } r_j \overset{\mathcal{L}}{=} -r_j.$$

Therefore,

$$\mathbf{E}(r_j r_k') = -\mathbf{E}(r_j r_k') = 0_{d \times d} \text{ and } \mathbf{E}(r_j) = 0.$$

On the other hand, $\text{cov}(r_j)(k, l) = \mathbf{E}(c_k(j)c_l(j))$. If $k \neq l$, our assumption that for any $j \neq k$ $(c_j, c_k) \overset{\mathcal{L}}{=} (c_j, -c_k)$ guarantees by the same argument as above that $\mathbf{E}(c_k(j)c_l(j)) = 0$. On the other hand, since the columns are exchangeable, it is clear that $\text{cov}(r_j)(k, k) = \mathbf{E}(c_k(j)c_k(j)) = \mathbf{E}(c_k(j)^2) = \text{cov}(r_j)(l, l)$. So

$$\mathbf{E}(r_j r_j') = \gamma j \mathbf{I}_d.$$

Our assumption that the rows are exchangeable guarantees that for all $j \neq k$, $\gamma j = \gamma k = \gamma$. \qed

C Distributional results for various models

C-1 Properties of the null distribution for VDM

We now consider the distribution of

$$g_{ij} = \arg\min_{g \in \mathbf{SO}_2} \|Z_i - g \circ Z_j\|_2^2. \quad \text{(C-3)}$$

Lemma C-1 (Null Case VDM). Suppose that $Z_i$ and $Z_j$ are independent. Suppose that each random variable has a distribution that is invariant under the action of $\mathbf{SO}_2$.

Then $g_{ij}$ and $Z_i$ are independent and so are $g_{ij}$ and $Z_j$. Furthermore, $g_{ij}$ is uniformly distributed on $\mathbf{SO}_2$.

Proof. Note that conditional on $Z_i$, if $Z_j \rightarrow O^{-1} \circ Z_j$, where $O \in \mathbf{SO}_2$, then $g_{ij} \rightarrow Og_{ij}$ by (C-3). Hence, using the assumption that $Z_j \overset{\mathcal{L}}{=} O \circ Z_j$ (i.e the law of $Z_j$ is invariant under the action of $\mathbf{SO}_2$) and $Z_j|Z_i \overset{\mathcal{L}}{=} O \circ Z_j|Z_i$ (this latter equality coming from independence of $Z_i$ and $Z_j$), we see that

$$g_{ij}|Z_i \overset{\mathcal{L}}{=} Og_{ij}|Z_i.$$

Since the only distribution on $\mathbf{SO}_2$ that is invariant by left-multiplication by an $\mathbf{SO}_2$ is the uniform distribution on $\mathbf{SO}_2$, we conclude that $g_{ij}|Z_i$ has the uniform distribution on $\mathbf{SO}_2$.  

28
Because the uniform distribution on $SO_2$ does not depend on $Z_i$, $g_{ij}$ and $Z_i$ are independent. Indeed, let $\Gamma$ be a function of $Z_i$ and $\omega$ be a function of $g_{ij}$. Note that since the distribution of $g_{ij}|Z_i$ does not depend on $Z_i$, we have

$$E(\omega(g_{ij})|Z_i) = \Omega = E(\omega(g_{ij})).$$

In other words, $\Omega$ is a constant (in particular, it does not depend on $Z_i$). Therefore, we have

$$E[\omega(g_{ij})\Gamma(Z_i)] = E[E(\omega(g_{ij})\Gamma(Z_i)|Z_i)] = E[\Gamma(Z_i)E(\omega(g_{ij})|Z_i)]$$

$$= E[\Gamma(Z_i)] \Omega = E[\Gamma(Z_i)] E[\omega(g_{ij})].$$

The same argument shows that $g_{ij}$ is also independent of $Z_j$. So we have established that $g_{ij}$ and $Z_i$ are independent. The same argument shows that $g_{ij}$ and $Z_j$ are independent. However the three random variables $g_{ij}$, $Z_i$ and $Z_j$ are not jointly independent. \hfill \Box

The previous lemma has the following useful consequence.

**Lemma C-2.** Suppose that $Z_i$, $Z_j$ and $Z_k$ are independent, each random variable having a distribution that is invariant under the action of $SO_2$.

Then $g_{ij}$ and $g_{ik}$ are independent, and so are $g_{ij}$ and $g_{jk}$.

Furthermore, the random variables $\{g_{ij}\}_{j=1}^n$ are jointly independent.

**Proof.** Let $f_1$ and $f_2$ be two functions. We have

$$E(f_1(g_{ij})f_2(g_{ik})) = E(E(f_1(g_{ij})f_2(g_{ik})|Z_i)).$$

Now, it is clear that $g_{ij}|Z_i$ is a function of $Z_j$ only. Similarly, $g_{ik}|Z_i$ is a function of $Z_k$ only. So $g_{ij}|Z_i$ is independent of $g_{ik}|Z_i$. Therefore,

$$E(f_1(g_{ij})f_2(g_{ik})|Z_i) = E(f_1(g_{ij})|Z_i) E(f_2(g_{ik})|Z_i).$$

Now recall that we have shown that $g_{ij}|Z_i \sim U$, where $U$ is a uniformly distributed random variable on $SO_2$; the same result applies to $g_{ik}$. Therefore,

$$E(f_1(g_{ij})|Z_i) E(f_2(g_{ik})|Z_i) = E(f_1(U)) E(f_2(U)).$$

Of course, our argument above shows that $E(f_1(g_{ij})) = E(f_1(U))$. We conclude that

$$E(f_1(g_{ij})f_2(g_{ik})) = E(f_1(U)) E(f_2(U)),$$

$$= E(f_1(g_{ij})) E(f_2(g_{ik})).$$

This shows that $g_{ij}$ and $g_{ik}$ are independent. The proof of joint independence of $\{g_{ij}\}_{j=1}^n$ follows exactly in the same manner: just start the proof with $f_1, \ldots, f_n$ and apply the same reasoning.

Our statement concerning $g_{ij}$ and $g_{jk}$ is also proven in a similar manner, by writing

$$E(f_1(g_{ij})f_2(g_{jk})) = E(E(f_1(g_{ij})f_2(g_{jk})|Z_j)),$$

and using the fact that $g_{ij}$ and $g_{jk}$ are independent conditionally on $Z_j$. The rest of the argument is similar to the one we gave above. \hfill \Box

So we have established some pairwise independence results, but we do not have joint independence for the three random variables $(g_{ij}, g_{ik}, g_{jk})$ or the random variables $(g_{ij})_{i<j}$.

**C-2 More details on the matrices drawn from $SL_d$**

We give more details about the stochastic properties of the matrices $B$ drawn according to the scheme described in Corollary 2.6.

29
C-2.1 Computing the joint density of the singular values

We consider the problem of understanding the singular values of the matrix

$$B = \frac{G}{|\det(G)|^{1/d}},$$

where $G$ is $d \times d$ with i.i.d Gaussian entries.

We write the an svd of $G$ as $G = UDV'$. By rotational invariance arguments, it is clear that $(U, V)$ and $D$ are independent. Furthermore, $U$ and $V$ are Haar-distributed on $O(d)$. Note that defining $U$ and $V$ as svd-representatives may induce some mild dependence between them, because of sign issues. It is possible to deal with this dependence issue but we do not discuss it further as our interest here is in singular values.

Let us call $s_i$ the singular values of $B$. Recall that $s_i \geq 0$. For simplicity, we seek to understand not the joint distribution of $s_i$’s but that of $s_i^2$. Note that if $d_i$’s are the singular values of $G$, we have the relationship

$$s_i = \frac{d_i}{(\prod_{1 \leq i \leq d} d_i)^{1/d}},$$

since $|\det(G)| = \sqrt{|\det(G')G|} = \sqrt{\prod_{1 \leq i \leq d} d_i^2}$.

In particular, we have, if we denote by $l_i$’s are the eigenvalues of $G'G$,

$$s_i^2 = \frac{l_i}{(\prod_{1 \leq i \leq d} l_i)^{1/d}}.$$ 

We have the following fact:

**Fact C.1.** The joint density of $l_1 > l_2 > \ldots > l_d$ is

$$f(l_1, \ldots, l_d) = C(d) \exp\left(-\frac{1}{2} \sum_{i=1}^{d} l_i \prod_{i=1}^{d} l_i^{-1/2} \prod_{i<j} (l_i - l_j) \right).$$ (C-4)

*Proof.* We note that $G'G$ is Wishart-distributed, specifically $\mathcal{W}(d, \text{Id}_d)$. The fact we mention is therefore just the content of Corollary 3.2.19 in [39]. The value of $C(d)$ is known explicitly:

$$C(d) = \frac{\pi^{d^2/2}}{2^{d^2/2} \Gamma(d/2)^2},$$

where $\Gamma_d(x) = \pi^{d(d-1)/4} \prod_{i=1}^{d} \Gamma[x - (i - 1)/2]$, provided $\text{Re}[x] > (m - 1)/2$ and $\Gamma$ is the ordinary Gamma function. For details, see [39], pp.61-62. $\square$

We now assume that $s_1 > s_2 > \ldots > s_d$. We call

$$y_i = s_i^2, 1 \leq i \leq d,$$

$$t = \left[ \prod_{1 \leq i \leq d} l_i \right]^{1/d}.$$ 

We would like to find $g(y_1, \ldots, y_{d-1})$, the density of the $d - 1$ largest eigenvalues of $B'B$. Note that $\det(B'B) = 1$, so $\prod_{1 \leq i \leq d} y_i = 1$.

Note also that if we keep the ordering, we must have $\left( \prod_{1 \leq i \leq d-1} y_i \right) y_{d-1} > 1$ to guarantee that there exists $y_d < y_{d-1}$ such that $\prod_{1 \leq i \leq d} y_i = 1$. This defines the subset of $\mathbb{R}^{d-1}$ where $(y_i)_{i=1}^{d-1}$ lives.

**Lemma C-3.** Let us call $\bar{y}$ the vector $(y_1, \ldots, y_{d-1})$, where $y_1 > y_2 > \ldots > y_{d-1} > 0$ and $\prod_{1 \leq i \leq d-1} y_i > 1/y_{d-1}$. We call $\mathcal{R}$ this subset of $\mathbb{R}^{d-1}$. 


Let \( \alpha = \frac{1}{\prod_{1 \leq i \leq d-1} y_i} \) and
\[
\gamma(\tilde{y}) = \frac{1}{2} \left( \sum_{1 \leq i \leq d-1} y_i + \alpha \right),
\]
\[
R(\tilde{y}) = \alpha \prod_{1 \leq i < j \leq d-1} (y_i - y_j) \prod_{1 \leq i \leq d-1} (y_i - \alpha).
\]

Then, the density of \( \tilde{y} \) over \( \mathcal{R} \) is
\[
g(y_1, \ldots, y_{d-1}) = \tilde{C}(d) \frac{R(\tilde{y})}{\gamma(\tilde{y})^{d/2}}.
\]

**Proof of Lemma C-3:** To find the density, we will use the following change of variables from \((l_1, \ldots, l_d)\) to \((y_1, \ldots, y_{d-1}, t)\):
\[
l_i = ty_i, \ 1 \leq i \leq d-1,
\]
\[
l_d = \frac{t}{\prod_{1 \leq i \leq d-1} y_i}.
\]

We call \( \alpha = \frac{1}{\prod_{1 \leq i \leq d-1} y_i} \).

Let us call \( \tilde{y} \) the \( d - 1 \times 1 \) vector with \( i \)-th entry \( y_i \). \( 1/\tilde{y} \) is the \( (d - 1) \times 1 \) vector with \( i \)-th entry \( 1/y_i \).

The Jacobian matrix for the change of variables we just discussed is
\[
M = \begin{pmatrix}
  t \mathbb{I}_{d-1} & \tilde{y} \\
  -\alpha/\tilde{y} & \alpha
\end{pmatrix}
\]

By multilinearity of the determinant, we therefore have
\[
\det(M) = \alpha \det \left( t \mathbb{I}_{d-1} - \frac{\tilde{y}}{\alpha} \right) = \alpha t^{d-1} \det \left( t \mathbb{I}_{d-1} - \frac{\tilde{y}}{1} \right).
\]

Now, let’s call \( y \) the \( d \times 1 \) vector such that
\[
y = \left( \begin{array}{c}
  \tilde{y} \\
  0
\end{array} \right).
\]

And let \( 1/y \) be the vector such that
\[
1/y = \left( \begin{array}{c}
  1/\tilde{y} \\
  0
\end{array} \right).
\]

We have, if \( e_d \) denotes the \( d \)-th canonical basis vector,
\[
\begin{pmatrix}
  \mathbb{I}_{d-1} & \tilde{y} \\
  -1/\tilde{y} & 1
\end{pmatrix} = \mathbb{I}_d + ye_d' - e_d 1/y'.
\]

From determinant theory ([31], Theorem I.3.2, p.9), we know that
\[
\det(\mathbb{I}_d + \sum_{1 \leq i \leq m} \phi_i \otimes f_i) = \det(\delta_{i,j} + \langle \phi_i, f_j \rangle)_{1 \leq i,j \leq m}.
\]

In the circumstances of interest to us, we have
\[
\langle \phi_i, f_j \rangle_{1 \leq i,j \leq 2} = \begin{pmatrix}
  0 & -(d-1) \\
  1 & 0
\end{pmatrix}.
\]

So we conclude that
\[
\det \begin{pmatrix}
  \mathbb{I}_{d-1} & \tilde{y} \\
  -1/\tilde{y} & 1
\end{pmatrix} = 1 + (d-1) = d.
\]
Hence, the Jacobian of our change of variable is
\[ J = \alpha t^{d-1} d \]

We conclude that the density of \((y_1, \ldots, y_{d-1}, t)\) is
\[ h(y_1, \ldots, y_{d-1}, t) = \frac{1}{\prod_{1 \leq i \leq d-1} y_i} t^{d-1} f(t y_1, \ldots, t y_{d-1}, t t^2 \prod_{1 \leq i \leq d-1} y_i). \]

Now,
\[ f(t y_1, \ldots, t y_{d-1}, t) = C(d) \exp(\frac{t}{2} \sum_{1 \leq i \leq d-1} y_i + \alpha) t^{-d/2} \prod_{1 \leq j < i \leq d-1} (y_i - y_j) \prod_{1 \leq i \leq d-1} (y_i - \alpha). \]

Therefore,
\[ h(y_1, \ldots, y_{d-1}, t) = C(d) \alpha t^{d-1-d/2-d(d-1)/2} \exp(\frac{t}{2} \sum_{1 \leq i \leq d-1} y_i + \alpha) \prod_{1 \leq j < i \leq d-1} (y_i - y_j) \prod_{1 \leq i \leq d-1} (y_i - \alpha). \]

where
\[ \gamma(\tilde{y}) = \frac{1}{2} \left( \sum_{1 \leq i \leq d-1} y_i + \alpha \right), \]
\[ R(\tilde{y}) = \alpha \prod_{1 \leq j < i \leq d-1} (y_i - y_j) \prod_{1 \leq i \leq d-1} (y_i - \alpha). \]

Now the joint density of \((y_1, \ldots, y_{d-1})\) is simply:
\[ g(y_1, \ldots, y_{d-1}) = \int_0^\infty h(y_1, \ldots, y_{d-1}, t) dt. \]

Note that \(\sum_{1 \leq i \leq d-1} y_i + \alpha > 0\) in the domain we consider, so there are no integrability problems. Also, if \(K\) is an integer,
\[ \int_0^\infty t^{K-1} \exp(-\beta t) dt = \Gamma(K) \beta^{-K} = (K-1)! \beta^{-K}. \]

Therefore, we finally have, for \(y_1 > y_2 > \ldots > y_{d-1}\),
\[ g(y_1, \ldots, y_{d-1}) = C(d) \frac{R(\tilde{y})}{(\gamma(\tilde{y}))^{d^2/2}}. \]

The Lemma is shown. \(\square\)

Let us apply the Lemma in the case \(d = 2\).

**Corollary C-1** (Case \(d = 2\)). Then, \(\gamma(\tilde{y}) = \frac{1}{2}(y_1 + 1/y_1)\) and \(R(\tilde{y}) = (1 - 1/y_1^2)\).

So, for \(y_1 > 1\),
\[ g(y_1) = C \frac{1 - y_1^{-2}}{(y_1 + 1/y_1)^2} \sim \infty C y_1^{-2}. \]

Therefore, \(y_1 = s_1^2\) has a \(1 - \epsilon\) moment for any \(\epsilon > 0\), but not 1 moment. In other words,
\[ \mathbb{E}(s_1^{2-\epsilon}) < \infty \text{ if } \epsilon > 0, \]

and \(\mathbb{E}(s_1^2) = \infty\).

This corollary shows that the square of the largest singular value of \(G/|\det(G)|^{1/d}\) has Cauchy-like behavior in the tail.
C-2.2 On the entries of $B'B$

We recall the famous Bartlett decomposition of a Wishart matrix (see [39], p.99).

**Theorem C.1** (Bartlett Decomposition). Let $A$ be $W_p(n, \text{Id}_p)$, with $n \geq p$ and write $A = T'T$, where $T$ is an upper-triangular $p \times p$ matrix with positive diagonal elements. Then the elements of $T$ are all independent, $T_{i,i}^2$ is $\chi^2_{n-i+1}$, for $1 \leq i \leq p$, and $T_{i,j}$ is $N(0, 1)$ for $1 \leq i < j \leq p$.

From now on, we call $T$ the upper-triangular matrix appearing in the Bartlett decomposition of $G'G$.

We have the following lemma.

**Lemma C-4.** We have

$$B'B = \tilde{T}'\tilde{T},$$

where

$$\tilde{T} = \frac{T}{\det(T)^{1/d}} = \frac{T}{\prod_{i=1}^{d} T_{i,i}^{1/d}}.$$ 

$T_{i,i}^2$ are independent and have distribution $\chi^2_{d-i+1}$, for $1 \leq i \leq d$.

Calling $T_i$ the $i$-th column of $T$, we therefore have

$$(B'B)_{i,j} = \tilde{T}'_i \tilde{T}_j.$$ 

If $i \leq j$, we have in particular

$$(B'B)_{i,j} = \sum_{k<i} T_{k,i} T_{k,j} \prod_{l=1}^{d} \frac{T_{l,l}^{2/d}}{T_{l,l}^{2/d}}.$$ 

More specifically,

1. when $i < j$,

$$(B'B)_{i,j} = \sum_{k<i} \frac{T_{k,i} T_{k,j}}{\prod_{l=1}^{d} T_{l,l}^{2/d}} + \frac{T_{i,i}^{1-2/d} T_{i,j}}{\prod_{l \neq i} T_{l,l}^{2/d}}.$$ 

(Because of independence properties of the $T_{i,j}$’s, computations of moments for $(B'B)_{i,j}$ is relatively simple.)

2. when $i = j$,

$$(B'B)_{i,i} = \sum_{k<i} \frac{T_{k,i}^2}{\prod_{l=1}^{d} T_{l,l}^{2/d}} + \frac{T_{i,i}^{2-2/d}}{\prod_{l \neq i} T_{l,l}^{2/d}}.$$ 

In particular, when $d = 2$, $(B'B)_{1,1} = \frac{T_{1,1}}{T_{2,2}} = \frac{\chi_2}{\chi_1}$, where the two $\chi$ random variables are independent. Since a Cauchy random variable is the ratio of two independent $\chi_1$ random variables we conclude that $(B'B)_{1,1}$ is stochastically larger than a Cauchy random variable.

D On the “Sparsity” of the kernel random matrix built from the signal

In continuing the discussion in Appendix A, in general, we shall call the element in the observation space, $X_i \in X$ the signal random vector, the added noise $Z_i$ the noise random vector, and the collected data

$$Y_i = X_i + Z_i$$

the noisy signal random vector. In this section, we discuss specifically the case when there is a manifold structure inside the signal random vector and the noise random vector does not exist. To be more precise, we assume $Z_i = 0$ in (D-5),

$$Y_i = X_i$$
Figure 7: The relationship between $\Omega$, $X$, $\mathbb{R}^p$, $\iota$ and $M$. The probability density function defined on the manifold $M$ is understood as a function defined on $M$.

and assume that $X_i$ is sampled from a manifold embedded in the Euclidean space. We introduce more notations.

Denote $M$ to be a $d$-dimensional compact smooth Riemannian manifold embedded in $\mathbb{R}^p$ via $\iota$. Denote the tangent bundle as $TM$. The tangent plane at $y \in M$ is denoted as $T_yM$. Introduce the metric $g$ on $M$ induced from the canonical metric of the ambient space $\mathbb{R}^p$. When the boundary $\partial M$ is not empty, it is assumed to be smooth, and we denote $M_\delta := \{ x \in M : d(x, \partial M) \leq \sqrt{\delta} \}$, where $d(\cdot, \cdot)$ is the geodesic distance. Denote by $\nabla$ the covariant derivative of the vector field, $\Delta_g$ the Laplace-Beltrami operator, $\nabla^2$ the connection Laplacian of the tangent bundle associated with $\nabla$, and by $\text{Ric}$ the Ricci curvature of $(M, g)$.

Let the signal random vector $X : \Omega \rightarrow \mathbb{R}^p$ be a measurable function with respect to the probability space $(\Omega, \mathcal{F}, P)$ so that its range is $\iota(M)$. The probability density function (p.d.f.) of $X$ is in general not defined when $d < p$. Here we employ the following definition which is based on the induced measure and change of variable [14]. Denote $\mathcal{B}$ to be the Borel sigma algebra on $\iota(M)$, and $\tilde{P}_X$ the probability measure of $X$, defined on $\mathcal{B}$, induced from $P$. Assume that $\tilde{P}_X$ is absolutely continuous with respect to the volume density on $\iota(M)$ associated with $g$, that is, $d\tilde{P}_X(x) = f(\iota^{-1}(x))\iota_\#dV(x)$, where $x \in \iota(M)$. Assume $f \in C(M)$. Thus, we have

$$
\mathbb{E}\zeta(X) = \int_{\Omega} \zeta(X(\omega))dP(\omega) = \int_{\iota(M)} \zeta(x)d\tilde{P}_X(x) = \int_{\iota(M)} \zeta(\iota(x))f(\iota^{-1}(x))\iota_\#dV(x) = \int_M \zeta(\iota(y))f(y)dV(y),
$$

(D-6)

where $\zeta : \iota(M) \rightarrow \mathbb{R}$ is integrable and the last one comes from the change of variable $x = \iota(y)$. In this sense we interpret $f$ as the p.d.f. of $X$ on $M$. See Figure 7 for an illustration.

To alleviate the notation, in the following we abuse the notation and will not distinguish between $\iota(M)$ and $M$. We use the same notation $\mathcal{X} = \{X_i\}_{i=1}^p$ to denote the sample we get from $X$. Then, build up the graph $G_M := (V, E)$ by taking $V = \mathcal{X}$, and $E = \{(X_i, X_j) : X_i \in \mathcal{X}\}$.

Pick up a kernel function $K \in C^2([0, 1])$ so that $K\big|_{(1, \infty)} = 0$. Denote $\mu_l^{(k)} := \int_{\mathbb{R}^d} ||x||^l K^{(k)}(||x||)dx$, where $k = 0, 1, 2$, $l \in \mathbb{N} \cup \{0\}$, and $K^{(k)}$ means the $k$-th order derivative of $K$. We assume $\mu_0^{(0)} = 1$. Choose a bandwidth $h > 0$. We assume that $\sqrt{h}$ is small enough, in particular, smaller than the reach $\text{reach}$ and the injectivity radius $\text{inj}$ of the manifold $M$. Denote $K_h(X_i, X_j) := K(\|X_i - X_j\|_{\mathbb{R}^p}/h)$

D-1 Diffusion Map

The first algorithm we discuss is the diffusion map (DM), which can be viewed as a special case of VDM in that the principal bundle we are working with is trivial [53]. We refer the reader to [17] for more

\footnotesize
\[\text{Sometimes we might build up a more sophisticated graph with sparser edges like } E = \{(X_i, X_j) : \|X_i - X_j\|_{\mathbb{R}^p} \leq \sqrt{h}\} \text{ for a given } h > 0.\]
details about the DM algorithm. We make more assumptions about the signal random vector $X$:

**Assumption D.1.** *(D1)* Suppose the diameter of $M$ is bounded by $D > 0$ and the Ricci curvature has a lower bound $k \in \mathbb{R}$;

*(D2)* The p.d.f. $f \in C^3(M)$ is uniformly bounded from below and above, that is, $0 < p_m \leq f(x) \leq p_M < \infty$. However, to simplify the exploration, we assume here that $f$ is uniform. Note that when $f$ is non-uniform, further normalization is needed. [17, 53]

Given the graph $G_M$ and the kernel $K$, build up a $n \times n$ weight matrix $W$ by

$$W(i,j) = K_h(X_i, X_j).$$  \(\text{\text{(D-7)}}\)

Then, build up a $n \times n$ diagonal matrix $D$ by

$$D(i,i) := \sum_{j=1}^{n} W(i,j),$$  \(\text{\text{(D-8)}}\)

and a $n \times n$ matrix

$$L := D^{-1}W_i,$$  \(\text{\text{(D-9)}}\)

which is the graph Laplacian built from the graph $G_M$ associated with the manifold. Note that (D-7) and (D-8) are the special case of (3) and (4) when $B = \mathcal{X}$, $\mathcal{X}$ is sampled from $M$ and $g : E \rightarrow SO(1) = \{1\}$. Then we evaluate the eigenstructure of $L$, and denote $u_{DM,i} \in \mathbb{R}^n$ as the $i$-th eigenvector with the associated eigenvalue $\nu_{DM,i}$, that is,$$Lu_{DM,i} = \nu_{DM,i}u_{DM,i}.$$  \(\text{\text{(D-9)}}\)

With DM, we are allowed to introduce a new metric between sampled points, which is referred to as diffusion distance (DD):

$$d_{DM,t,n}(X_i, X_j) := \| \Phi_{t,n}(X_i) - \Phi_{t,n}(X_j) \|_{\mathbb{R}^{n-1}}.$$  \(\text{\text{(D-10)}}\)

In practice, when there is noise, that is, when $Z_i \neq 0$, we replace $X_i$ above by $Y_i$. The theoretical properties of DM and DD will be clear when $n \rightarrow \infty$.

We claim that $L$ satisfies some “sparsity” property which we define now.

**Definition 4.** The graph Laplacian built from a given weighted graph, denoted as a $n \times n$ matrix $L$, satisfies the graph sparsity property if its eigenvalues $\nu_{L, \ell}$, $\ell = 1, \ldots, n$, satisfy

$$\nu_{L, \ell} = O(e^{-\ell^{\gamma}})$$

for some $\gamma > 0$. In addition, under Assumption [D.1], the built $n \times n$ matrix $L$ in (D-9) satisfies the manifold sparsity property if its eigenvalues satisfy

$$\nu_{DM, \ell} = O(e^{-\ell^{2/D}}),$$

where the constants depends on the dimension $d$, the lower bound of the Ricci curvature $k$ and the diameter $D$ of the manifold $M$.

We claim that asymptotically as $n \rightarrow \infty$, $L$ satisfies the manifold sparsity property.

To show this claim, we need the following two Theorems. These theorems summarize the asymptotic behavior of $L$ in the pointwise sense and in the spectral sense.
Theorem D.2 (DM Pointwise Convergence). Suppose Assumption D.1 hold and let \( q \) be a function in \( C^4 \). For all \( X_i \) away from the boundary, that is, \( X_i \notin M_{\sqrt{n}} \) with high probability (w.h.p.)

\[
(Lq)[i] = q(X_i) + \frac{h^2}{2d} \Delta_q q(X_i) + O\left(h^2 + n^{-\frac{1}{2}}h^{-\frac{d-2}{4}}\right)
\]

where \( q \in \mathbb{R}^n \) and \( q(i) = q(X_i) \). For all \( X_i \) near the boundary, that is, \( X_i \in M_{\sqrt{n}} \), we have w.h.p.

\[
(Lq)[i] = q(X_i) + O(\sqrt{h})\nabla_{\partial_d} q(x_0) + O\left(h + n^{-\frac{1}{2}}h^{-\frac{d-2}{4}}\right),
\]

where \( x_0 = \arg\min_{y \in \partial M} d(X_i, y) \) and \( \nabla_{\partial_d} \) is the derivative in the normal direction.

To state the following spectral convergence result \([5, 53]\), we need a new notation. Define \( T_{DM,h} : C(M) \to C(M) \):

\[
T_{DM,h}q(y) := \frac{\sum_{j=1}^{n} K_h(y, X_j)q(X_j)}{\sum_{j=1}^{n} K_h(y, X_j)},
\]

where \( q \) is a continuous function defined on \( M \). Note that when \( M = \mathbb{R} \) and is isometrically embedded into \( \mathbb{R} \), \( T_{DM,h}q \) is the well-known Nadaraya-Watson kernel regression. In our current setup, we can view it as a generalization of the Nadaraya-Watson kernel regression to the manifold setup. We refer the reader to \([14, \text{Section 6}] \) for more details about its relationship with the regression setup. A simple argument shows that the eigenvalues and eigenfunctions of \( T_{DM,h} \) are equivalent to the eigenvalues and eigenvectors of \( L \).

Theorem D.3 (DM Spectral Convergence). Suppose Assumption D.1 and fix \( t > 0 \). Denote \( \nu_{DM,t,h,i} \) to be the \( i \)-th eigenvalue of \( T_{DM,h}^{t/h} \) with the associated eigenvector \( q_{DM,t,h,i} \). Also denote \( \nu_{t,i} \) to be the \( i \)-th eigenvalue of the heat kernel of the Laplace-Beltrami operator \( e^{t\Delta_g} \) with the associated eigen function \( q_i \).

We assume \( \nu_{DM,t,h,i} \) and \( \nu_{t,i} \) decreases as \( i \) increases, respecting the multiplicity. Fix \( i \in \mathbb{N} \). Then there exists a sequence \( h_n \to 0 \) such that

\[
\lim_{n \to \infty} \nu_{DM,t,h_n,i} = \nu_{t,i}
\]

and

\[
\lim_{n \to \infty} \|q_{DM,t,h_n,i} - q_i\|_{L^2(M)} = 0
\]

in probability.

Denote \( -\gamma_i \) to be the \( i \)-th eigenvalue of \( \Delta_g \), where \( \gamma_i \geq 0 \), which is related to \( \nu_{t,i} \) by \( \nu_{t,i} = e^{-t\gamma_i} \). Recall the well known Weyl’s theorem \([8]\), that is

\[
N(\gamma) \sim \frac{1}{(4\pi)^{d/2}\Gamma(d/2 + 1)}\gamma^{d/2},
\]

where \( N(\gamma) \) is the number of eigenvalues of \( \Delta_g \) less than \( \gamma > 0 \), and the consequence that \([7]\)

\[
\gamma_j \geq A(d, k, D)j^{2/d}, \tag{D-11}
\]

where \( j \in \mathbb{N} \) and \( A(d, k, D) \) is the universal constant depending only on \( d \), the lower bound of the Ricci curvature \( k \) and the diameter \( D \). In other words, we have \( \nu_{t,j} = O(e^{-tj^{2/d}}) \). With Theorem D.2 Theorem D.3 and (D-11), we conclude that asymptotically \( L \) satisfies the manifold sparsity property.

Recall that for several applications of DM, we need only the first few eigenvalues and eigenvectors of the graph Laplacian. For example, in the spectral clustering problem \([58]\), we only need to find the first \( k \) trivial eigenfunctions; in the cryo-EM problem, if we want to reconstruct the rotational position of each projection image, we need the first 9 non-trivial eigenfunctions \([28]\); in the random tomography problem \([52]\), only the first 2 non-trivial eigenfunctions are needed; in the orientability detection problem \([51]\), we need the first eigenfunction. More algorithms depending on the eigenstructure of the graph Laplacian can be found, to mention but a few, in \([18, 37, 44, 47, 48, 56]\).
In yet another cases, we may need to estimate the geodesic distance among points, and DD is proposed for this goal. We need to discuss some spectral geometry in order to quantify DD and understand its relationship with the geodesic distance [7]. Consider the following class of closed Riemannian manifolds \( \mathcal{M}_{d,k,D} \) with prescribed constraints:

\[
\mathcal{M}_{d,k,D} := \{(M, g)| \dim(M) = d, \text{Ric}(g) \geq (d-1)kg, \text{diam}(M) \leq D\},
\]

where Ric is the Ricci curvature and diam is the diameter. Then embed \( M \in \mathcal{M}_{d,k,D} \) into the \( \ell^2 \) space of real-valued, square integrable series by considering the heat kernel of the Laplace-Beltrami operator \( \Delta_g \) of \((M, g)\). Indeed, the classical elliptic theory allows us to decompose \( L^2(M) \) as \( L^2(M) = \bigoplus_{k=1}^{\infty} E_k \), where \( E_k \) is the eigenspace of \( \Delta_g \) corresponding to increasing eigenvalues, denoted as \( \gamma_k \) [29]. Denote by \( m(\gamma_k) \) the multiplicity of \( \gamma_k \). It is well known that \( m(\gamma_k) \) is finite. Denote \( B(E_k) \) the set of bases of \( E_k \), which is identical to the orthogonal group \( O(m(\gamma_k)) \). Denote the set of the corresponding orthonormal bases of \( L^2(M) \) by

\[
B(M, g) = \prod_{k=1}^{\infty} B(E_k).
\]

Fix \( a \in B(M, g) \) and \( t > 0 \), the authors in [7] define the map \( \Phi^a_t \), which maps \( x \in M \) to the Hilbert space \( \ell^2 \) by:

\[
\Phi^a_t : x \mapsto \left(\frac{1}{(4\pi)^{d/2}t^{(d+1)/2}}e^{-\gamma t} \phi_{\ell}(x)\right)_{\ell=1}^{\infty},
\]

where \( \phi_{\ell} \) is the eigenfunction of \( \Delta_g \) with eigenvalue \( \gamma_{\ell} \). Note that the DM is the discretization of \( \Phi^a_t \), so we may abuse the notation and call \( \Phi^a_t \) DM. With the map \( \Phi^a_t \), we are able to define a new affinity between pairs of points:

\[
d_{DM,t}(x,y) := \|\Phi^a_t(x) - \Phi^a_t(y)\|_{\ell^2}.
\]

Again, note that the DD is the discretization of \( d_{DM,t} \), so we may abuse the notation and call \( d_{DM,t}(x,y) \) the DD between \( x \) and \( y \). In addition to showing that we can define a metric on \( \mathcal{M}_{d,k,D} \) so that \( \mathcal{M}_{d,k,D} \) becomes precompact, it has been shown that the diffusion mapping satisfies the following “almost isometric” property [51]:

**Theorem D.4.** Let \((M, g) \in \mathcal{M}_{d,k,D} \). For all \( t > 0 \), \( \Phi^0_t \) is diffeomorphic. Furthermore, suppose \( x, y \in M \) so that \( x = \exp_y v \), where \( v \in T_y M \). When \( \|v\|^2 \ll t \ll 1 \) we have the following asymptotic expansion of DD:

\[
d_{DM,t}(x,y) = \|v\|^2 + O(t\|v\|^2).
\]

This theorem, when combined with the above spectral convergence theorem, says that the DD provides an accurate estimation of the geodesic between two close points.

Theorem D.4 and the manifold sparsity property lead to the following practical fact – if we are allowed a positive small error when we estimate the geodesic distance, we do not need to recover the whole eigenstructure. Instead, the first few eigenvalues and eigenvectors are enough. To be more precise, if we allow the error in the geodesic distance estimate to be \( \delta > 0 \) so that \( t\|v\|^2 \ll \delta^2 < \|v\|^2 \), then we need only the first \( n' \) eigenvalues and eigenfunctions so that

\[
\frac{1}{(4\pi)^{d/2}t^{d+1}} \sum_{\ell=n'+1}^{\infty} e^{-2\gamma_{\ell} t} \leq \delta^2.
\]

The above results partially explains why the commonly applied truncated diffusion map of the dataset with diffusion time \( t \) and truncation \( \delta' \) is stable to noise. In particular, the truncated DM, denoted as \( \Phi_{t,n}^{\delta'} \), is defined as:

\[
\Phi_{t,n}^{\delta'} : X_i \mapsto (e^{-\nu_{DM,t,n}(i)}^{m})_{i=2}^{m},
\]

where \( m > 0 \) is the first eigenvalue satisfies \( \nu_{DM,m} \leq \delta' \).
D-2 Vector Diffusion Map

As a generalization of DM, VDM shares several similar properties of DM. We discuss the VDM on the frame bundle and its associated tangent bundle here \[^{[51]}\]. For VDM associated with the more general principal bundle structure, we refer the reader to \[^{[53]}\]. Suppose the signal is noise free. Note that when the signal is contaminated by noise, unlike the DM algorithm, the noise model is complicated by the way we prepare the input to the VDM algorithm. One particular example is the class averaging algorithm based on VDM shown in Section \[^{[A]}\]. Assume that Assumption \[^{[D.1]}\] holds and the graph \(G_M\) is given. In addition, we assume a group-valued function \(b : V \rightarrow O(d)\), and define \(g(i, j) := b(i)^T P_{X_i, X_j} b(j)\), where \(P_{X_i, X_j}\) presents the parallel transport \[^{[22]}\] of the vector field from \(X_i\) to \(X_j\). Define \(w(i, j) := K_h(X_i, X_j)\), for a given \(h > 0\). Build up \(S\) and \(D\) from the graph \(G_M\) by \(^{[5]}\) and \(^{[4]}\). The matrix \(C := D^{-1}S\) is referred to as Graph Connection Laplacian \(^{[2]}\). It is an operator acting on \(v \in \mathbb{R}^{nd}\) by

\[
(Cv)[i] = \frac{\sum_{j: (i, j) \in E} K_h(X_i, X_j) g(i, j) v[j]}{\sum_{j: (i, j) \in E} K_h(X_i, X_j)},
\]

(D-15)

where \(v[j] := (v((j - 1)d + 1), \ldots, v(jd)) \in \mathbb{R}^d\). The geometrical meaning of the above quantities deserves some discussions. First, note that the tangent plane \(T_{X_i} M\) is isomorphic to \(\mathbb{R}^d\) \(^{[22]}\), but they are different. \(b(i) \in O(d)\) is actually a basis of the tangent plane of \(T_{X_i} M\), which maps \(\mathbb{R}^d\) isomorphically to \(T_{X_i} M\). The parallel transport \(P_{X_i, X_j}\) is a geometrical generalization of the notion “parallel translation” in the Euclidean space. By definition, it is an isometric mapping \(T_{X_i} M\) to \(T_{X_j} M\). As a result, \(g(i, j) \in O(d)\) is an isometric map from \(\mathbb{R}^d\) to \(\mathbb{R}^d\), and geometrically it maps the coordinate of a vector field at \(X_j\) to the vector field \(b(j)v[j]\), parallel transport \(b(j)v[j]\) to \(X_i\), and then evaluate the coordinate of \(P_{X_i, X_j} b(j)v[j]\) with related to the basis \(b(i)\).

We consider the following symmetric matrix which is similar to \(C\)

\[
\tilde{S} = D^{-1/2}SD^{-1/2}.
\]

Since \(\tilde{S}\) is symmetric, it has a complete set of eigenvectors \(v_{\text{VDM},1}, v_{\text{VDM},2}, \ldots, v_{\text{VDM},nd}\) and eigenvalues \(\mu_{\text{VDM},1}, \mu_{\text{VDM},2}, \ldots, \mu_{\text{VDM},nd}\), where the eigenvalues are ordered by \(|\mu_{\text{VDM},1}| \geq |\mu_{\text{VDM},2}| \geq \ldots \geq |\mu_{\text{VDM},nd}|\). Then, fix \(t > 0\), we define the vector diffusion mapping (VDM) \(V_{t,n}\), which is a map from \(X\) to \(\mathbb{R}^{(nd)^2}\) by

\[
V_{t,n} : X_i \mapsto ((\mu_{\text{VDM},1} \mu_{\text{VDM},r})^i (v_{\text{VDM},l}[i], v_{\text{VDM},r}[i]))_{l,r=1}^{nd},
\]

where \(v_{\text{VDM},l}[i]\) is a \(d\)-dim vector containing the \((i - 1)d + 1\)-th entry to the \(id\)-th entry of \(v_{\text{VDM},l}\). With this map, the Hilbert-Schmidt norm block of \(\tilde{S}\) satisfies

\[
\|\tilde{S}^{2t}(i, j)\|_H^2 = (V_{n,n}(X_i), V_{n,n}(X_j)),
\]

that is, \(\|\tilde{S}^{2t}(i, j)\|_H^2\) becomes an inner product for the finite dimensional Hilbert space. The reason we need to consider \(\tilde{S}^{2t}\) and not \(\tilde{S}^{t}\) for instance is that all eigenvalues \(\mu_{\text{VDM},t}\) of \(\tilde{S}\) reside in the interval \([-1, 1]\), and we can not guarantee the positivity of \(\mu_{\text{VDM},t}\) when \(n\) is finite. We can then define the vector diffusion distance (VDD) to quantify the affinity between nodes \(i\) and \(j\):

\[
d_{\text{VDM},t,n} := \|V_{t,n}(X_i) - V_{t,n}(X_j)\|^2.
\]

We claim that asymptotically \(C\) also satisfies the manifold sparsity property, that is,

\[
\mu_{\text{VDM},t} = O(e^{-\ell^2/d}),
\]

where the constants depends \(d\), the lower bound of the Ricci curvature \(k\) and the diameter \(\ell\).

As is discussed in the DM, we need the pointwise convergence and the spectral convergence of \(C\).

**Theorem D.5** (VDM Pointwise Convergence). Suppose Assumption \[^{[D.1]}\] hold and \(X \in C^4(TM)\). For all \(X_i \notin M_{\sqrt{n}}\) with high probability (w.h.p.)

\[
b(i)(C\overline{X})[i] = X(X_i) + h\mu_2 \nabla^2 X(X_i) + O(h^2) + O\left(\frac{1}{n^{1/2}h^{d/4-1/2}}\right)
\]

38
where \( \mathbf{X} \in \mathbb{R}^{n,d} \) and \( \mathbf{X}[i] = b(i)^{-1} \mathbf{X}(X_i) \). For all \( X_i \in M_{\sqrt{n}} \), we have w.h.p.

\[
    b(i)(C\mathbf{X})[i] = \mathbf{X}(X_i) + O(\sqrt{n})P_{x_0}x_0\nabla_{\partial_i} \mathbf{X}(x_0) + O(h) + O\left(\frac{1}{n^{1/2}h^{d/4-1/2}}\right),
\]

where \( x_0 = \text{argmin}_{y \in \mathcal{M}} d(X_i, y) \) and \( \nabla_{\partial_i} \) is the derivative in the normal direction.

Define an operator \( T_{\text{VDM},h} : C(TM) \rightarrow C(TM) \),

\[
    T_{\text{VDM},h} \mathbf{X}(y) := \frac{\sum_{j:(i,j) \in E} K_h(y, X_j)P_{y,X_j} \mathbf{X}(X_j)}{\sum_{j:(i,j) \in E} K_h(y, X_j)},
\]

where \( \mathbf{X} \in C(TM) \).

**Theorem D.6 (VDM Spectral Convergence).** Suppose Assumption \[D.7\] and fix \( t > 0 \). Denote \( \mu_{\text{VDM},t,h,i} \) to be the \( i \)-th eigenvalue of \( T_{\text{VDM},h} \) with the associated eigenvector \( \mathbf{X}_{\text{VDM},t,h,i} \). Also denote \( \mu_{t,i} > 0 \) to be the \( i \)-th eigenvalue of the heat kernel of the Connection Laplacian \( e^{t\nabla^2} \) with the associated eigen-vector field \( \mathbf{X}_{t,i} \). We assume that both \( \mu_{\text{VDM},t,h,i} \) and \( \mu_{t,i} \) decrease as \( i \) increases, respecting the multiplicity. Fix \( i \in \mathbb{N} \). Then there exists a sequence \( h_n \rightarrow 0 \) such that

\[
    \lim_{n \rightarrow \infty} \mu_{\text{VDM},t,h_n,i} = \mu_{t,i}
\]

and

\[
    \lim_{n \rightarrow \infty} \| \mathbf{X}_{\text{VDM},t,h_n,i} - \mathbf{X}_{t,i} \|_{L^2(TM)} = 0
\]

in probability.

We denote the spectrum of \( \nabla^2 \) by \( \{-\lambda_l\}_{l=0}^{\infty} \), where \( 0 = \lambda_0 \leq \lambda_1 \leq \ldots \), and the corresponding eigenspaces by \( F_l := \{ \mathbf{X} \in L^2(TM) : \nabla^2 \mathbf{X} = -\lambda_l \mathbf{X}\} \), \( l = 0, 1, \ldots \). Note that the Weyl’s theorem holds for the connection Laplacian, that is,

\[
    \mathcal{N}(\mu) \sim \frac{1}{(4\pi)^{d/2}(2d/2 + 1)!} \mu^{d/2},
\]

where \( \mathcal{N}(\mu) \) is the number of eigenvalues of \( \nabla^2 \) less than \( \mu > 0 \). Thus, we have the consequence that \[60\]

\[
    \lambda_j \geq A'(d, k, D) j^{2/d},
\]

where \( j \in \mathbb{N} \) and \( A'(d, k, D) \) is the universal constant depending only on \( d \), the lower bound of the Ricci curvature \( k \) and the diameter \( D \). Note that \( \mu_{t,i} = e^{-t\lambda_i} \). In other words, we have \( \mu_{t,j} = O(e^{-t^{2/d}}) \), and hence the claim.

We have similar results for VDM and VDD based on spectral geometry \[60\]. It is well known \[29\] that \( \text{dim}(F_l) < \infty \), the eigen-vector-fields are smooth and form a basis for \( L^2(TM) \), that is, \( L^2(TM) = \bigoplus_{l \in \mathbb{N} \cup \{0\}} F_l \), i.e., it is the completion of \( \bigoplus_{l \in \mathbb{N} \cup \{0\}} F_l \) according to the measure associated with \( g \). To simplify the statement, we assume that \( \lambda_l \) for each \( l \) are simple and \( X_l \) is a normalized basis of \( F_l \). Note that in general \( \lambda_0 \) may not exist: a simple example is found considering \( S^2 \) with the standard metric. Denote \( B(F_k) \) the set of bases of \( F_k \), which is identical to the orthogonal group \( O(\text{dim}(F_k)) \). Denote the set of the corresponding orthonormal bases of \( L^2(TM) \) by

\[
    B(TM, g) = \Pi_{k=1}^\infty B(F_k).
\]

Fix \( a \in B(TM, g) \) and \( t > 0 \), define the **vector diffusion map** (VDM), denoted as \( V_t^a \), which maps \( x \in M \) to the Hilbert space \( \ell^2 \) by

\[
    V_t^a : x \mapsto \left( \frac{1}{\sqrt{\text{det}(4\pi)^{d/2}(d+1)/2}} e^{-(\lambda_n + \lambda_m)t/2} \langle \mathbf{X}_n(x), \mathbf{X}_m(x) \rangle \right)_{n,m=1}^\infty,
\]

where \( \mathbf{X}_m \) is the eigen-vector field of \( \nabla^2 \) with the eigenvalue \( \lambda_m \). With the VDM, we define a new affinity referred to as **vector diffusion distance** (VDD) between pairs of points by

\[
    d_{\text{VDM},t}(x, y) := \| V_t^a(x) - V_t^a(y) \|_{\ell^2},
\]

It has been shown that the VDM satisfies the following “almost isometric” property \[51\ [60]:
Theorem D.7. Let \((M,g) \in \mathcal{M}_{d,k,D}\). For all \(t > 0\), the VDM \(V_t^a\) is diffeomorphic. Furthermore, suppose \(x,y \in M\) so that \(x = \exp_y v\), where \(v \in T_y M\). When \(\|v\|^2 < t \ll 1\) we have the following asymptotic expansion of VDD:

\[
d^2_{VDM,t}(x,y) = \|v\|^2 + O(t\|v\|^2)
\]

(D-18)

References

[1] T. W. Anderson. *An introduction to multivariate statistical analysis*. Wiley Series in Probability and Statistics. Wiley-Interscience [John Wiley & Sons], third edition, 2003.

[2] A. S. Bandeira, A. Singer, and D. A. Spielman. A Cheeger Inequality for the Graph Connection Laplacian. *submitted*, 2013. arXiv:1204.3873 [math.SP].

[3] M. Belkin and P. Niyogi. Laplacian Eigenmaps for Dimensionality Reduction and Data Representation. *Neural. Comput.*, 15(6):1373–1396, June 2003.

[4] M. Belkin and P. Niyogi. Towards a theoretical foundation for Laplacian-based manifold methods. In *Proceedings of the 18th Conference on Learning Theory (COLT)*, pages 486–500, 2005.

[5] M. Belkin and P. Niyogi. Convergence of laplacian eigenmaps. In *Advances in Neural Information Processing Systems 19: Proceedings of the 2006 Conference*, volume 19, page 129. The MIT Press, 2007.

[6] P. Bérand. *Spectral Geometry: Direct and Inverse Problems*. Springer, 1986.

[7] P. Bérand, G. Besson, and S. Gallot. Embedding riemannian manifolds by their heat kernel. *Geom. Funct. Anal.*, 4:373–398, 1994. 10.1007/BF01896401.

[8] N. Berline, E. Getzler, and M. Vergne. *Heat Kernels and Dirac Operators*. Springer, 2004.

[9] R. Bhatia. *Matrix analysis*, volume 169 of *Graduate Texts in Mathematics*. Springer-Verlag, 1997.

[10] R. L. Bishop and R. J. Crittenden. *Geometry of Manifolds*. Amer Mathematical Society, 2001.

[11] M. Bolla. Distribution of the eigenvalues of random block-matrices. *Linear Algebra Appl.*, 377:219–240, 2004.

[12] S. Chatterjee. A simple invariance principle. 2005. Available at http://arxiv.org/abs/math/0508213.

[13] S. Chatterjee. A generalization of the Lindeberg principle. *Ann. Probab.*, 34(6):2061–2076, 2006.

[14] M.-Y. Cheng and H.-T. Wu. Local linear regression on manifolds and its geometric interpretation. *J. Am. Stat. Assoc.*, 2013. Accepted for publication.

[15] X. Cheng and A. Singer. The Spectrum of Random Inner-product Kernel Matrices. *submitted*, 2012.

[16] Y. S. Chow and H. Teicher. *Probability theory*. Springer Texts in Statistics. Springer-Verlag, New York, third edition, 1997. Independence, interchangeability, martingales.

[17] R. R. Coifman and S. Lafon. Diffusion maps. *Appl. Comput. Harmon. Anal.*, 21(1):5–30, 2006.

[18] R. R. Coifman and M. Maggioni. Diffusion wavelets. *Appl. Comput. Harmon. Anal.*, 21(1):53 – 94, 2006.

[19] M. Cucuringu, A. Singer, and D. Cowburn. Eigenvector synchronization, graph rigidity and the molecule problem. *Information and Inference: A Journal of the IMA*, 1:21–67, 2012.

[20] L. de Haan and A. Ferreira. *Extreme value theory*. Springer Series in Operations Research and Financial Engineering. Springer, New York, 2006. An introduction.
[21] H. Dette and B. Reuther. Random block matrices and matrix orthogonal polynomials. J. Theoret. Probab., 23(2):378–400, 2010.

[22] M.P. do Carmo and F. Flaherty. Riemannian Geometry. Birkhauser Boston, 1992.

[23] M. L. Eaton. Multivariate statistics. Institute of Mathematical Statistics Lecture Notes—Monograph Series, 53. Institute of Mathematical Statistics, 2007. Reprint of the 1983 original.

[24] N. El Karoui. Concentration of measure and spectra of random matrices: Applications to correlation matrices, elliptical distributions and beyond. Ann. Appl. Probab., 19(6):2362–2405, December 2009.

[25] N. El Karoui. On information plus noise kernel random matrices. Ann. Stat., 38(5):3191–3216, 2010.

[26] N. El Karoui. The spectrum of kernel random matrices. Ann. Stat., 38(1):1–50, February 2010.

[27] N. El Karoui and H. Koesters. Geometric sensitivity of random matrix results: consequences for shrinkage estimators of covariance and related statistical methods. Submitted to Bernoulli, 2011. Available at arXiv:1105.1404 (68 pages).

[28] D. Giannakis, P. Schwander, and A. Ourmazd. The symmetries of image formation by scattering. i. theoretical framework. Opt. Express, 20(12):12799–12826, Jun 2012.

[29] P. Gilkey. The Index Theorem and the Heat Equation. Princeton, 1974.

[30] V. L. Girko. A matrix equation for the resolvents of random matrices with independent blocks. Teor. Veroyatnost. i Primenen., 40(4):741–753, 1995.

[31] I. Gohberg, S. Goldberg, and N. Krupnik. Traces and determinants of linear operators, volume 116 of Operator Theory: Advances and Applications. Birkhäuser Verlag, 2000.

[32] F. Götze and A. N. Tikhomirov. Limit theorems for spectra of random matrices with martingale structure. In Stein’s method and applications, volume 5 of Lect. Notes Ser. Inst. Math. Sci. Natl. Univ. Singap., pages 181–193. Singapore Univ. Press, Singapore, 2005.

[33] R. Hadani and A. Singer. Representation Theoretic Patterns in Three-Dimensional Cryo-Electron Microscopy II. The Class Averaging Problem. Found. Comput. Math., 11(5):589–616, 2011.

[34] M. Hein, J. Audibert, and U. von Luxburg. From graphs to manifolds - weak and strong pointwise consistency of graph Laplacians. In Proceedings of the 18th Conference on Learning Theory (COLT), pages 470–485, 2005.

[35] R. A. Horn and C. R. Johnson. Matrix analysis. Cambridge University Press, Cambridge, 1990. Corrected reprint of the 1985 original.

[36] D. Martinec and T. Pajdla. Robust rotation and translation estimation in multiview reconstruction. In CVPR. IEEE Computer Society, 2007.

[37] F. Méromi. A spectral notion of gromov-wasserstein distance and related methods. Appl. Comput. Harmon. Anal., 30(3):363 – 401, 2011.

[38] F. Mezzadri. How to Generate Random Matrices from the Classical Compact Groups. Notices of the AMS, 54(5):592–604, 2006.

[39] R. J. Muirhead. Aspects of multivariate statistical theory. John Wiley & Sons Inc., 1982. Wiley Series in Probability and Mathematical Statistics.

[40] B. Nadler, S. Lafon, R. R. Coifman, and I. G. Kevrekidis. Diffusion maps, spectral clustering and eigenfunctions of Fokker-Planck operators. In Y. Weiss, B. Schölkopf, and J. Platt, editors, Adv. Neur. In., volume 18, pages 955–962, Cambridge, MA, 2006. MIT Press.
[41] B. Nadler, S. Lafon, R. R. Coifman, and I. G. Kevrekidis. Diffusion maps, spectral clustering and reaction coordinates of dynamical systems. *Appl. Comput. Harmon. Anal.*, 21(1):113–127, 2006.

[42] P. Niyogi, S. Smale, and S. Weinberger. Finding the homology of submanifolds with high confidence from random samples. In *Twentieth Anniversary Volume*, pages 1–23. Springer New York, 2009.

[43] T. Oraby. The spectral laws of Hermitian block-matrices with large random blocks. *Electron. Comm. Probab.*, 12:465–476, 2007.

[44] M. Ovsjanikov, J. Sun, and L. Guibas. Global intrinsic symmetries of shapes. In *Proceedings of the Symposium on Geometry Processing*, SGP ’08, pages 1341–1348. Eurographics Association, 2008.

[45] A. Pajor and L. Pastur. On the limiting empirical measure of eigenvalues of the sum of rank one matrices with log-concave distribution. *Studia Math.*, 195(1):11–29, 2009.

[46] R. Rashidi Far, T. Oraby, W. Bryc, and R. Speicher. On slow-fading MIMO systems with nonseparable correlation. *IEEE Trans. Inform. Theory*, 54(2):544–553, 2008.

[47] M. Reuter. Hierarchical Shape Segmentation and Registration via Topological Features of Laplace-Beltrami Eigenfunctions. *Int. J. Comput. Vision*, 89:287–308, 2010.

[48] R. M. Rustamov. Laplace-beltrami eigenfunctions for deformation invariant shape representation. In *Proceedings of the fifth Eurographics symposium on Geometry processing*, SGP ’07, pages 225–233. Eurographics Association, 2007.

[49] J. W. Silverstein and Z. D. Bai. On the empirical distribution of eigenvalues of a class of large-dimensional random matrices. *J. Multivariate Anal.*, 54(2):175–192, 1995.

[50] A. Singer. From graph to manifold Laplacian: The convergence rate. *Appl. Comput. Harmon. Anal.*, 21(1):128–134, 2006.

[51] A. Singer and H.-T. Wu. Vector diffusion maps and the connection Laplacian. *Comm. Pure Appl. Math.*, 65(8):1067–1144, 2012.

[52] A. Singer and H.-T. Wu. 2-d tomography from noisy projections taken at unknown random directions. *SIAM J. Imaging Sci.*, 6(1):136–175, 2013.

[53] A. Singer and H.-T. Wu. Spectral convergence of the connection laplacian from random samples. submitted, 2013.

[54] A. Singer, Zhao Z., Shkolnisky Y., and Hadani R. Viewing angle classification of cryo-electron microscopy images using eigenvectors. *SIAM J. Imaging Sci.*, 4(2):723–759, 2011.

[55] B. Sonday, A Singer, and I. G. Kevrekidis. Noisy dynamic simulations in the presence of symmetry: data alignment and model reduction. *Computers & Mathematics with Applications*, 65(10):1535–1557, 2013.

[56] J. Sun, M. Ovsjanikov, and L. Guibas. A concise and provably informative multi-scale signature based on heat diffusion. In *Proceedings of the Symposium on Geometry Processing*, SGP ’09, pages 1383–1392. Eurographics Association, 2009.

[57] M. Talagrand. *Spin glasses: a challenge for mathematicians*, volume 46 of *Ergebnisse der Mathematik und ihrer Grenzgebiete. 3. Folge. A Series of Modern Surveys in Mathematics [Results in Mathematics and Related Areas. 3rd Series. A Series of Modern Surveys in Mathematics]*. Springer-Verlag, Berlin, 2003. Cavity and mean field models.

[58] U. von Luxburg, M. Belkin, and O. Bousquet. Consistency of spectral clustering. *Ann. Stat.*, 36(2):555–586, 2008.
[59] L. Wang and A. Singer. Exact and Stable Recovery of Rotations for Robust Synchronization. *Information and Inference: A Journal of the IMA*, 2013. accepted for publication.

[60] H.-T. Wu. Embedding riemannian manifolds by the heat kernel of the connection laplacian. *submitted*, 2012.