bioRxiv preprint doi: https://doi.org/10.1101/2020.05.31.126649; this version posted June 1, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

ROBUST AND SCALABLE MANIFOLD LEARNING VIA LANDMARK DIFFUSION FOR LONG-TERM MEDICAL SIGNAL PROCESSING

CHAO SHEN, YU-TING LIN, AND HAU-TIENG WU

Abstract. Motivated by analyzing long-term physiological time series, like 24 hours electrocardiogram, we design a computationally efficient spectral embedding algorithm that is suitable to handle "big data". We measure the affinity between any pair of two points *via a set of landmarks*, which is composed of a small number of points, and "diffuse" on the dataset via the landmark set to achieve a spectral embedding. We coined the algorithm RObust and Scalable Embedding via LANdmark Diffusion (ROSELAND). The algorithm is applied to study the arterial blood pressure waveform dynamics during a liver transplant operation lasting for 12 hours long. The proposed Roseland could help researchers handle big datasets or long-term physiological signals if the spectral embedding is considered, and the landmark idea beyond Roseland could be applied to speed up other algorithms.

1. Introduction

Learning from data has been an intriguing topic in many scientific fields, particularly the biomedical field. A learning procedure is taking existing data as the "knowledge" or "experience" to interpret incoming data and help field experts make decision. In the clinical medicine, it has been widely debated that eventually a well trained artificial intelligence (AI) by a proper learning procedure could replace physicians. While it is not impossible, we believe that there is still a long way to go. No matter how, a general consensus is that such a system would augment physicians, and release physicians' precious time by automatizing some routine and time-consuming workflow so that physicians.

Motivated by its importance, various datasets of different data types from clinics have been extensively explored from this learning perspective, like electronic health record [31], medical imaging [21], genomic data [37], biomedical waveforms (or time series, signals) [19], etc. To our knowledge, however, how to extract intrinsic dynamics underlying biomedical waveforms for clinical usage is relatively less discussed, and most of existing literature focus on simplifying the waveform information into few scalars [8, 35, 2, 44]. While this has been successfully applied to clinical medicine, we may loss information encoded in the original waveform.

One solution to depict intrinsic dynamics directly from the original waveform is obtaining as many "features" as possible, and applying machine learning algorithms [23]. Another solution is applying manifold learning algorithms to the original physiological waveforms in a recent study [29, 47]. The basic idea in [29, 47] is truncating the physiological waveform into pieces according to some rules, and then apply the spectral embedding algorithm, like the diffusion maps (DM) [10], to embed those pieces into a finite dimensional Euclidean space, which represents

bioRxiv preprint doi: https://doi.org/10.1101/2020.05.31.126649; this version posted June 1, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

the intrinsic dynamics. If the physiological waveform is embedded into the three dimensional Euclidean space, the physiological waveform is converted into a three dimensional image so that users can visualize the waveform from a different perspective.

It is commonly believed that more data lead to more comprehensive "knowledge" in the learning procedure, and hence more informative features and better visualization of the intrinsic dynamics. We thus hypothesize that if the technique shown in [29, 47] could be applied to analyze long-term physiological waveforms of length on the order of days or weeks, it will be beneficial to supporting the short-term memory of the human brain when handling long-term physiological waveforms. Specifically, due to the short-term memory limitation, it is easy to overlook information hidden in the long-term waveform; for example, what is the relationship between the waveform in the first hour of the operator and the 7th hour? Will the surgeon remember the useful information in the first hour after 5 hours? While this idea is natural, unfortunately, it is prohibited by the computational complexity inherited in most spectral decomposition based machine learning algorithm, including the DM applied in [47].

Recall that the DM is based on the eigendecomposition of the graph Laplacian (GL) matrix. The algorithm has been shown to perform well when the database is "tiny", like in the order of $10^3 \sim 10^4$. However, when the database gets larger, like in the order of 10^6 or above, the algorithm is challenged by the scalability issue. Take electrocardiogram (ECG) into account. There are $10^3 \sim 10^4$ cycles in 1 hour long ECG, and roughly 10^6 cycles in 14 days. In practice, it is natural to consider subsampling the dataset, however, we may loss information. As a result, although the DM works well and provides alternative clinical information [47], it is limited to datasets of length about one hour. In this paper, we propose a novel spectral embedding algorithm to resolve this issue. The solution will be generic and not limited to analyze physiological waveforms.

1.1. **Existing solutions.** There have been several solutions toward this scalability challenge. One usual technique is the k-nearest neighbor (kNN) scheme. However, it is not robust to noise. Specifically, when the dataset is noisy and the neighboring information is *not* provided, obtaining a reliable kNN information is challenging. A randomized kNN approach is recently considered in [30]. Another practical solution is directly subsampling the dataset, and then recovering the information of interest by the *Nyström extension* [13]. This approach is also called the *Nyström low-rank approximation* [7], the *kernel extension method* [18], or in general the *interpolative decomposition* [32]. This approach has several theoretical backups, for example [7], and has been widely applied. While it works well for some missions, this approach is limited by the information loss during the subsampling process. Yet another approach is speeding up the matrix decomposition by taking randomization into account. For example, we can construct a thin matrix by taking a random subset of columns of the GL matrix and speed up the algorithm by taking the singular value decomposition (SVD) into account [32]. While this approach has been widely applied, to the best of our knowledge, we have limited knowledge about how it helps the spectral embedding algorithms, and how robust it is to the inevitable noise.

bioRxiv preprint doi: https://doi.org/10.1101/2020.05.31.126649; this version posted June 1, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

1.2. **Our solution/contribution.** Unlike the above, in this paper we propose a novel algorithm that resolves two common challenges when we apply spectral embedding algorithms—robustness and scalability. In short, we measure the affinity between any pair of two points via a *landmark set*, which is composed of a small number of points. With this affinity, the embedding is carried out via the spectral decomposition. We coin the proposed algorithm the *RObust and Scalable Embedding via LANdmark Diffusion* (Roseland).

The algorithm is intuitive and can be summarized in three steps. First, we find a "small" subset of points from the whole dataset, either randomly or by design, or collect a separate clean point cloud of small size, which we call a *landmark set*. Second, we construct an affinity matrix recording the affinities between points in the whole dataset and the landmark set, and normalize it properly. This normalized affinity matrix is thin; that is, there are fewer columns than rows. Third, evaluate the singular vectors and singular values of the normalized affinity matrix, and embed the dataset using the singular vectors and singular values. As we will make clear soon, this algorithm is directly related to the diffusion process.

From the application perspective, in addition to providing a series of numerical simulations comparing the Roseland with other algorithms, we also demonstrate how it works in the physiological time series collected from a liver transplant operation lasting for more than 10 hours.

1.3. **Organization of the paper.** In Section 2 we recall the DM algorithm, introduce the Roseland, and summarize a closely related algorithm, the *HKC* algorithm, which stands for initials of three authors in [20]. In Section 3, we provide numerical results with simulated datasets. In Section 4, we illustrate how to apply the proposed algorithm to study the arterial blood pressure waveform during the liver transplant surgery. In Section 5, discussion and conclusion are provided.

2. The proposed Roseland algorithm and relevant algorithms

In this section, we assume that we have a data set $\mathcal{X} = \{x_i\}_{i=1}^n \subseteq \mathbb{R}^q$. Take a set $\mathcal{Y} = \{y_k\}_{k=1}^m$, which might or might not be a subset of $\mathcal{X}$. We call $\mathcal{Y}$ the *landmark set*. Fix a non-negative kernel function $K : \mathbb{R}_{\geq 0} \to \mathbb{R}_+$ with proper decay and regularity; for example, a Gaussian function.

2.1. **The proposed algorithm – Roseland.** We introduce the Roseland. First we construct a *landmark-set affinity matrix* $W^{(\mathrm{r})} \in \mathbb{R}^{n\times m}$, which is defined as

$$W^{(\mathrm{r})}_{ik} = K_\epsilon(x_i, y_k) := K\left(\frac{\|x_i - y_k\|_{\mathbb{R}^q}}{\sqrt{\epsilon}}\right). \tag{1}$$

That is, the (i,k)-th entry of $W^{(\mathrm{r})}$ is the similarity between the i-th data point and j-th landmark, and clearly the larger the distance between two points (or the two points are less similar), the smaller the similarity. Next compute a diagonal matrix $D^{(\mathrm{R})}$ as

$$D^{(\mathrm{R})}_{ii} := e_i^\top W^{(\mathrm{r})}(W^{(\mathrm{r})})^\top \mathbf{1}\,, \tag{2}$$

where $\mathbf{1}$ is a $n\times 1$ vector with all entries 1, and e_i is the unit vector with 1 in the i-th entry. $D^{(\mathrm{R})}_{ii}$ is called the degree of the i-th data point x_i. Intuitively, it represents

bioRxiv preprint doi: https://doi.org/10.1101/2020.05.31.126649; this version posted June 1, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

how strong x_i is attached to the data set. With $W^{(\mathrm{r})}$ and $D^{(\mathrm{R})}$, we evaluate the SVD of $(D^{(\mathrm{R})})^{-1/2}W^{(\mathrm{r})}$:

$$(D^{(\mathrm{R})})^{-1/2}W^{(\mathrm{r})} = U\Lambda V^\top, \tag{3}$$

where the singular values $\sigma_1 \geq \sigma_2 \geq \ldots \geq \sigma_m \geq 0$ are on the diagonal of the diagonal matrix Λ. Set $\bar{U} := (D^{(\mathrm{R})})^{-1/2}U$. Take $q' \in \mathbb{N}$ so that $q' \leq m$. Let $\bar{U}_{q'} \in \mathbb{R}^{n\times q'}$ to be a matrix consisting of the second to the $(q'+1)$-th columns of $\bar{U}$ and $L_{q'} := \mathtt{diag}(\sigma_2^2, \ldots, \sigma_{q'+1}^2)$. Finally we define the Roseland embedding as

$$\Phi_t^{(\mathrm{R})} : x_i \mapsto e_i^\top \bar{U}_{q'}(L_{q'})^t\,, \tag{4}$$

where $t > 0$ is the chosen diffusion time, in other words, the i-th data point x_i is embedded using the i-th row of $\bar{U}_{q'}$ entry-wisly rescaled by $[\sigma_2^{2t}, \cdots, \sigma_{q'+1}^{2t}]$. See Algorithm 1 for a summarization of the Roseland algorithm. We thus define the associated *Roseland diffusion distance* (RDD) by

$$D_t^{(\mathrm{R})}(x_i, x_j) := \|\Phi_t^{(\mathrm{R})}(x_i) - \Phi_t^{(\mathrm{R})}(x_j)\|_{\mathbb{R}^{q'}}\,. \tag{5}$$

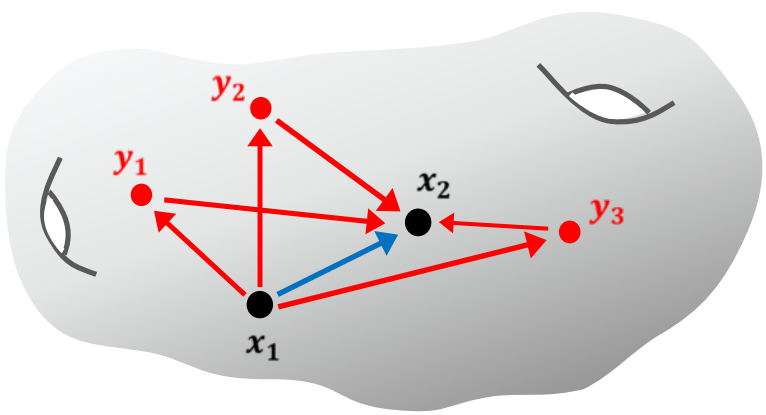

FIGURE 1. Main idea of Roseland: to measure the similarity between x_1 to x_2, instead of diffuse from x_1 to x_2 directly, we take a detour and first diffuse x_1 to the landmarks y_1, y_2, y_3, and then diffuse from the landmarks to x_2.

Note that the Roseland induces a new affinity matrix on the data set $\mathcal{X}$ via

$$W^{(\mathrm{R})} := W^{(\mathrm{r})}(W^{(\mathrm{r})})^\top \in \mathbb{R}^{n\times n}\,, \tag{6}$$

where $W^{(\mathrm{r})}$ is the landmark-set affinity matrix (1). We call $W^{(\mathrm{R})}$ the *landmark-affinity matrix*, which is positive and positive-definite. We remark that traditional affinity matrices between data points are often constructed from *one* global pre-fixed kernel K, while in Roseland we cannot find a global fixed kernel $\bar{K}$ and a bandwidth $\bar{\epsilon} > 0$ so that $W_{ij}^{(\mathrm{R})} = \bar{K}(\|x_i - x_j\|/\bar{\epsilon})$ for all i, j in general. Hence Roseland can provide us with more dynamic similarity measurements between data points based on the landmarks. Also note that $A^{(\mathrm{R})} := (D^{(\mathrm{R})})^{-1}W^{(\mathrm{R})}$ is a transition matrix on $\mathcal{X}$, which means each of the row in $A^{(\mathrm{R})}$ sums to 1. This is because $D^{(\mathrm{R})}$ is the degree matrix associated with the landmark-affinity matrix $W^{(\mathrm{R})}$ by construction. Hence, we can view the similarity measure between data points x_i and x_j as a Markov or diffusion process through the landmarks. A systematic convergence

bioRxiv preprint doi: https://doi.org/10.1101/2020.05.31.126649; this version posted June 1, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Algorithm 1 The pseudo-code of Roseland.

procedure INPUT(data set $\mathcal{X} = \{x_i\}_{i=1}^n \subset \mathbb{R}^q$, landmark set $\mathcal{Y} = \{y_i\}_{i=1}^m \subset \mathbb{R}^q$, kernel K, bandwidth $\sigma > 0$, embedded dimension $q' \in \mathbb{N}$, $0 \leq \alpha \leq 1$, and diffusion time $t > 0$)
  Construct the affinity matrix $W^{(\mathrm{r})}$ w.r.t. the landmark set.
  Construct the degree matrix $D^{(\mathrm{R})}$.
  Run SVD $(D^{(\mathrm{R})})^{-1/2} W^{(\mathrm{r})} = U \Lambda V^\top$, where $U \in O(n)$ and $V \in O(m)$. Denote singular values as $\sigma_1 \geq \sigma_2 \geq \ldots \geq 0$.
  Set $\bar{U} = (D^{(\mathrm{R})})^{-1/2} U$.
  Let $\bar{U}_{q'} \in \mathbb{R}^{n \times q'}$ be the second to the $(q'+1)$-th columns of $\bar{U}$. Set $L_{q'} := \mathtt{diag}(\sigma_2^2, \sigma_3^2, \ldots, \sigma_{q'+1}^2)$.
  Embed $\mathcal{X}$ via $\Phi_t^{(\mathrm{R})} : x_i \mapsto e_i^\top \bar{U}_{q'} L_{q'}^t$
  Output $\{\Phi_t^{(\mathrm{R})}(x_i)\}_{i=1}^n \subset \mathbb{R}^{q'}$.
end procedure

analysis and robustness analysis of Roseland under the manifold assumption can be found in [39].

2.2. **Graph Laplacian and Diffusion Maps.** We now review the GL and compare Roseland with the well-known algorithm diffusion map (DM) [10]. First, pre-fix a kernel function K and a bandwidth parameter $\epsilon > 0$. Then, compute the affinity matrix $W \in \mathbb{R}^{n \times n}$ by

$$W_{ij} := K\left(\frac{\|x_i - x_j\|_{\mathbb{R}^q}}{\sqrt{\epsilon}}\right) \tag{7}$$

and the corresponding *degree matrix* $D \in \mathbb{R}^{n \times n}$, which is a diagonal matrix defined as $D_{ii} := \sum_{j=1}^n W_{ij}$. For a fixed $\alpha \in [0,1]$, the α-*normalized* affinity matrix $W^{(\alpha)} \in \mathbb{R}^{n \times n}$ [10] is defined as $W_{ij}^{(\alpha)} := \frac{W_{ij}}{D_{ii}^\alpha D_{jj}^\alpha}$, where $W_{ij}^{(\alpha)}$ is called the α-*normalized affinity* between x_i and x_j. Note that $W^{(0)} = W$ defined in (7). In some applications when we want to remove the density effect caused by data sampling, we set $\alpha = 1$. With the α-normalized affinity matrix $W^{(\alpha)}$, one can analogously define the associated degree matrix $D^{(\alpha)} \in \mathbb{R}^{n \times n}$ by $D_{ii}^{(\alpha)} := \sum_{j=1}^n W_{ij}^{(\alpha)}$. The GL is defined as $L^{(\alpha)} := I - A^{(\alpha)}$, where

$$A^{(\alpha)} := (D^{(\alpha)})^{-1} W^{(\alpha)} \tag{8}$$

is the associated transition matrix. Clearly, $A^{(\alpha)}$ is row stochastic, and it defines a random walk on the dataset $\mathcal{X}$. We mention that the transition matrix $(D^{(\mathrm{R})})^{-1} W^{(\mathrm{R})}$ in the Roseland algorithm can be viewed as an alternative way of constructing a Markov process on the dataset $\mathcal{X}$. Since $A^{(\alpha)}$ is similar to $(D^{(\alpha)})^{-1/2} W^{(\alpha)} (D^{(\alpha)})^{-1/2}$ we can find its eigendecomposition with eigenvalues $1 = \lambda_1 > \lambda_2 \geq \ldots \geq \lambda_n$ and the associated eigenvectors $\phi_1, \ldots, \phi_n$. Denote ϕ_i the i-th right eigenvector of $A^{(\alpha)}$. Recall in Roseland, we perform the SVD decomposition in (3), which is a parallel step of the eigen-decomposition of $\bar{A}^{(0)}$.

Among various algorithms, we focus on the well-known DM algorithm that we shall see to be closely related to Roseland. With the spectral decomposition of the GL, the chosen normalization α, embedding dimension q' and *diffusion time* t, the

bioRxiv preprint doi: https://doi.org/10.1101/2020.05.31.126649; this version posted June 1, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

DM embeds $\mathcal{X}$ via the map

$$\Phi_t : x_i \mapsto e_i^\top \tilde{U}_{q'}^{(\alpha)} (\Lambda_{q'}^{(\alpha)})^t \in \mathbb{R}^{q'}, \tag{9}$$

where $\tilde{U}_{q'}^{(\alpha)} \in \mathbb{R}^{n\times q'}$ to be a matrix consisting of the second to the $(q'+1)$-th columns of $\tilde{U}^{(\alpha)}$ and $\Lambda_{q'}^{(\alpha)} := \mathtt{diag}(\lambda_2, \ldots, \lambda_{q'+1})$. The diffusion distance (DD) with the diffusion time $t > 0$ is defined as

$$D_t(x_i, x_j) := \|\Phi_t(x_i) - \Phi_t(x_j)\|_{\mathbb{R}^{q'}} . \tag{10}$$

Recall the Roseland embedding (4) and the RDD (5) to notice the close relation between the DM embedding and the DD.

As it is expensive to perform eigen-decomposition of dense matrices, one common practice of DM or general spectral embedding methods is to use the kNN scheme to construct a rather sparse affinity matrix; that is, set $W_{ij} = 0$ when x_j is not within the first k nearest neighbors of x_i, where k is chosen by the user. Another way is to use a compactly supported kernel K. For example, $K(t)$ is 1 when $t \in [0, 1]$ and 0 when $t > 1$.

2.3. **The HKC algorithm.** HKC was proposed for the texture separation problem. The authors proposed to first divide an image into a collection of small patches, from which to choose a subset consists of specific patterns of interest as the *reference set*. Note that in [20], the reference set plays the same role as the landmark set in Roseland. Then one can construct an affinity matrix associated of the patches based on the landmark set. HKC is the closest algorithm to the Roseland among others. However, the normalization in HKC is different from the Roseland, and this difference turns out to be significant.

We now summarize the HKC algorithm [20]. Firstly, form the affinity matrix between the data set and the landmark set just like (1) in the Roseland; that is, set $W^{(\mathrm{HKC})} = W^{(\mathrm{r})}$. HKC then compute a $n \times n$ diagonal matrix by $D_{ii}^{(\mathrm{HKC})} = \sum_{j=1}^m W_{i,j}^{(\mathrm{HKC})}$, where $i = 1, \ldots, n$. Then, convert $W^{(\mathrm{HKC})}$ to be row stochastic by:

$$A^{(\mathrm{HKC})} = (D^{(\mathrm{HKC})})^{-1} W^{(\mathrm{HKC})} \in \mathbb{R}^{n\times m} . \tag{11}$$

One should notice the difference between Roseland and HKC when computing the degree matrix. In Roseland, the degree matrix is computed from the row sum of the matrix $W^{(\mathrm{HKC})}(W^{(\mathrm{HKC})})^\top$ instead of the row sum of $W^{(\mathrm{HKC})}$. Therefore, Roseland defines a Markov process on the data set, but HKC does not. Indeed, due to the normalization (11), $\bar{W}^{(\mathrm{HKC})}$ can only be viewed as a new affinity matrix on $\mathcal{X}$, and is different from the $(D^{(\mathrm{R})})^{-1/2} W^{(\mathrm{R})} (D^{(\mathrm{R})})^{-1/2}$ in Roseland. This normalization step plays a significant role. Finally, HKC embeds the data via the eigenvectors ψ_j of the matrix $\bar{W}^{(\mathrm{HKC})} = A^{(\mathrm{HKC})}(A^{(\mathrm{HKC})})^\top \in \mathbb{R}^{n\times n}$, which can be computed efficiently by

$$\psi_j^{(\mathrm{HKC})} = (\lambda_j^{(\mathrm{HKC})})^{-1/2} A^{(\mathrm{HKC})} \phi_j^{(\mathrm{HKC})} \tag{12}$$

where $\phi_j^{(\mathrm{HKC})}$ is the j-th eigenvectors of the matrix $\tilde{W}^{(\mathrm{HKC})} := (A^{(\mathrm{HKC})})^\top A^{(\mathrm{HKC})} \in \mathbb{R}^{m\times m}$ associated with the eigenvalue $\lambda_j^{(\mathrm{HKC})}$. The denominator $(\lambda_j^{(\mathrm{HKC})})^{1/2}$ is to ensure that $\|\psi_j^{(\mathrm{HKC})}\|_2 = \|\phi_j^{(\mathrm{HKC})}\|_2$. In summary, we see that the HKC algorithm is close to Roseland with a different normalization.

bioRxiv preprint doi: https://doi.org/10.1101/2020.05.31.126649; this version posted June 1, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

2.4. **Nyström Extension.** Another widely applied algorithm aiming to scale up spectral embedding is the Nyström extension [3, 18, 49, 4, 11]. The idea is simple but effective. First, run the eigen-decomposition on a subset of the given dataset. Then, extend the eigenvectors to the whole dataset. There are some variants of the Nyström extension method, for example, [5]. A direct application is out of sample embedding.

We are interested in applying the Nyström extension to the eigen-decomposition of the transition matrix defined in (8) for the spectral embedding purpose. Similarly, we consider the symmetric kernel matrix $M = D^{-1/2}WD^{-1/2}$. Note that if we want to apply Nyström extension directly on M, we would have to compute the affinity matrix W and the degree matrix D, which is expensive and the kNN scheme is needed. We thus follow the existing literature [26, 41] and apply the following modified Nyström extension.

Suppose we have n data points. First, run the DM on a chosen subset, also called the landmark set, which is of size $L = n^\beta$, where $\beta \in (0, 1)$. Denote the affinity matrix associated with this landmark set as W_L, and run the eigen-decomposition of the matrix $D_L^{-1/2}W_L D_L^{-1/2} = V_L \mathcal{L}_L V_L^\top$, where D_L is the degree matrix associated with W_L and $\mathcal{L}_L = \mathtt{diag}\begin{bmatrix}\ell_1 & \dots & \ell_L\end{bmatrix} \in \mathbb{R}^{L\times L}$. Let $\widetilde{U}_L = D_L^{-1/2}V_L$ to be the eigenvectors of $D_L^{-1}W_L$. We then extend it to the rest $n - L$ points by:

$$\check{U}_{\text{ext}} = D_{n-L}^{-1} E \widetilde{U}_L \mathcal{L}_L^{-1} \in \mathbb{R}^{(n-L)\times L}\,, \tag{13}$$

where $E \in \mathbb{R}^{(n-L)\times L}$ is the affinity matrix between the remaining $n-L$ data points and the landmark set. In other words, $E_{i,j}$ is the similarity between point x_i in the remaining dataset and x_j in the landmark set; D_{n-L} is a $(n-L)\times(n-L)$ diagonal matrix such that $D_{n-L}(i,i) = \sum_{j=1}^{L} E(i,j)$. Hence the eigenvectors to be used to embed the whole dataset is:

$$\check{U} = \begin{bmatrix}\widetilde{U}_L \\ \check{U}_{\text{ext}}\end{bmatrix} = \begin{bmatrix} D_L^{-1} & \\ & D_{n-L}^{-1}\end{bmatrix}\begin{bmatrix} W_L \\ E\end{bmatrix}\widetilde{U}_L \mathcal{L}_L^{-1}\,. \tag{14}$$

Roughly speaking, the embedding coordinates of a data point x outside the landmark set is simply the average of all of the landmarks' embeddings, weighted by the similarity between x and all the landmarks.

While it is slightly different from the original Nyström extension, we still call it the Nyström extension. Note that in practice, we only need to calculate W_L and E instead of W and D, which is more efficient in the sense of both time and spatial complexities. With the estimated eigenvectors on the whole dataset, we can define the associated embedding and hence the distance just as in Roseland and DM. Specifically, suppose we have $\check{U} = \begin{bmatrix}\tilde{u}_1 & \dots & \tilde{u}_L\end{bmatrix} \in \mathbb{R}^{N\times L}$ and $\check{\mathcal{L}} = \mathtt{diag}\begin{bmatrix}\ell_1 & \dots & \ell_L & 0 & \dots & 0\end{bmatrix} \in \mathbb{R}^{n\times n}$. Then we can define the associated embedding by

$$\Phi_t^{(\text{Nyström})} : x_i \mapsto e_i^\top \check{U}_{q'} \check{\mathcal{L}}_{q'}^t\,, \tag{15}$$

where $t > 0$ is the chosen diffusion time, $\check{U}_{q'} \in \mathbb{R}^{n\times q'}$ to be a matrix consisting of $\tilde{u}_2, \dots, \tilde{u}_{q'+1}$ and $\check{\mathcal{L}}_{q'} := \mathtt{diag}(\ell_2, \dots, \ell_{q'+1})$.

2.5. **Complexity analysis.** Let n be the size of the dataset and n^β the size of the landmark set, where $\beta \le 1$ throughout and set $k = n^\beta$ whenever kNN scheme is applied. For dense affinity matrix, the spatial complexity of the DM is $O(n^2)$. If

bioRxiv preprint doi: https://doi.org/10.1101/2020.05.31.126649; this version posted June 1, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

kNN scheme is applied, the spatial complexity of the DM becomes $O(n^{1+\beta})$. On the other hand, no matter what kernel is chosen, compactly supported or not, the spatial complexity of the Roseland and the Nyström extension is $O(n^{1+\beta})$.

For the computational complexity, it can be divided into two parts. The first part is forming the affinity matrix and the corresponding degree matrix; the second part is performing the eigen-decomposition or SVD. In the ordinary DM, the construction of the affinity matrix and the degree matrix is $O(n^2)$. If the kNN construction is considered and the k-d tree based algorithm is applied, the averaged time complexity of constructing the affinity matrix and the degree matrix is $O(n\log(n) + n^{1+\beta}) = O(n^{1+\beta})$. In the Nyström extension, the construction of the W_L and hence its degree matrix is $O(n^{2\beta})$ when $L = n^\beta$ for $\beta \leq 1$, while the construction of E and D_{n-L} is $O(n^{1+\beta})$. Thus, the first part complexity for the Nyström extension is $O(n^{2\beta} + n^{1+\beta})$. In the Roseland, the construction of the landmark-set affinity matrix and its associated degree matrix is $O(n^{1+\beta})$. For the second part, it falls in the discussion of the complexity of the general eigen-decomposition and SVD. For a symmetric kernel matrix $M \in \mathbb{R}^{N\times N}$, the eigen-decomposition complexity is usually $O(N^3)$,[1] and when M is k sparse, where $k \leq n$, the complexity can be improved to $O(N^{2+\eta'})$ for an arbitrary $\eta' > 0$ when $k \leq N^{0.14}$ [50]. In our application, even if we make $k \leq N^{0.14}$, the eigen-decomposition of the M is roughly $O(N^{2+\eta'})$. On the other hand, for a matrix of size $N \times N'$, where $N \geq N'$, then the complexity of the SVD for is $O(NN'^2)$. Hence, the overall computational complexity for the ordinary DM is $O(n^3)$ and is $O(n^\omega)$ for the DM with the kNN scheme, where $\omega > 2$ depends on the chosen β, $O(n^{1+\beta} + n^{3\beta})$ for the Nyström extension, and $O(n^{1+2\beta})$ for the Roseland. The complexity of the HKC is the same as that of the Roseland. To summarize, both the Nyström extension and the Roseland are more efficient than the ordinary DM with or without the kNN scheme. While the Roseland is not faster than the traditional Nyström extension approach, it is comparable, particularly for small β.

2.6. **Roseland is robust to noise.** It is well known that when the dataset is noisy, spectral embedding algorithms might lead us to a bad, or even misleading result, due to inevitable noise, unless the parameters are properly chosen. In [16], efforts have been made to stabilize the algorithm; for example, consider the complete graph or choose a large number of kNN, and force the random walk non-lazy. It is clear that this solution is not scalable. Roseland, on the other hand, automatically enjoys the desired robustness property since Roseland measures similarities between data points by diffusing through *all* landmarks. This step can be viewed as a surrogate of knowing *true neighbors* in the kNN scheme, and it explains the robustness of Roseland.

2.7. **Reference set as subset of the data.** In cases that we may not be able to acquire additional data points as landmark set but have to select the landmark set from the available dataset, we propose to first sample m landmarks, denoted by $\mathcal{Y}$, from $\mathcal{X}$ so that $\mathcal{Y}$ is independent of $\mathcal{X} \setminus \mathcal{Y}$. Then, we apply the Roseland on $\mathcal{X} \setminus \mathcal{Y}$ using $\mathcal{Y}$ as landmarks, and extend the embedding to $\mathcal{Y}$ by the Nyström extension.

[1]Theoretically, it can reach $O(N^{\omega+\eta})$, where the N^ω part comes from the algorithm of matrix multiplication, and an arbitrary $\eta > 0$ [14]. Note that when M is dense, $\omega = \omega_0 \approx 2.376$ [12]. However, the implied constant in these asymptotic is too large and cannot be practical [27].

bioRxiv preprint doi: https://doi.org/10.1101/2020.05.31.126649; this version posted June 1, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

When $|\mathcal{Y}| \ll |\mathcal{X}|$, the discrepancy of this approach and the independence setup with original Roseland is negligible, and will asymptotically vanish.

3. Numerical Simulation Results

To illustrate how the Roseland performs, in addition to showing the dimension reduction and geometric recovery results, we also compare the results with the Nyström extension and the HKC. For a fair comparison, in all the following simulations, the subset used in the Nyström extension and the reference set used in the HKC to embed the dataset are the same as the landmark set used in the Roseland. As a result, the ranks of the matrices associated with the Roseland, the HKC and the Nyström extension are the same.

3.1. **Scalability of the Roseland.** We take the dataset consists of random projections of the two-dimensional Shepp-Logan phantom [42]. A phantom is a 2-dim image function ψ compactly supported on $\mathbb{R}^2$ without any symmetry assumption. It is commonly applied in medical imaging society as a benchmark. Suppose we uniformly sample n points from S^1, $\theta_1 \dots \theta_n \in S^1$, as the projection angles. Then we generate a high dimensional data set by taking the Radon transform of ψ, denoted as $R_\psi : S^1 \to L^2(\mathbb{R})$, followed by discretizing the projection image into $p \in \mathbb{N}$ points; that is, we have the dataset $\mathcal{X} := \{D_p R_\psi(\theta_i)\}_{i=1}^n \subset \mathbb{R}^p$, where D_p is the discretization operator. We refer readers with interest to [42] for details. In this simulation, we fix the number of discretization points $p = 128$ and let the number of projections n vary. We run the DM, the Roseland, the HKC and the Nyström extension with $n = 10,000$ and $m = n^\beta$, where $\beta = 0.5$, and show the 3-dim embedding of $\mathcal{X}$ in Figure 2. Clearly, both the DM and the Roseland recover the S^1 structure, while the Roseland is distorted. On the other hand, the HKC and the Nyström extension are confused and lead to erroneous embeddings. The computational times of different algorithms with $\beta = 0.3$ are also shown for a comparison. When $n = 1,280,000$, the Roseland can finish in about 2.5 minutes in an ordinary laptop.

3.2. **Robustness of the Roseland.** We compare performance of the Nyström extension, the HKC, and the Roseland from the aspect of spectral embedding when the data is noisy. We consider the standard S^1 model, which is the one-dimensional canonical S^1 embedded in the first two coordinates of $\mathbb{R}^{100}$, since all ground truths can be analytically calculated. Specifically, we uniformly sample $n = 90,000$ points from the S^1 to be the dataset and independently sample another $m = 300$ points uniformly to be the landmark set; that is, $\beta = 0.5$. Then, embed all points to $\mathbb{R}^p$, where $p = 100$, and add independent Gaussian noise ϵ_i to both the dataset and the landmark set, where ϵ_i are i.i.d. sampled from $\mathcal{N}(0, \frac{1}{\sqrt{p}} I_{p\times p})$. The visualization results are shown in Figures 3. Clearly, while the Nyström extension and the HKC embed S^1 successfully, the embedding by the Roseland is cleaner.

Next, the recovered eigenvectors are shown in Figures 4 and 5. Clearly, the Nyström method can only successfully recover first few eigenfunctions of the Laplace-Beltrami operator (visually, only the first 8 look reasonably), while the HKC and the Roseland can recover more eigenfunctions (visually, the first 10 are reasonably well). Since HKC is not designed to recover the Laplace-Beltrami operator of S^1, we do not consider it in Figure 5. Compared with eigenvectors, only the first 7

bioRxiv preprint doi: https://doi.org/10.1101/2020.05.31.126649; this version posted June 1, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

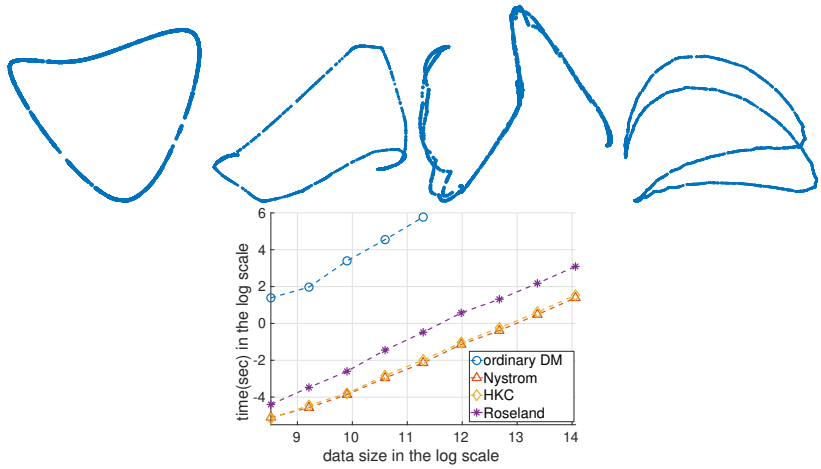

FIGURE 2. The random projection data from the phantom image, where the data size is $n = 10,000$ and the dimension is $p = 128$. In the top row, we take $\beta = 0.5$ for the Roseland, the Nyström, and the HKC. In the bottom row, we show the relationship between the computational time and the data size. Top row, from left to right: the DM embedding, the Roseland, the HKC, and the Nyström extension. All embeddings are 3-dim, and have been rotated to optimize the visualization. Bottom row: the runtime comparison of various algorithms when $\beta = 0.3$ The x-axis is in the natural log unit, and the largest database size is $1,280,000$.

FIGURE 3. Left: noisy data and noisy subset (only the first two coordinately are shown). Middle left: the Nyström embedding. Middle right: the HKC embedding. Right: the Roseland embedding.

or 8 eigenvalues of Laplace-Beltrami operator can be well approximated in both Nyström method and Roseland.

"Visually" the qualities of the first two non-trivial eigenfunctions of the Nyström extension, HKC and Roseland are similar, but the qualities of embeddings are different. To understand this discrepancy, we consider the following quantities. Note that the first two non-trivial eigenvectors, $v_1, v_2 \in \mathbb{R}^n$ from either the Nyström extension, the HKC or Roseland, if successfully recovered the eigenfunctions of the Laplace-Beltrami operator, should be $\sin(\theta+\phi)$ and $\cos(\theta+\phi)$ for some $\phi \in (0, 2\pi]$ respectively. Here, the phase ϕ comes from the uncertainty nature of the spectral embedding methods. We then plot $\arctan(v_1(i)/v_2(i))$ and $\sqrt{v_1(i)^2 + v_2(i)^2}$ against θ_i, where θ_i is the angle of the i-th sampled point. The results are shown in Figure

bioRxiv preprint doi: https://doi.org/10.1101/2020.05.31.126649; this version posted June 1, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

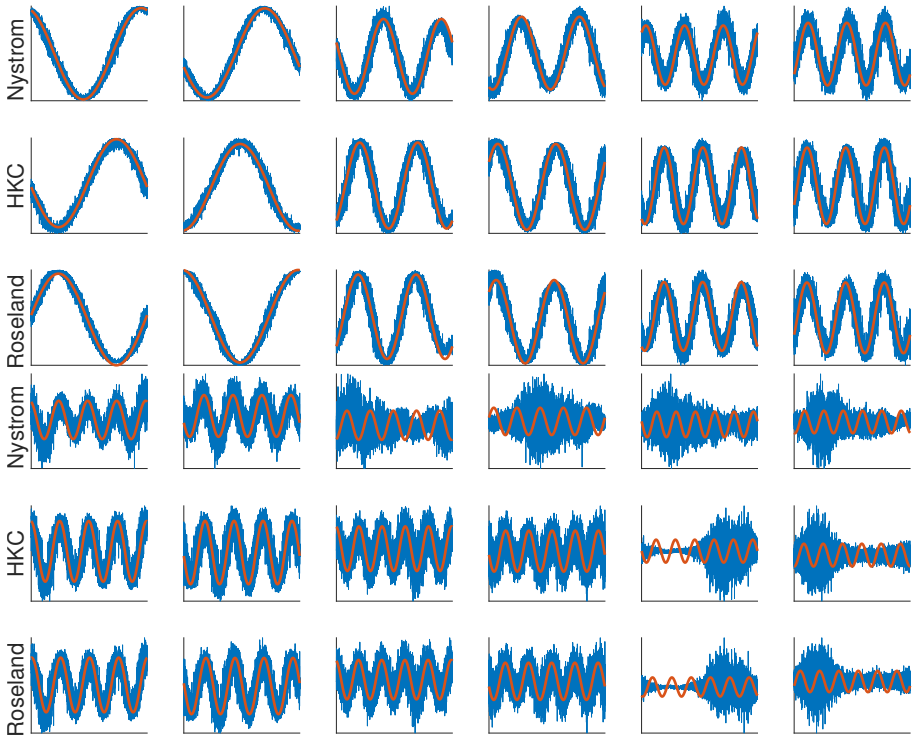

FIGURE 4. Noisy data set and subset. Superimpose the top 12 non-trivial eigenvectors by the Nyström, the HKC and the Roseland with the ground truth (superimposed in red). Top three rows: the top 6 eigenvectors; bottom three rows: the 7th to the 12th eigenvectors.

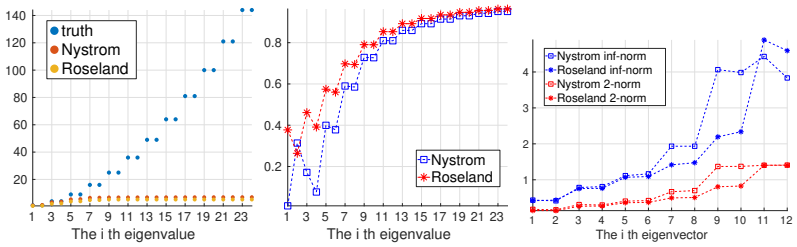

FIGURE 5. Illustration of the Nyström method and the Roseland on the noisy dataset and landmark set. Left: the top 18 non-trivial eigenvalues by the Nyström and Roseland with the ground truth. Middle: relative error of eigenvalues. Right: relative L^∞ and L^2 error of the top 12 non-trivial eigenvectors by the Nyström and Roseland with the ground truth.

6. Clearly, the amplitude eigenvectors of the Nyström extension and the HKC fluctuates more than those of the Roseland, while the phase recovery qualities are

bioRxiv preprint doi: https://doi.org/10.1101/2020.05.31.126649; this version posted June 1, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

similar. This difference comes from the different normalization steps of Roseland and HKC.

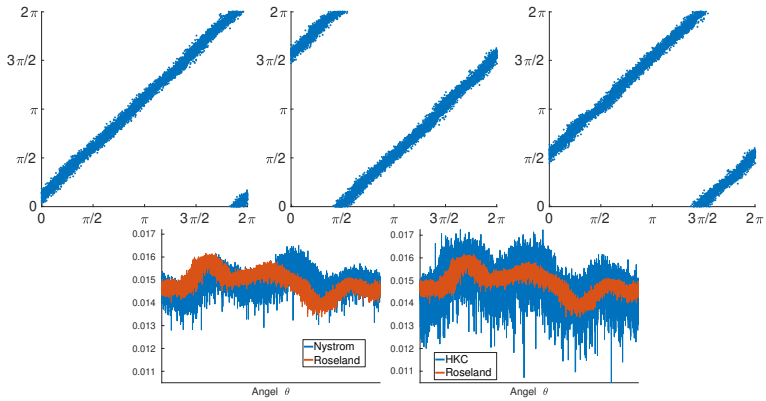

FIGURE 6. Noisy data set and subset. Top left: the phase of the embedding by the Roseland. Top middle: the phase of the embedding by the HKC. Top right: the phase of the embedding by the Nyström extension. Bottom left: the amplitude of the embedding by the Roseland and the Nyström. Bottom right: the amplitude of the embedding by the Roseland and the HKC. The phase of the embedding is determined by $\arctan(v_1(i)/v_2(i))$ and the amplitude is determined by $\sqrt{v_1(i)^2 + v_2(i)^2}$, where v_1 and v_2 are the first non-trivial eigenvectors determined by the Nyström extension or Roseland.

4. Liver transplant analysis

4.1. **Background.** Liver transplant surgery is the only life-saving treatment for patients in certain medical conditions. It is a challenging surgical procedure, and significant medical resource, experience and dedication are needed. During the surgery, the clamping of major vessels and the subsequent vascular anastomosis bring huge impacts on the recipient's circulation system [38]. As a better understanding of the cardiovascular dynamics during the procedure may help optimize the intraoperative management, commercial monitoring instruments based on real-time arterial blood pressure (ABP) waveform analysis has been introduced. However, they been questioned subsequently for its performance in liver transplant surgery [6, 45, 40]. Thus, obtaining useful information from the ABP waveform in liver transplant is still a challenging problem.

We shall briefly detour to discuss the ABP waveform analysis before showing our results. It is well known that the ABP waveform morphology reflects the physical effect of fluid dynamics, like wave reflection. Deep into the root of the matter, the waveform morphology reflects more complicated dynamical interaction over the physiological network [33]. For example, the sympathetic neural activity in response to surgical stress produces vasoconstriction and increases cardiac contractility. Epinephrine and norepinephrine releasing may increase the blood volume returning to the heart to enhance heart contractility by the Frank-Starling law,

bioRxiv preprint doi: https://doi.org/10.1101/2020.05.31.126649; this version posted June 1, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

and the subsequent increased blood pressure induces baroreflex to reduce the heart beat rate. The renin-angiotensin system also work in tendem with the autonomic nerve system to regulate the vascular wall tension while the fluid volume changes drastically during the vascular procedure of liver transplantation. The above physiological mechanisms contain sequential and parallel control loops forming a complex network-like interaction. Traditionally, various extracted features, either landmark measurements in the time domain [34] or quantities in the frequency domain [48], serve as the input for the subsequent pulse waveform analysis. These designated features are supposed to reflect underlying physiological information, or those parameters driving the network interaction.

However, it is reasonable to suspect that information hidden in the finer scale might be ignored via the above approach, and hence finer structure of network dynamics is overlooked, particularly when the physiology is disturbed. It is thus reasonable to argue that taking the whole waveform into account might provide more complimentary information compared with those traditional parameters. On the other hand, due to the short-term memory nature of human brain, it is challenging to visualize and directly utilize the dynamics encoded in the ABP waveform on the large scale. Motivated by handling the above challenges, including finding finer information on the short scale, and exploring the dynamics on the large scale, in our previous research, we reported a solution under the manifold learning framework, and showed that the DM can extract rich information directly from the *raw cardiovascular waveform* [47]. The novelty in [47] capturing subtle morphological changes that might be overlooked by the designed features. However, due to the computational barrier intrinsic to the DM, the approach is limited to relatively small dataset.

In this study, we hypothesize that with the help of Roseland, the manifold learning approach shown in [47] can be applied to study the ABP waveform during the liver transplant procedure, and provide hemodynamic information on both the small and large scales. Note that the Roseland is critical in this analysis since the whole period of the surgery can yield more than 10^5 continuous pulses as data points in high dimensional space for a pulse-to-pulse waveform analysis.

4.2. **Material.** The data was collected from an observational study per institutional ethic regulation. We collected physiological signals via the data collection software, S5 collect (GE Healthcare, Chicago, Illinois, United States) from the standard patient monitor instrument, GE CARESCAPE^{TM}B850 (GE Healthcare, Chicago, Illinois, United States). The recorded ABP signal was uniformly sampled at 300 Hz in the instrument and resampled at 500 Hz via the cubic spline interpolation for off-line processing. The signal is of 78,350s long spanning the whole surgical procedure and contains 120,725 pulses.

4.3. **Data analysis.** Denote the ABP waveform as $x^A \in \mathbb{R}^N$. We used the maximum of the first derivation during the ascent of each ABP pulse waveform as a fiducial point. A legitimate ABP pulse is determined by a two-pass algorithm using the following measurements automatically: the peak maximum, the trough minimum, the minimum of difference between the maximum and minimum within the pulse, the pulse width, and the duration to the previous pulse. The thresholds for those measurements are automatically adjusted by a feedback

bioRxiv preprint doi: https://doi.org/10.1101/2020.05.31.126649; this version posted June 1, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

mechanism. Suppose there are L legitimate cycles in x^A. Denote the i-th fiducial point as n_i. Break x^A into $L-1$ segments so that the i-th segment is the i-th ABP pulse containing one waveform cycle. Denote the i-th segment as $\bar{x}_i^A := [x^A(n_i), x^A(n_i+1), \ldots, x^A(n_{i+1})]^T$. Since the duration of each pulse is not constant, we truncated them into an uniform size according to their minimal length $q = \min\{n_{i+1} - n_i + 1\} \in \mathbb{N}$, and get $\hat{x}_i^A := [x^A(n_i), x^A(n_i+1), \ldots, x^A(n_i+q-1)]^T$. Next, normalize $\hat{x}_i^A$ by removing the mean and setting the variance to 1 to separate the blood pressure information from the normalized ABP pulse, and denote the normalized ABP pulse as $x_i^A \in \mathbb{R}^q$. Derived from the ABP signal, we get the data set $\mathcal{X}^A = \{[(x_i^A)^T\ (x_{i-1}^A)^T\ (x_{i-2}^A)^T]^T\}_{i=3}^L \subseteq \mathbb{R}^{3q}$. We assume that $\mathcal{X}^A$ can be well approximated by a low dimensional manifold, referred to as the *wave-shape manifold* [29]. To apply the Roseland, the landmark set $\mathcal{Y}^A = \{y_k^A\}_{k=1}^m$, where $m = \lfloor n/600 \rfloor$, was chosen from $\mathcal{X}^A$ via setting $y_k^A = x_{600k}^A$.

4.4. **Results.** The dataset consists of $L = 120,725$ legitimate cycles of length $q = 334$, and hence $120,723$ data points in $\mathcal{X}^A$. Thus, $m = 201$ and note that $201 = 120,723^{0.453}$. The total computation time of the Roseland algorithm is less than 40s on an ordinary personal computer (CPU: Intel Core i5-7500, operation system: Microsoft Windows 10 64-bit home edition, programming platform and language: Microsoft Visual Studio Community version 2019, .NET framework 4.8, and C#, LAPACK software library: Intel Math Kernel Library 2020 Initial Release), while the estimated computation time using traditional DM algorithm based on eigendecomposition would be more than one day.

The embedding result is shown in Figure 7. The successive pulses evolve with time and constitute a trajectory on the manifold presented as a 3D embedding (Fig.7, panel A). The trajectory visits different locations during different steps of the liver transplant procedure. Moreover, there is a "clustering" effect in the embedding, which is enhanced by the imposed color that encodes the temporal information. We can thus visualize the relationship among different hemodynamic status during different surgery steps. This relationship provides physiological dynamics on the large scale. We emphasize that while we can easily read the waveform, but it is not easy to perceive the dynamics and organize them with only human eyes and brain (Fig.7, panel B and C), particularly when the signal is long.

We further quantify the trajectory in different surgical phases as well as the phase transition periods in which the trajectory moves in fast pace. We consider the following different hemodynamic phases during the liver transplant procedure, particularly those that phase transition happens with violent physiological changes take place—First, the occlusion of blood inflow to the "old" liver (to be replaced), performed by the cross-clamp of the inferior vena cava, the largest vein of the human body; second, the start of the circulatory connection from the graft (the new liver organ) to the circulation system as the blood flow starts in the portal vein; third, the start of the connection between hepatic artery and the graft. All these transitions drastically affect the cardiovascular system via the changes of fluid volume and electrolyte composition.

For each hemodynamic phase, we embed the ABP waveforms by the Roseland embedding into 10-dim Euclidean space; that is, $q' = 10$, and evaluate the geometric center of all beats during that phase. Then, we define the distance between two hemodynamic phases groups by measuring the RDD between their geometric centers. The quantitative measurement is expressed as mean and 95 % confidence

bioRxiv preprint doi: https://doi.org/10.1101/2020.05.31.126649; this version posted June 1, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

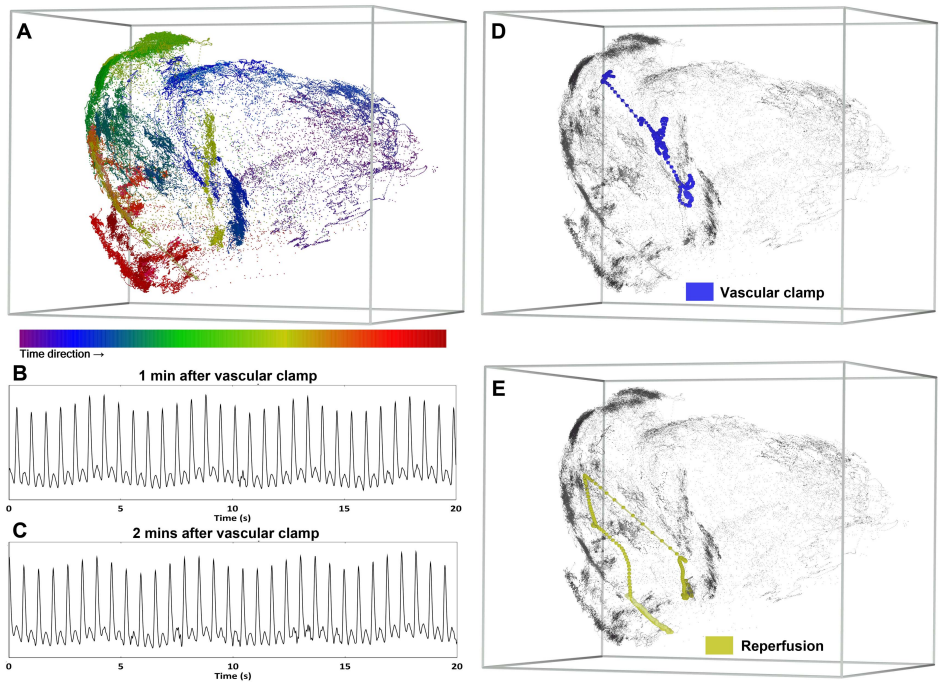

FIGURE 7. The 3D embedding (panel A) of pulse-to-pulse pressure waveforms collected from a 78,350s ABP signal (120,725 pulses) during the liver transplantation procedure. The embedded pulses are labeled by colors encoding the time. The color helps visualize the ever-changing trajectory formed from successive pulse waveforms. The embedding is clustered and different clusters are related to different stages. During transition phases of inferior vena cava cross clamping (panel B, C), the ABP tracings provide little clues with respect to the subtle waveform information and its long-term evolving, while the Roseland algorithm reveals the fast paced movements (panel D). To signify the physiological dynamics associated with the vascular clamp (panel D) and reperfusion (panel E) events, pulses in transition phases are labeled with colored linked dots while the rest pulses are not colored. In panels A, D, and E, the grids are drawn to enhance the 3D visualization, and an online supplementary video is provided for more details.

interval after bootstrap resampling without replacement in 100,000 samples. In light of the Newtonian mechanics, we consider two quantities. The *two-point velocity* measures the dynamics on the large scale. The two-point velocity is defined as the ratio of the RDD between two points and the time difference between the two points. We also consider the *trajectory speed*, which measures the dynamics on the small scale. The trajectory speed is defined as the ratio of the path length of the trajectory and the time different between the beginning and ending of the trajectory. The quantitative result (Table 1) shows that during the phase transition, the trajectory moves faster macroscopically (on the large scale) when quantified

bioRxiv preprint doi: https://doi.org/10.1101/2020.05.31.126649; this version posted June 1, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

TABLE 1. Quantitative results of phases and phase transitions from liver transplantation data

Surgical phases	two point velocity	trajectory speed	RDD to previous phase
dissection phase before clamp	0.59	32.77	na
vascular clamp (transition)	5.34	32.32	na
anhepatic phase after clamp	0.84	36.58	2152.6 (2127.9, 2177.2)
anhepatic phase before reperfusion	0.74	17.53	409.8 (386.1, 434.0)
reperfusion (transition)	6.27	41.41	na
neohepatic phase after reperfusion	0.64	13.26	984.8 (976.4, 993.3)
before artery anastomosis	0.21	16.65	439.3 (422.5, 456.3)
artery anastomosis (transition)	2.47	22.23	na
after artery anastomosis	0.34	11.87	1818.9 (1802.8, 1834.9)

All numbers expressed by 10^{-5}; speed and velocity unit: s^{-1}

by the two-point velocity, particularly when compared with that within each surgical phase. However, the trajectory speeds, which represents hemodynamics on the microscopic scale (small scale), are similar during the phase transition and within each surgical phase. The numeric results are consistent with visualization from the 3D embedding (Figure 7 and 8).

5. Discussion and conclusion

In this paper, we introduce a new spectral embedding algorithm based on the landmark set. We provide a series of numerical simulation and a real-world application to support the potential of the algorithm.

5.1. **Related work – scalability and robustness.** The recent paper [13] contains a comprehensive review of numerical acceleration techniques for nonlinear dimension reduction, and we refer readers with interest to that work. To handle scalability, an intuitive approach is accelerating the kNN search step. See [13] for a summary and a recently proposed randomized kNN approach [30]. However, it is well known that the kNN scheme is not robust when the dataset is noisy when the neighboring information is not provided. Specifically, it is challenging to estimate pairwise distance robustly, unless we have extra structure to design a robust metric, for example, in the image analysis [9]. If the tangent plane is known, it can help us determine neighbors [46]; however, when the dataset is noisy, the local principle component analysis approach to estimate the tangent space is biased [25]. In short, the kNN is only useful when we have an accurate information about the neighbors.

Another natural approach to handle scalability is accelerating the eigen-decomposition step. For example, we can approximate the kernel decomposition by classical iteration-based algorithms [22]. We can also evaluate the matrix decomposition by designing a randomized algorithm [36].

For the robustness issue, one naive idea is "denoising" the dataset before applying any algorithm. However, it is in general an independent challenging problem. Under the manifold setup, researchers have proposed several algorithms to denoise the dataset. For example, the "reverse diffusion" scheme [24] and the manifold fitting

bioRxiv preprint doi: https://doi.org/10.1101/2020.05.31.126649; this version posted June 1, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

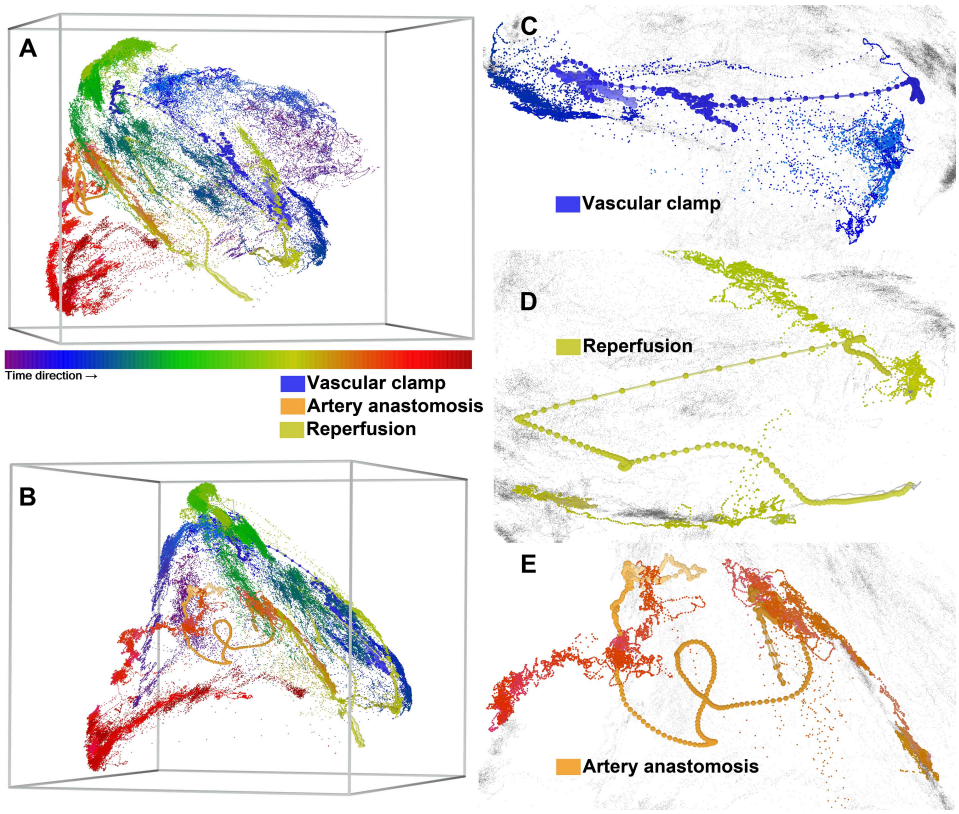

FIGURE 8. The 3D embedding highlights the locations of transition phases (in linked dot) in relation to the whole period of ABP waveforms (smaller dot) in a liver transplant surgery. The color labels the time sequence. Panel B is the horizontally rotated Panel A by 60 degrees for a better visualization of the artery anastomosis transition. These views show geographic relationship between surgical stages and transition phases. Zoom-in views of the 3D embedding of pulse-to-pulse ABP waveform include transition phases of major vascular cross-clamping (panel A), new liver graft reperfusion (panel B), and hepatic artery anastomosis (panel C), which shows fast paced movement in transition phases (enlarged colored dots with line-link). On the other hand, in the statuses immediate before and after transition (colored dots without line-link), we see less movement and the embeddings are clustered. The rest pulses (uncolored small dots) appears in the background.

scheme [17]. We mention that the algorithm might not be scalable, but not too much is known at this moment. Another approach is modifying the random walk scheme to a non-lazy random walk via diffusion to obtain a self-consistency Markov chain. But it is under the assumption that the edge information is known [43], which is not possible in many applications. To our knowledge, the general theory for the robustness of kernel methods was first studied in [15], and the analysis

bioRxiv preprint doi: https://doi.org/10.1101/2020.05.31.126649; this version posted June 1, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

was extended to the large noise setup [16]. The authors proved that the spectral embedding methods can be efficiently stabilized by forcing the random walk to be a non-lazy one on the complete graph. Unfortunately, while it could help stabilize the noise impact, the algorithm is not scalable.

5.2. **Application of the Roseland.** The idea of landmark set have several applications. Here we mention two of them. The VDM [41] is a generalization of DM that aims to encode the group structure when comparing objects. The VDM suffers from the expensive computational cost more than the DM, since the group structure is usually represented as a matrix, which inflates the matrix size. Specifically, if the group structure is represented as a $q \times q$ matrix and we have n objects to compare, then we need to eigendecompose a $nq \times nq$ kernel matrix in the VDM. We expect the landmark idea can be generalized to accelerate the VDM. We will explore this possibility in our future work.

Spectral clustering methods are known to perform well when the classical clustering methods such as k-means and linkage fail [1]. It is well known that the more clusters we need to determine, the more eigenvectors we need [1, 28]. As is shown in the numerical section, the Roseland has the ability to recover more and better eigenvectors, at least compared with the Nystöm extension. This shows the potential of applying the Roseland for the multiway spectral clustering purpose.

5.3. **Implication in Medical Signal Processing.** The liver-transplant example shows the benefit and potential of the Roseland algorithm. As more (longer) data leads to a richer knowledge base, we now have an unprecedented signal processing tool for a long-period signal with complex underlying physiology. In the liver transplant example, the 3D embedding reflects the complex relationship among different surgical phases without *ad hoc* pulse waveform knowledge. This suggests the practicability of handling data governed by complex physiological mechanism. As there is room to be improved in monitoring the hemodynamic status in liver transplant surgery [6, 45, 40], we expect that the proposed waveform analysis would lead to more insights into the hemodynamic status of liver transplant surgery to improve the patient's outcome. We will report research in this direction in our future work. Certainly, the source of knowledge base is not limited to the ABP waveform. Different physiological waveforms can be considered to further enrich the knowledge base. How to simultaneously utilize multimodal physiological waveforms, particularly when the recording is long, is a relatively white area, and we expect that the proposed waveform analysis would form a base toward this goal. We also expect that the similar principal could be applied to study other medical datasets for different medical problems, for example, the long-term outcome of the patient underwent organ transplantation with respect to the immune function, or the genetic predisposition and environment factors with respect to the cancer occurrence.

6. Acknowledgement

The authors acknowledge Dr. Shen-Chih Wang for the fruitful discussion and suggestion. The work of Yu-Ting Lin was supported by the National Science and Technology Development Fund (MOST 108-2115-M-075-001) of Ministry of Science and Technology, Taipei, Taiwan.

bioRxiv preprint doi: https://doi.org/10.1101/2020.05.31.126649; this version posted June 1, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

References

[1] C. Alzate and J.A.K. Suykens. Multiway spectral clustering with out-of-sample extensions through weighted kernel pca. *IEEE transactions on pattern analysis and machine intelligence*, 32(2):335–347, 2008.

[2] A. P. Avolio, L. M. Van Bortel, P. Boutouyrie, and others. Role of pulse pressure amplification in arterial hypertension: experts opinion and review of the data. *Hypertension*, 54(2):375–383, 2009.

[3] M.A. Belabbas and P.J. Wolfe. On landmark selection and sampling in high-dimensional data analysis. *Philosophical Transactions of the Royal Society of London A: Mathematical, Physical and Engineering Sciences*, 367(1906):4295–4312, 2009.

[4] M.A. Belabbas and P.J. Wolfe. Spectral methods in machine learning and new strategies for very large datasets. *Proceedings of the National Academy of Sciences*, 106(2):369–374, 2009.

[5] A. Bermanis, A. Averbuch, and R.R. Coifman. Multiscale data sampling and function extension. *Applied and Computational Harmonic Analysis*, 34(1):15–29, 2013.

[6] G. Biancofiore, L.A.H. Critchley, A. Lee, X.-x. Yang, L. M. Bindi, M. Esposito, M. Bisà, L. Meacci, R. Mozzo, and F. Filipponi. Evaluation of a new software version of the flotrac/vigileo (version 3.02) and a comparison with previous data in cirrhotic patients undergoing liver transplant surgery. *Anesthesia & Analgesia*, 113(3):515–522, 2011.

[7] L.B. Chang, Z. Bai, S.Y. Huang, and C.R. Hwang. Asymptotic error bounds for kernel-based nyström low-rank approximation matrices. *Journal of Multivariate Analysis*, 120:102–119, 2013.

[8] C.-H. Chen, E. Nevo, B. Fetics, P. H. Pak, F. C.P. Yin, W. L. Maughan, and D. A. Kass. Estimation of central aortic pressure waveform by mathematical transformation of radial tonometry pressure: validation of generalized transfer function. *Circulation*, 95(7):1827–1836, 1997.

[9] B. Cheng, J. Yang, S. Yan, Y. Fu, and T.S. Huang. Learning with ℓ^1-graph for image analysis. *IEEE transactions on image processing*, 19(4):858–866, 2009.

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

[11] R.R. Coifman and S. Lafon. Geometric harmonics: a novel tool for multiscale out-of-sample extension of empirical functions. *Applied and Computational Harmonic Analysis*, 21(1):31–52, 2006.

[12] D. Coppersmith and S. Winograd. Matrix multiplication via arithmetic progressions. *J. Symb. Comput.*, 9:251–80, 1990.

[13] W. Czaja, T. Doster, and A. Halevy. An overview of numerical acceleration techniques for nonlinear dimension reduction. In *Recent Applications of Harmonic Analysis to Function Spaces, Differential Equations, and Data Science*, pages 797–829. Springer, 2017.

[14] J. Demmel, I. Dumitriu, and O. Holtz. Fast linear algebra is stable. *Numer. Math.*, 108:59–91, 2007.

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

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

[17] C. Fefferman, S. Ivanov, Y. Kurylev, M. Lassas, and H. Narayanan. Fitting a putative manifold to noisy data. In *Conference On Learning Theory*, pages 688–720, 2018.

[18] C. Fowlkes, S. Belongie, F. Chung, and J. Malik. Spectral grouping using the nystrom method. *IEEE transactions on pattern analysis and machine intelligence*, 26(2):214–225, 2004.

[19] N. Gambarotta, F. Aletti, G. Baselli, and M. Ferrario. A review of methods for the signal quality assessment to improve reliability of heart rate and blood pressures derived parameters. *Medical & biological engineering & computing*, 54(7):1025–1035, 2016.

[20] A. Haddad, D. Kushnir, and R.R. Coifman. Texture separation via a reference set. *Applied and Computational Harmonic Analysis*, 36(2):335–347, 2014.

[21] S. S. Halabi, L. M. Prevedello, J. Kalpathy-Cramer, A. B. Mamonov, A. Bilbily, M. Cicero, I. Pan, L. A. Pereira, R. T. Sousa, and N. Abdala. The rsna pediatric bone age machine learning challenge. *Radiology*, 290(2):498–503, 2019.

[22] N. Halko, P.G. Martinsson, and J.A. Tropp. Finding structure with randomness: Probabilistic algorithms for constructing approximate matrix decompositions. *SIAM review*, 53(2):217–288, 2011.

bioRxiv preprint doi: https://doi.org/10.1101/2020.05.31.126649; this version posted June 1, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

[23] F. Hatib, Z. Jian, S. Buddi, C. Lee, J. Settels, K. Sibert, J. Rinehart, and M. Cannesson. Machine-learning algorithm to predict hypotension based on high-fidelity arterial pressure waveform analysis. *Anesthesiology*, 129(4):663–674, 2018.

[24] M. Hein and M. Maier. Manifold denoising. In B. Schölkopf, J. C. Platt, and T. Hoffman, editors, *Advances in Neural Information Processing Systems 19*, pages 561–568. MIT Press, 2007.

[25] I. M. Johnstone. High dimensional statistical inference and random matrices. In *Proceedings of the International Congress of Mathematicians Madrid, August 22–30, 2006*, pages 307–333, 2007.

[26] S. Lafon, Y. Keller, and R.R. Coifman. Data fusion and multicue data matching by diffusion maps. *IEEE Transactions on pattern analysis and machine intelligence*, 28(11):1784–1797, 2006.

[27] F. Le Gall. Faster algorithms for rectangular matrix multiplication. In *2012 IEEE 53rd annual symposium on foundations of computer science*, pages 514–523. IEEE, 2012.

[28] J.R. Lee, S.O. Gharan, and L. Trevisan. Multiway spectral partitioning and higher-order cheeger inequalities. *Journal of the ACM (JACM)*, 61(6):37, 2014.

[29] Y.-T. Lin, J. Malik, and H.-T. Wu. Wave-shape oscillatory model for biomedical time series with applications. *arXiv preprint arXiv:1907.00502*, 2019.

[30] G. C. Linderman, G. Mishne, Y. Kluger, and S. Steinerberger. Randomized near neighbor graphs, giant components, and applications in data science. *arXiv preprint arXiv:1711.04712*, 2017.

[31] J. L. Marcus, L. B. Hurley, D. S. Krakower, S. Alexeeff, M. J. Silverberg, and J. E. Volk. Use of electronic health record data and machine learning to identify candidates for hiv pre-exposure prophylaxis: a modelling study. *The Lancet HIV*, 6(10):e688–e695, 2019.

[32] P.G. Martinsson, V. Rokhlin, and M. Tygert. A randomized algorithm for the decomposition of matrices. *Applied and Computational Harmonic Analysis*, 30(1):47–68, 2011.

[33] D. A. McDonald. *Blood flow in arteries.* , Williams & Wilkins, 1974.

[34] G. F. Mitchell, S.-J. Hwang, R. S. Vasan, M. G. Larson, M. J. Pencina, N. M. Hamburg, J. A. Vita, D. Levy, and E. J. Benjamin. Arterial stiffness and cardiovascular events: the framingham heart study. *Circulation*, 121(4):505, 2010.

[35] M. F. O'Rourke and W. W. Nichols. Changes in wave reflection with advancing age in normal subjects. *Hypertension (Dallas, Tex.: 1979)*, 44(6):e10–1, 2004.

[36] V. Rokhlin, A. Szlam, and M. Tygert. A randomized algorithm for principal component analysis. *SIAM Journal on Matrix Analysis and Applications*, 31(3):1100–1124, 2009.

[37] C. F. Rowlands, D. Baralle, and J. M. Ellingford. Machine learning approaches for the prioritization of genomic variants impacting pre-mrna splicing. *Cells*, 8(12):1513, 2019.

[38] M. R. Rudnick, L. De Marchi, and J. S. Plotkin. Hemodynamic monitoring during liver transplantation: A state of the art review. *World journal of hepatology*, 7(10):1302, 2015.

[39] C. Shen and H.-T. Wu. Scalability and robustness of spectral embedding: landmark diffusion is all you need. *arXiv preprint arXiv:2001.00801*, 2020.

[40] B-F Shih, P-H Huang, H-P Yu, F-C Liu, C-C Lin, PC-H Chung, C-Y Chen, C-J Chang, and Y-F Tsai. Cardiac output assessed by the fourth-generation arterial waveform analysis system is unreliable in liver transplant recipients. In *Transplantation proceedings*, volume 48, pages 1170–1175. Elsevier, 2016.

[41] A. Singer and H.-T. Wu. Vector diffusion maps and the connection laplacian. *Communications on pure and applied mathematics*, 65(8):1067–1144, 2012.

[42] A. Singer and H-T Wu. Two-dimensional tomography from noisy projections taken at unknown random directions. *SIAM journal on imaging sciences*, 6(1):136–175, 2013.

[43] S. Steinerberger. A Filtering Technique for Markov Chains with Applications to Spectral Embedding. *Applied and Computational Harmonic Analysis*, 40:575–587, 2016.

[44] J.-L. Teboul, B. Saugel, M. Cecconi, et al. Less invasive hemodynamic monitoring in critically ill patients. *Intensive care medicine*, 42(9):1350–1359, 2016.

[45] Y-F Tsai, B-C Su, C-C Lin, F-C Liu, W-C Lee, and H-P Yu. Cardiac output derived from arterial pressure waveform analysis: validation of the third-generation software in patients undergoing orthotopic liver transplantation. In *Transplantation proceedings*, volume 44, pages 433–437. Elsevier, 2012.

[46] J. Wang, Z. Zhang, and H. Zha. Adaptive manifold learning. In *Advances in neural information processing systems*, pages 1473–1480, 2005.

bioRxiv preprint doi: https://doi.org/10.1101/2020.05.31.126649; this version posted June 1, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

[47] S.-C. Wang, H.-T. Wu, P.-H. Huang, C.-H. Chang, C.-K. Ting, and Y.-T. Lin. Novel imaging revealing inner dynamics for cardiovascular waveform analysis via unsupervised manifold learning. *Anesthesia & Analgesia*, 130(5):1244–1254, 2020.

[48] Y.-Y L. Wang, T.-L. Hsu, M.-Y. Jan, W.-K. Wang, et al. Theory and applications of the harmonic analysis of arterial pressure pulse waves. *Journal of Medical and Biological Engineering*, 30(3):125–131, 2010.

[49] C.K.I. Williams and M. Seeger. Using the nyström method to speed up kernel machines. In *Advances in neural information processing systems*, pages 682–688, 2001.

[50] R. Yuster and U. Zwick. Fast sparse matrix multiplication. *Lect. Notes Comput. Sci.*, 322:604–15, 2004.

Department of Mathematics, Duke University, Durham, NC, USA

Department of Anesthesiology, Taipei Veteran General Hospital, Taipei, Taiwan

Department of Mathematics, Duke University, Durham, NC, USA; Department of Statistical Science, Duke University, Durham, NC, USA; Mathematics Division, National Center for Theoretical Sciences, Taipei, Taiwan

E-mail address: hauwu@math.duke.edu