Semi-Supervised High Dimensional Clustering by Tight Wavelet Frames

Bin Dong*[a] and Ning Hao [b]

[a] Beijing International Center for Mathematical Research (BICMR), Peking University, Beijing, China; [b] Department of Mathematics, The University of Arizona, Tucson, AZ, USA

ABSTRACT

High-dimensional clustering arises frequently from many areas in natural sciences, technical disciplines and social medias. In this paper, we consider the problem of binary clustering of high-dimensional data, i.e. classification of a data set into 2 classes. We assume that the correct (or mostly correct) classification of a small portion of the given data is known. Based on such partial classification, we design optimization models that complete the clustering of the entire data set using the recently introduced tight wavelet frames on graphs.[1] Numerical experiments of the proposed models applied to some real data sets are conducted. In particular, the performance of the models on some very high-dimensional data sets are examined; and combinations of the models with some existing dimension reduction techniques are also considered.

Keywords: Tight wavelet frames, sparse representation on graphs, spectral graph theory, graph clustering, high-dimensional data analysis.

1. INTRODUCTION

Recent advances in information and computer technology contribute greatly to the exponential growth of data. In particular, the dimension that data reside in has increased tremendously which inevitably increases the needs of efficient techniques to properly handle, process and analyze high-dimensional data sets. High-dimensional clustering arises frequently from bioinformatics such as disease classifications using high throughput data like micorarrays or SNPs, machine learning such as document classification and image recognition, and many other areas in natural sciences, technical disciplines and social medias. The objective of clustering analysis is to discover groups, or clusters, of similar objects. The objects are usually represented as points in a multidimensional space. The data points are often regarded as vertices of a certain graph with properly defined weight functions on the edges depicting similarities among the vertices. Graphical framework is frequently used to exploit underlying similarities in the data.[2–6] In this paper, we consider the problem of binary clustering of high-dimensional data, i.e. classification of a data set into 2 classes. We assume that the correct (or mostly correct) classification of a small portion of the given data is known. Based on such partial classification, we design optimization models that complete the clustering of the entire data set using the recently introduced tight wavelet frames on graphs.[1]

In image processing and analysis, redundant systems such as wavelet frames have been implemented with excellent results in both classical and more challenging problems. Their applications in classical image restoration problems include image inpainting,[7] super-resolution,[8] deblurring,[9–12] demosaicing[13] and enhancement.[14] Wavelet frames are also applied to more challenging image restoration problems such as blind deblurring,[15,16] blind inpainting,[17] and denoising with unknown noise type.[18] Wavelet frame related algorithms have been developed to solve medical and biological image processing problems as well, e.g. X-ray computer tomography (CT) image reconstruction,[19,20] and protein molecule 3D reconstruction from electron microscopy images.[21] Furthermore, wavelet frames have been successfully used in video processing,[22] 4D CT image reconstruction,[23,24] image segmentation[25,26] and classifications.[27,28]

Another class of methods for image processing and analysis that is rather successful is the PDE based approach,[29–31] which includes variational and (nonlinear) PDE based methods. Recently, fundamental connections between wavelet frame based approach and PDE based approach were established.[32–34] In particular, connections

* E-mail: dongbin@math.pku.edu.cn; phone: (8610)62744091.

to the total variation model,[32] nonlinear evolution PDEs,[33] and the Mumford-Shah model[34] were established. The series of three papers[32–34] showed that wavelet frame transforms are discretization of differential operators in both variational and PDE frameworks, and such discretization is superior to some of the traditional finite difference schemes for image restoration. This new understanding essentially merged the two seemingly unrelated areas: wavelet frame based approach and PDE based approach. It also gave birth to many innovative and more effective image restoration models and algorithms.

Recently, there has been a growing interest in formulating graph clustering problems as variational problems, such as the non-local total variation[35–39] and Ginzburg-Landau functional based models.[40–42] Many of these models are motivated by variational and PDE-based methods used in image processing. Since wavelet frame based approach and PDE based approach have strong connections,[32–34] it is natural to ask whether wavelet frame based approach can be generalized to process and analyze data on graphs. This motivates the recent work[1] where the author introduced a new (constructive) characterization of tight wavelet frames in both continuum setting, i.e. on manifolds, and discrete setting, i.e. on graphs; discussed how fast tight wavelet frame transforms can be computed and how they can be effectively used to process graph data. In particular, a semi-supervised graph clustering model was proposed, preliminary numerical experiments on synthetic and real data sets were conducted, and comparisons with some state-of-the-art clustering methods were provided.

In this paper, we present two new clustering models based on the tight wavelet frame on graphs[1] that are specially designed to handle two different scenarios: (1) the known partial classification of the data set is accurate; and (2) the known partial classification of the data set is mostly accurate. Numerical experiments of the two models applied to some real data sets are conducted. In particular, the performance of the models on some very high-dimensional data sets are examined; and combinations of the models with some existing dimension reduction techniques are also considered.

The rest of the paper is organized as follows. A brief review of the tight wavelet frames on graphs and fast transformation algorithms[1] is given in Section 2. Two clustering models with their associated fast numerical algorithms are introduced in Section 3. Numerical experiments are presented in Section 4.

2. TIGHT WAVELET FRAMES ON GRAPHS

In this section, we briefly review the characterization and construction of tight wavelet frame systems on graphs, and the fast transformation algorithms proposed in one of the authors' earlier work.[1] Interested readers should consult the original article for full details.

2.1 Preliminaries

We denote a graph as $G := \{E, V, w\}$, where $V := \{v_k \in \mathbb{R}^s : k = 1, \ldots, K\} \subset \mathbb{R}^s$, $E \subset V \times V$ is an edge set, and $w : E \mapsto \mathbb{R}^+$ denotes a weight function. In this paper, we choose the following commonly used weight function

$$w(v_k, v_{k'}) := e^{-\|v_k - v_{k'}\|_2^2/\sigma}, \quad \sigma > 0.$$

Let $A := (a_{k,k'})$ be the adjacency matrix

$$a_{k,k'} := \begin{cases} w(v_k, v_{k'}) & \text{if } v_k \text{ and } v_{k'} \text{ are connected by an edge in } E \\ 0 & \text{otherwise,} \end{cases}$$

and $D := \mathrm{diag}\{d[1], d[2], \ldots, d[K]\}$ where $d[k]$ is the degree of node v_k defined by $d[k] := \sum_{k'} a_{k,k'}$. Let $\mathcal{L}$ be the (unnormalized) graph Laplacian, which takes the following form

$$\mathcal{L} := D - A.$$

Denote $\{(\lambda_k, u_k)\}_{k=0}^{K-1}$ the set of pairs of eigenvalues and eigenfunctions of $\mathcal{L}$. Assuming the graph is connected, then we have $0 = \lambda_0 < \lambda_1 \leq \lambda_2 \leq \cdots \leq \lambda_{K-1}$. The eigenfunctions form an orthonormal basis for all functions on the graph:

$$\langle u_k, u_{k'} \rangle := \sum_{n=1}^{K} u_k[n] u_{k'}[n] = \delta_{k,k'}.$$

Let $f_G : V \mapsto \mathbb{R}$ be a function on the graph G. Then its Fourier transform is given by

$$\widehat{f_G}[k] := \sum_{n=1}^{K} f_G[n] u_k[n].$$

2.2 Discrete Tight Wavelet Frame Transforms on Graph G

Given a set of masks $\{a_j : \ 0 \le j \le r\} \subset \ell_0(\mathbb{Z})$, we denote the Fourier series of a_j as

$$\widehat{a}_j(\xi) := \sum_{k \in \mathbb{Z}} a_j[k] e^{ik\xi}.$$

We define the *discrete L-level tight wavelet frame decomposition* as

$$\boldsymbol{W} f_G := \{W_{j,l} f_G : (j,l) \in \mathbb{B}\}$$

with

$$\mathbb{B} := \{(1,1),(2,1),\ldots,(r,1),(1,2),\ldots,(r,L)\} \cup \{(0,L)\} \tag{2.1}$$

and

$$\widehat{W_{j,l} f_G}[k] := \begin{cases} \widehat{a}_j^*(2^{-N}\lambda_k)\widehat{f_G}[k] & l = 1, \\ \widehat{a}_j^*(2^{-N+l-1}\lambda_k)\widehat{a}_0^*(2^{-N-l+2}\lambda_k)\cdots\widehat{a}_0^*(2^{-N}\lambda_k)\widehat{f_G}[k] & 2 \le l \le L. \end{cases} \tag{2.2}$$

The dilation scale N is chosen as the smallest integer such that

$$\lambda_{\max} := \lambda_{K-1} \le 2^N \pi.$$

Note that the scale N is selected such that $2^{-N}\lambda_k \in [0,\pi]$ for $0 \le k \le K-1$. The index j denotes the band of the transform with $j = 0$ the low frequency component and $1 \le j \le r$ the high frequency components. The index l denotes the level of the transform.

Given a graph function f_G, let $\boldsymbol{\alpha} := \boldsymbol{W} f_G := \{\alpha_{j,l} : \ (j,l) \in \mathbb{B}\}$, with $\alpha_{j,l} := W_{j,l} f_G$, be its tight wavelet frame coefficients. We denote the *discrete L-level tight wavelet frame reconstruction* as $\boldsymbol{W}^\top \boldsymbol{\alpha}$, which is defined by the following iterative procedure in frequency domain

$$\widehat{\alpha}_{0,l-1}[k] = \sum_{j=0}^{r} \widehat{a}_j(2^{-N+l-1}\lambda_k)\widehat{\alpha}_{j,l}[k] \qquad \text{for } l = L, L-1, \ldots, 1, \tag{2.3}$$

where $\alpha_{0,0} := \boldsymbol{W}^\top \boldsymbol{\alpha}$ is the reconstructed graph data from $\boldsymbol{\alpha}$. Note that $\boldsymbol{W}$ is obviously a linear transformation, and it is easy to verify that the linear transformation $\boldsymbol{W}^\top$ defined by (2.3) is indeed the adjoint of $\boldsymbol{W}$ satisfying $\langle \boldsymbol{W} f_G, \boldsymbol{\alpha} \rangle = \langle f_G, \boldsymbol{W}^\top \boldsymbol{\alpha} \rangle$ for all f_G and $\boldsymbol{\alpha}$. The following theorem states that $\{a_j : 0 \le j \le r\}$ generates a discrete tight wavelet frame on graphs.[1]

Theorem 2.1. *Given a set of masks $\{a_j : 0 \le j \le r\} \subset \ell_0(\mathbb{Z})$, suppose the following condition is satisfied:*

$$\sum_{j=0}^{r} |\widehat{a}_j(\xi)|^2 = 1. \tag{2.4}$$

Then, the discrete tight wavelet frame transforms $\boldsymbol{W}$ and $\boldsymbol{W}^\top$ defined on $G = \{E, V, w\}$ by (2.2) *and* (2.3) *satisfy*

$$\boldsymbol{W}^\top \boldsymbol{W} f_G = f_G, \quad \textit{for all } f_G : V \mapsto \mathbb{R}.$$

There are many sets of masks satisfying condition (2.4). For example, all masks constructed from the unitary extension principle[43] satisfy (2.4), and hence generate tight frame systems on graphs by Theorem 2.1. Here, we present two examples, i.e "Haar" and "Linear".[1]

Examples.

1. **Haar.**
$$\widehat{a}_0(\xi) = \cos(\xi/2) \quad \text{and} \quad \widehat{a}_1(\xi) = \sin(\xi/2).$$

2. **Linear.**
$$\widehat{a}_0(\xi) = \cos^2(\xi/2), \quad \widehat{a}_1(\xi) = \frac{1}{\sqrt{2}}\sin(\xi) \quad \text{and} \quad \widehat{a}_2(\xi) = \sin^2(\xi/2).$$

2.3 Fast Tight Wavelet Frame Transform on Graphs (WFTG)

The discrete tight wavelet frame transforms given by (2.2) and (2.3) require the full set of eigenvectors and eigenvalues of the graph Laplacian, which is computationally expensive to obtain for large graphs. A solution to such computation challenge is to use polynomial approximation of the masks, such as the Chebyshev polynomials, so that eigenvalue decomposition of the graph Laplacian is not needed.[1,44] Note that, the masks $\widehat{a}_j(\xi)$ that we use, as well as those constructed in the literature from the unitary extension principle,[43] are trigonometric polynomials. Therefore, $\widehat{a}_j(\xi)$ can be accurately approximated by *low-degree* Chebyshev polynomials[45] which significantly reduces the computation cost of the decomposition and reconstruction algorithms. In our simulations, we choose $n = 8$ which is sufficient for the applications we consider. Note that, if a higher order B-spline tight wavelet frame system is used, Chebyshev polynomials with a higher degree may be needed to achieve a given approximation accuracy. However, just as in image processing, only the lower order systems are mostly used because they offer a good balance between quality and computation efficiency.

Given mask $\widehat{a}_j(\xi)(\xi)$ with $\xi \in [0, \pi]$, we denote its Chebyshev polynomial approximation as

$$\widehat{a}_j(\xi) \approx \mathcal{T}_j^n(\xi) = \frac{1}{2}c_{j,0} + \sum_{k=1}^{n-1} c_{j,k}T_k, \qquad c_{j,k} = \frac{2}{\pi}\int_0^\pi \cos(k\theta)\widehat{a}_j\left(\frac{\pi}{2}(\cos(\theta)+1)\right)d\theta.$$

We denote the Chebyshev polynomial approximation of $\widehat{a}_j^*$ as $\mathcal{T}_j^{n*}$. Since the masks $\{\widehat{a}_j\}$ we will be using are real-valued, we have $\mathcal{T}_j^n = \mathcal{T}_j^{n*}$. If we substitute the approximation $\widehat{a}_j(\xi) \approx \mathcal{T}_j^n(\xi)$ in (2.2), and use the fact that $\mathcal{T}_j^n$ are polynomials and Fourier transform on G is unitary, we obtain the WFTG:

Fast Tight Wavelet Frame Transform on Graphs (WFTG)[1]

$\boldsymbol{W} f_G = \{W_{j,l} f_G : (j,l) \in \mathbb{B}\}$ where

$$W_{j,l} f_G := \begin{cases} \mathcal{T}_j^{n*}(2^{-N}\mathcal{L})f_G & l = 1, \\ \mathcal{T}_j^{n*}(2^{-N+l-1}\mathcal{L})\mathcal{T}_0^{n*}(2^{-N-l+2}\mathcal{L})\cdots\mathcal{T}_0^{n*}(2^{-N}\mathcal{L})f_G & l \geq 2. \end{cases} \tag{2.5}$$

Reconstruction transform $\boldsymbol{W}^\top$ can be defined similarly and we have $\boldsymbol{W}^\top\boldsymbol{W} \approx \boldsymbol{I}$. Note that, for computational efficiency, the operations $\mathcal{T}_j^*(s\mathcal{L})f_G$ for decomposition and $\mathcal{T}_j(s\mathcal{L})f_G$ for reconstruction are computed via the iterative definition of the Chebyshev polynomials. Therefore, in WFTG, only matrix-vector multiplications are involved.

3. SEMI-SUPERVISED GRAPH CLUSTERING MODELS AND ALGORITHMS

We consider semi-supervised clustering, where the labeling, i.e. clusters, of a small subset of the data is provided in advance. We will introduce two optimization models and their associated fast numerical algorithms based on the WFTG for two different clustering scenarios. The proposed models and algorithms are motivated by the earlier work on wavelet frame based image segmentation[25] and surface reconstruction,[46] as well as the variational frameworks for image segmentation[47,48] and graph clustering.[40,42]

Given graph $G = \{E, V, w\}$, let $|V| = K$ and $\Gamma \subset \Omega := \{1, 2, \ldots, K\}$ be the set of labels where, for $i \in \Gamma$, we know which cluster v_i belongs to. Let $\Gamma := \Gamma_0 \cup \Gamma_1$ where Γ_0 is the index set for cluster 1 and Γ_1 for cluster 2. Define the cluster-indicator function $f : \Gamma \mapsto \mathbb{R}$ as

$$f[i] := \begin{cases} 0 & \text{for } i \in \Gamma_0, \\ 1 & \text{for } i \in \Gamma_1. \end{cases} \tag{3.1}$$

Let $\mathcal{R}_\Gamma$ denote the restriction operator on set Γ, i.e. $\mathcal{R}_\Gamma u : \Gamma \mapsto \mathbb{R}$ for $u : \Omega \mapsto \mathbb{R}$ with $\mathcal{R}_\Gamma u = u$ on Γ. Our objective is to recover a function $u : \Omega \mapsto [0, 1]$ with $\mathcal{R}_\Gamma u \approx f$, such that the two sets $\{i \mid u[i] \geq \beta\}$ and $\{i \mid u[i] < \beta\}$ provide a good clustering of the given graph for some $\beta \in (0, 1)$. In our experiments, we select $\beta = 0.5$.

In the recent work,[1] the following model was proposed to find u

$$\min_{u\in[0,1]} \quad \|\boldsymbol{\nu} \cdot \boldsymbol{W} u\|_{1,G} + \frac{1}{2}\|u_{|\Gamma} - f\|_{2,G}^2, \tag{3.2}$$

where the first term impose a regularization of the level sets of u making sure u has regular level sets and is close to a binary function on G, while the second is the fidelity term making sure that $u_{|\Gamma} \approx f$. Positive numerical results on both simulated and real data sets were reported and comparison with state-of-the-art methods were discussed.

In practice, when data f is collected, it is reasonable to assume that we are in one of the following two situations: (1) f is perfectly accurate; (2) f contains some miss-labelling while the percentage of miss-labelling is small. Then, model 3.2 is not ideal to model either of the two cases, although decent clustering results may still be achieved.[1] In this section, we will propose an optimization model and its associated fast algorithm for each of the aforementioned scenarios.

3.1 Exact Model

Assume that the labelling-indicator function f is error-free. We propose the following model which will be referred to as the "exact model":

$$\begin{array}{cl} \min_u & \|\boldsymbol{\nu} \cdot \boldsymbol{W} u\|_{1,G} \\ \text{s.t.} & \mathcal{R}_\Gamma u = f, \end{array} \tag{3.3}$$

where

$$\nu_{j,l} = \begin{cases} 4^{-l+1}\nu & \text{for } j \neq 0, 1 \leq l \leq L, \\ 0 & \text{for } j = 0, l = L, \end{cases} \tag{3.4}$$

where $\nu > 0$ is a scalar tuning parameter.

To solve problem (3.3), we first convert it to its equivalent form:

$$\begin{array}{cl} \min_{u,d} & \|\boldsymbol{\nu} \cdot d\|_{1,G} \\ \text{s.t.} & \mathcal{R}_\Gamma u = f, \ \boldsymbol{W} u = d. \end{array} \tag{3.5}$$

Problem (3.5) can be solved using the augmented Lagrangian method:[49–52]

$$\begin{cases} (u^{k+1}, d^{k+1}) = \arg\min\limits_{u,d} \ \|\boldsymbol{\nu} \cdot d\|_{1,G} + \frac{\mu_1}{2}\|\mathcal{R}_\Gamma u - f + b_1^k\|_2^2 + \frac{\mu_2}{2}\|\boldsymbol{W} u - d^k + b_2^k\|_2^2, \\ b_1^{k+1} = b_1^k + \mathcal{R}_\Gamma u^{k+1} - f, \\ b_2^{k+1} = b_2^k + \boldsymbol{W} u^{k+1} - d^{k+1}. \end{cases}$$

If we solve the above subproblem alternatively using only one step for each of the variable u and d, we have the following algorithm which is known as the alternating direction method of multipliers (ADMM)[52–54] and the split Bregman algorithm:[9, 55]

$$\begin{cases} u^{k+1} = \arg\min_u \ \frac{\mu_1}{2}\|\mathcal{R}_\Gamma u - f + b_1^k\|_2^2 + \frac{\mu_2}{2}\|\boldsymbol{W} u - d^k + b_2^k\|_2^2, \\ d^{k+1} = \arg\min_d \ \|\boldsymbol{\nu} \cdot d\|_{1,G} + \frac{\mu_2}{2}\|d - (\boldsymbol{W} u^{k+1} + b_2^k)\|_2^2, \\ b_1^{k+1} = b_1^k + \mathcal{R}_\Gamma u^{k+1} - f, \\ b_2^{k+1} = b_2^k + \boldsymbol{W} u^{k+1} - d^{k+1}. \end{cases} \tag{3.6}$$

3.2 Robust Model

If there is miss-labelling when data is collected, then the labelling-indicator function f contains errors. Thus, we cannot enforce that $\mathcal{R}_\Gamma = f$. Assuming that the percentage of the miss-labelling is relatively small, we propose the following "robust model":

$$\min_u \ \|\boldsymbol{\nu} \cdot \boldsymbol{W} u\|_{1,G} + \|\mathcal{R}_\Gamma u - f\|_{1,G}, \tag{3.7}$$

where $\boldsymbol{\nu}$ is given by (3.4). Here, the use of the ℓ_1-norm in the second term of (3.7) is to properly handle miss-labelling in f. The use of the ℓ_1-norm in data fidelity terms is common in variational methods of image restoration to handle sparse impulse noise or other sparse corruptions.[17, 56–65]

To solve problem (3.7), we convert it to its equivalent form as

$$\begin{array}{cl} \min_{u,d_1,d_2} & \|\boldsymbol{\nu} \cdot d_1\|_{1,G} + \|d_2\|_{1,G} \\ \text{s.t.} & \boldsymbol{W} u = d_1, \ d_2 = \mathcal{R}_\Gamma u - f. \end{array} \tag{3.8}$$

Same as (3.5), problem (3.8) can be solved using the augmented Lagranging method:

$$\begin{cases} (u^{k+1}, d_1^{k+1}, d_2^{k+1}) = \arg\min\limits_{u,d} \ \|\boldsymbol{\nu} \cdot d_1\|_{1,G} + \|d_2\|_{1,G} + \frac{\mu_1}{2}\|\boldsymbol{W} u - d_1^k + b_1^k\|_2^2 + \frac{\mu_2}{2}\|\mathcal{R}_\Gamma u - f - d_2^k + b_2^k\|_2^2, \\ b_1^{k+1} = b_1^k + \boldsymbol{W} u^{k+1} - d_1^{k+1}, \\ b_2^{k+1} = b_2^k + \mathcal{R}_\Gamma u^{k+1} - f - d_2^{k+1}, \end{cases}$$

which leads to the the ADMM algorithm:

$$\begin{cases} u^{k+1} = \arg\min_u \ \frac{\mu_1}{2}\|\boldsymbol{W} u - d_1^k + b_1^k\|_2^2 + \frac{\mu_2}{2}\|\mathcal{R}_\Gamma u - f - d_2^k + b_2^k\|_2^2, \\ d_1^{k+1} = \arg\min_{d_1} \ \|\boldsymbol{\nu} \cdot d_1\|_{1,G} + \frac{\mu_1}{2}\|d - (\boldsymbol{W} u^{k+1} + b_1^k)\|_2^2, \\ d_2^{k+1} = \arg\min_{d_2} \ \|d_2\|_{1,G} + \frac{\mu_2}{2}\|d_2 - (\mathcal{R}_\Gamma u - f + b_2^k)\|_2^2, \\ b_1^{k+1} = b_1^k + \boldsymbol{W} u^{k+1} - d_1^{k+1}, \\ b_2^{k+1} = b_2^k + \mathcal{R}_\Gamma u^{k+1} - f - d_2^{k+1}. \end{cases} \tag{3.9}$$

Convergence of the algorithm (3.6) and (3.9) have been discussed in the literature where (rate of) convergence of these algorithms is also available under suitable assumptions of the objective function.[9, 66–70]

4. NUMERICAL SIMULATIONS

In this section, we apply the exact and robust models and their associated algorithms to some real data sets. In particular, we shall observe how the proposed methods perform on high dimensional data sets, especially when the dimensions of the data are noticeably higher than the cardinalities of the data sets. Under such situation, we also observe how the combination of dimension reduction techniques with the proposed method perform on these high dimensional data sets. Throughout this section, we use "Haar" tight wavelet frame system and the level of decomposition L is chosen to be 1.

4.1 Low-Dimensional Clustering

We test the exact and robust model (3.3) and (3.7) on two real data sets. These two data sets are considered as "low-dimensional" since the cardinalities of the data sets are relatively large comparing to the dimensions of the data. We will compare the performances of the two models on the two data sets under two scenarios: with and without miss-labelling.

The first real data set we use is the MNIST data set,[71] which is available at http:// yann.lecun.com/ exdb/ mnist/. It contains 70000 28×28 images of handwritten digits from 0-9. Since our clustering method is binary (2-classes clustering), we choose the subset with digits 4 and 9 to classify since these digits are harder to distinguish. This created a data set of 13782 image vectors, which is either 4 or 9. In our numerical simulations, we randomly draw 500 image vectors from the data set (about 3.62%) as training set and use them to create the known label set Γ and graph data f of (3.2). We repeat this process 100 times and report the average classification errors (in %) and computation time.

The second data set is the banknote authentication data set from the UCI machine learning repository.[72] It is a data set of 1372 features extracted from images (400 × 400 pixels) of genuine and forged banknotes using wavelet transforms. The goal is to separate the banknotes into being either genuine or forged. In our numerical simulations, we randomly draw 50 data vectors from the entire data set (about 3.64%) as training set and use them to create the known label set Γ and graph data f of (3.2). We repeat this process 100 times and report the average classification errors (in %) and computation time.

Numerical results of the exact and robust model on the two data sets with and without miss-labelling are presented in Table 1 and 2. As we can see, when there is not miss-labelling, the robust model performs comparably to the exact model. When there is miss-labelling, robust model outperforms the exact model for both cases. Also, recall that the clustering errors of model (3.2) were 2.76% and 1.64% for MNIST and banknote data sets without miss-labelling (under the exact experimental settings),[1] while the errors of the exact model without miss-labelling are 2.65% and 1.45%. This shows that considering the exact model (3.3) has some advantage over the earlier proposed model (3.2) when there is no miss-labelling.

Table 1. Clustering errors (%) of MNIST data set. Average computation time (in seconds) for both exact and robust model is included in parentheses.

Errors (Time)	Exact Model	Robust Model
w/o miss-labelling	2.65% (10.9 sec.)	2.66% (9.6 sec.)
20% miss-labelling	7.57% (11.3 sec.)	7.29% (11.5 sec.)

Table 2. Clustering errors (%) of banknote data set. Average computation time (in seconds) for both exact and robust model is included in parentheses.

Errors (Time)	Exact Model	Robust Model
w/o miss-labelling	1.45% (4.2 sec.)	1.45% (3.7 sec.)
10% miss-labelling	7.54% (3.9 sec.)	5.89% (4.2 sec.)

4.2 High-Dimensional Clustering

Given a data set containing two classes $X^{(1)} = \{x_j^{(1)} \in \mathbb{R}^s : 1 \leq j \leq n_1\}$ and $X^{(2)} = \{x_j^{(2)} \in \mathbb{R}^s : 1 \leq j \leq n_2\}$, we consider the situation when $n_1, n_2 \ll s$ and will refer to the cluster of such data set as high-dimensional clustering. High-dimensional clustering is a challenging problem. Many classical methods such as the traditional linear discriminant analysis (LDA) methods do not perform well. The reason is that, with limited number of observations n_1 and n_2, it is impossible to estimate too many parameters simultaneously and accurately. To overcome the lack of observations, sparsity based LDA methods were proposed, such as the nearest shrunken centroids (NSC),[73] IR,[74] FAIR,[75] ROAD,[76] LPD[77] and the rotate-and-solve (RS) strategy.[78] The first part of this subsection is to apply the proposed exact model (3.3) to benchmark data sets, i.e. gene expression data sets, that is commonly used for high-dimensional clustering, and compare its performance with state-of-the-art sparsity based LDA methods that are known to be effective on the gene expression data sets.

The main challenge of the gene expression data sets is that the dimension of the data s is far larger than the number of observations n_1 and n_2. When the dimension s of the given data set is moderately larger than n_1 and n_2, we can apply dimension reduction techniques to reduce the original dimension of the data to a dimension that is smaller than n_1, n_2, and then apply the exact model (3.3). The second part of this subsection is to experiment on the performances of different dimension reduction techniques combining with our clustering method. Throughout this section, we assume that all labels given are accurate, i.e. no miss-labelling.

4.2.1 Gene Expression Data

We consider the two popular gene expression data set: Leukemia[79] and lung cancer.[80] The two data sets come with separate training and testing sets of data vectors. The Leukemia data set contains $s = 7129$ genes with $n_1 = 27$ acute lymphoblastic leukemia (ALL) and $n_2 = 11$ acute myeloid leukemia (AML) vectors in the training set. The testing set includes 20 ALL and 14 AML vectors. The Lung Cancer data set contains $s = 12533$ genes with $n_1 = 16$ adenocarcinoma (ADCA) and $n_2 = 16$ mesothelioma training vectors. The testing set has 134 ADCA and 15 mesothelioma vectors.

For Leukemia data set, we put all the 47 (27 training + 20 testing data) ALL vectors and 25 (11 training + 14 testing data) AML vectors together and randomly select 23 ALL and 12 AML as training set (about 50% of data) and the rest as testing set. We repeat the experiments 20 times. We conduct a similar experiment on Lung cancer data by randomly select 75 ADCA and 15 mesothelioma data vector as training set (about 50% of data) and the rest as testing set, and repeat 20 times.

The clustering results are presented in Table 3. As we can see that the proposed exact model (3.3) is comparable to the state-of-the-art methods that are specially designed for high-dimensional clustering. For Leukemia, the exact model is slightly worse than IR and RS-ROAD; while for Lung, the exact model is better than the other methods. To see the different between the two data sets, we use t-SNE method[81] to visualize the data sets in $\mathbb{R}^3$, which is given in Figure 1. Note that, all LDA methods are based on the assumption that both of the clusters $X^{(1)}$ and $X^{(2)}$ are sampled from two s-dimensional normal distributions with the same covariance matrix. As we can see from Figure 1 that such assumption is somewhat valid on Leukemia data set and a linear classifier is sufficient; while this assumption is clearly invalid on Lung data set. From the plot of Lung data set, it seems that a nonlinear classifier is more suitable than a linear classifier, which is consistent with our numerical results in Table 3.

Table 3. Clustering errors and standard deviations for gene expression data sets.

Errors % (std %)	IR[74]	NSC[73]	ROAD[76]	RS-ROAD[78]	Exact Model
Leukemia	4.2708 (2.9998)	8.5135 (8.4232)	6.3514 (5.9650)	4.4595 (3.0721)	5.5714 (4.1945)
Lung Cancer	3.4669 (1.4381)	10.4396 (7.2675)	1.3736 (1.0621)	0.9341 (0.8931)	0.5889 (0.6004)

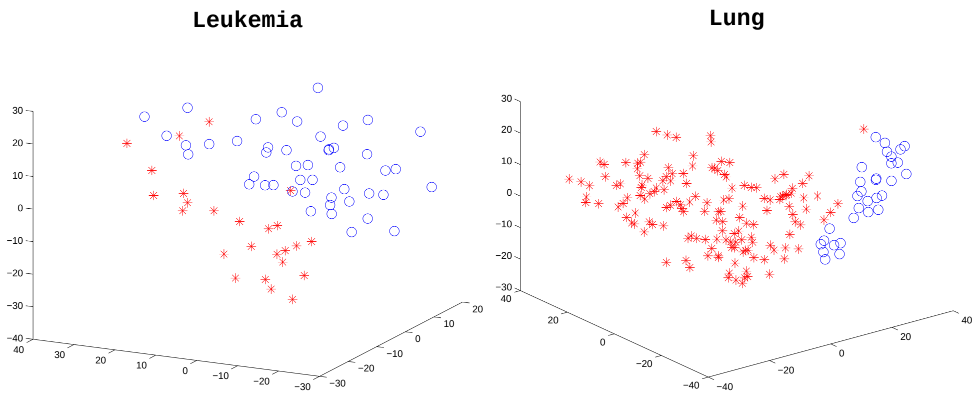

Figure 1. 3D plots of data set Leukemia and Lung.

4.2.2 Combining with Dimension Reduction Techniques

When the dimension of a given data set s is moderately bigger than numbers of observations n_1 and n_2, we may apply dimension reduction techniques before applying the proposed clustering models. Dimension reduction techniques are often needed to capture important features, discard unwanted noise, and downsize the data set without information loss before a sophisticated and computationally expensive analysis. Here, we experiment on a few existing dimension reduction methods in combination of the exact model (3.3). The dimension reduction methods we select for our experiments are: PCA,[82] classical MDS,[83] Laplacian eigenmap (LapEigen)[84] and the RS method.[78] The dimension reduction methods PCA, MDS and LapEigen are implemented in MATLAB using the dimension reduction toolbox written by Laurens van der Maaten (http://lvdmaaten.github.io/drtoolbox/#usage).[81,85]

We select MNIST data set (numbers 4 and 9) for our tests. To create the situations where the dimension s is greater than number of observations n_1, n_2, we randomly select subsets of the data set with different cardinalities (i.e. $n_1 + n_2$). We use 50% of the data as training and the rest as testing. For each given sub-sampled data set, we use the exact same experimental setting as in Section 4.1. For each run, we first apply one of the aforementioned dimension reduction methods and then the clustering model (3.3). All parameters are selected manually for optimal clustering results. Numerical results are presented in Table 4. As we can see that, when the total number of observations decreases, the clustering quality of the exact model degrades, which is typical for most clustering methods. Also, it seems from Table 4 that the best combo of dimension reduction techniques with our clustering model (balancing between quality and computation efficiency) is simple clustering methods such as PCA and MDS. The RS method also works well when number of observations are relatively high. One may use more sophisticated dimension reduction methods. However, when number of observations is small, a sophisticated dimension reduction method may not improve the final clustering results such as LapEign. However, to be more conclusive, experiments on more data sets using more dimension reduction methods need to be done in our future studies.

Table 4. Clustering errors for MNIST data set (number 4 and 6) with varied numbers of observations n_1 and n_2. The dimension of the data is $s = 783$. The best result among all the methods for each case is made in bold.

Errors %	No DR	PCA	MDS	LapEigen	RS
$n_1 + n_2 = 276$	10.29	**7.88**	8.69	17.65	10.23
$n_1 + n_2 = 413$	9.38	7.04	**6.96**	13.13	8.62
$n_1 + n_2 = 689$	5.72	4.30	**4.14**	7.03	4.56
$n_1 + n_2 = 965$	4.93	4.43	4.54	5.73	**4.39**
$n_1 + n_2 = 1378$	4.11	3.53	**3.39**	5.72	3.65

REFERENCES

[1] Dong, B., "Sparse representation on graphs by tight wavelet frames and applications," *accepted by Applied and Computational Harmonic Analysis* (2015).

[2] Zhou, D. and Schölkopf, B., "A regularization framework for learning from graph data," in [*ICML workshop on statistical relational learning and Its connections to other fields*], **15**, 67–68 (2004).

[3] Zhou, D., Bousquet, O., Lal, T. N., Weston, J., and Schölkopf, B., "Learning with local and global consistency," *Advances in neural information processing systems* **16**(16), 321–328 (2004).

[4] Belkin, M., Niyogi, P., and Sindhwani, V., "Manifold regularization: A geometric framework for learning from labeled and unlabeled examples," *The Journal of Machine Learning Research* **7**, 2399–2434 (2006).

[5] Chapelle, O., Schölkopf, B., Zien, A., et al., "Semi-supervised learning," (2006).

[6] Wang, J., Jebara, T., and Chang, S.-F., "Graph transduction via alternating minimization," in [*Proceedings of the 25th international conference on Machine learning*], 1144–1151, ACM (2008).

[7] Cai, J., Chan, R., and Shen, Z., "A framelet-based image inpainting algorithm," *Applied and Computational Harmonic Analysis* **24**(2), 131–149 (2008).

[8] Chan, R., Chan, T., Shen, L., and Shen, Z., "Wavelet algorithms for high-resolution image reconstruction," *SIAM Journal on Scientific Computing* **24**(4), 1408–1432 (2003).

[9] Cai, J., Osher, S., and Shen, Z., "Split Bregman methods and frame based image restoration," *Multiscale Modeling and Simulation: A SIAM Interdisciplinary Journal* **8**(2), 337–369 (2009).

[10] Cai, J., Osher, S., and Shen, Z., "Linearized Bregman iterations for frame-based image deblurring," *SIAM J. Imaging Sci* **2**(1), 226–252 (2009).

[11] Zhang, Y., Dong, B., and Lu, Z., "ℓ_0 minimization of wavelet frame based image restoration," *Mathematics of Computation* **82**, 995–1015 (2013).

[12] Dong, B. and Zhang, Y., "An efficient algorithm for ℓ_0 minimization in wavelet frame based image restoration," *Journal of Scientific Computing* **54 (2-3)**, 350–368 (2013).

[13] Liang, J., Li, J., Shen, Z., and Zhang, X., "Wavelet frame based color image demosaicing," *Inverse Problems and Imaging* **7**(3), 777–794 (2013).

[14] Hou, L., Ji, H., and Shen, Z., "Recovering over-/underexposed regions in photographs.," *SIAM J. Imaging Sciences* **6**(4), 2213–2235 (2013).

[15] Cai, J., Ji, H., Liu, C., and Shen, Z., "Blind motion deblurring using multiple images," *Journal of Computational Physics* **228**(14), 5057–5071 (2009).

[16] Cai, J., Ji, H., Liu, C., and Shen, Z., "Blind motion deblurring from a single image using sparse approximation," in [*Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*], 104–111, IEEE (2009).

[17] Dong, B., Ji, H., Li, J., Shen, Z., and Xu, Y., "Wavelet frame based blind image inpainting," *accepted by Applied and Computational Harmonic Analysis* **32**(2), 268–279 (2011).

[18] Gong, Z., Shen, Z., and Toh, K.-C., "Image restoration with mixed or unknown noises," *Multiscale Modeling & Simulation* **12**(2), 458–487 (2014).

[19] Jia, X., Dong, B., Lou, Y., and Jiang, S., "GPU-based iterative cone-beam CT reconstruction using tight frame regularization," *Physics in Medicine and Biology* **56**, 3787–3807 (2011).

[20] Dong, B., Li, J., and Shen, Z., "X-ray CT image reconstruction via wavelet frame based regularization and Radon domain inpainting," *Journal of Scientific Computing* **54 (2-3)**, 333–349 (2013).

[21] Li, M., Fan, Z., Ji, H., and Shen, Z., "Wavelet frame based algorithm for 3d reconstruction in electron microscopy," *SIAM Journal on Scientific Computing* **36**(1), B45–B69 (2014).

[22] Ji, H., Huang, S., Shen, Z., and Xu, Y., "Robust video restoration by joint sparse and low rank matrix approximation," *SIAM Journal on Imaging Sciences* **4**, 1122 (2011).

[23] Gao, H., Cai, J., Shen, Z., and Zhao, H., "Robust principal component analysis-based four-dimensional computed tomography," *Physics in Medicine and Biology* **56**, 3181 (2011).

[24] Cai, J., Jia, X., Gao, H., Jiang, S., Shen, Z., and Zhao, H., "Cine cone beam ct reconstruction using low-rank matrix factorization: Algorithm and a proof-of-principle study.," *IEEE transactions on medical imaging* **33**(8), 1581–1591 (2014).

[25] Dong, B., Chien, A., and Shen, Z., "Frame based segmentation for medical images," *Communications in Mathematical Sciences* **9(2)**, 551–559 (2010).

[26] Tai, C., Zhang, X., and Shen, Z., "Wavelet frame based multiphase image segmentation," *SIAM Journal on Imaging Sciences* **6**(4), 2521–2546 (2013).

[27] Wendt, H., Abry, P., Jaffard, S., Ji, H., and Shen, Z., "Wavelet leader multifractal analysis for texture classification," in [*Image Processing (ICIP), 2009 16th IEEE International Conference on*], 3829–3832, IEEE (2009).

[28] Bao, C., Ji, H., Quan, Y., and Shen, Z., "ℓ_0 norm based dictionary learning by proximal methods with global convergence," in [*Computer Vision and Pattern Recognition (CVPR), 2014 IEEE Conference on*], 3858–3865, IEEE (2014).

[29] Sapiro, G., [*Geometric partial differential equations and image analysis*], Cambridge University Press (2001).

[30] Osher, S. and Fedkiw, R., [*Level set methods and dynamic implicit surfaces*], Springer (2003).

[31] Chan, T. and Shen, J., [*Image processing and analysis: variational, PDE, wavelet, and stochastic methods*], Society for Industrial Mathematics (2005).

[32] Cai, J., Dong, B., Osher, S., and Shen, Z., "Image restorations: total variation, wavelet frames and beyond," *Journal of American Mathematical Society* **25(4)**, 1033–1089 (2012).

[33] Dong, B., Jiang, Q., and Shen, Z., "Image restoration: Wavelet frame shrinkage, nonlinear evolution pdes, and beyond," *UCLA CAM Report* **13-78** (2013).

[34] Cai, J., Dong, B., and Shen, Z., "Image restorations: a wavelet frame based model for piecewise smooth functions and beyond," *Preprint* (2014).

[35] Gilboa, G. and Osher, S., "Nonlocal operators with applications to image processing," *Multiscale Model Sim* **7**, 1005–1028 (Jan 2008).

[36] Gilboa, G. and Osher, S., "Nonlocal linear image regularization and supervised segmentation," *Multiscale Modeling and Simulation* **6**(2), 595–630 (2008).

[37] Bresson, X. and Chan, T. F., "Non-local unsupervised variational image segmentation models," *UCLA CAM Report* **8-67** (2008).

[38] Houhou, N., Bresson, X., Szlam, A., Chan, T. F., and Thiran, J.-P., "Semi-supervised segmentation based on non-local continuous min-cut," in [*Scale Space and Variational Methods in Computer Vision*], 112–123, Springer (2009).

[39] Zhang, X., Burger, M., Bresson, X., and Osher, S., "Bregmanized nonlocal regularization for deconvolution and sparse reconstruction," *SIAM Journal on Imaging Sciences* **3**, 253–276 (2010).

[40] Bertozzi, A. L. and Flenner, A., "Diffuse interface models on graphs for classification of high dimensional data," *Multiscale Modeling & Simulation* **10**(3), 1090–1118 (2012).

[41] Merkurjev, E., Kostic, T., and Bertozzi, A. L., "An mbo scheme on graphs for classification and image processing," *SIAM Journal on Imaging Sciences* **6**(4), 1903–1930 (2013).

[42] Merkurjev, E., Bae, E., Bertozzi, A., and Tai, X. C., "Global binary optimization on graphs for classification of high dimensional data," *UCLA CAM Report* **14-72** (2014).

[43] Ron, A. and Shen, Z., "Affine systems in $L_2(\mathbb{R}^d)$: The analysis of the analysis operator," *Journal of Functional Analysis* **148**(2), 408–447 (1997).

[44] Hammond, D. K., Vandergheynst, P., and Gribonval, R., "Wavelets on graphs via spectral graph theory," *Applied and Computational Harmonic Analysis* **30**(2), 129–150 (2011).

[45] Mason, J. C. and Handscomb, D. C., [*Chebyshev polynomials*], CHAPMAN & HALL/CRC Press (2010).

[46] Dong, B. and Shen, Z., "Frame based surface reconstruction from unorganized points," *Journal of Computational Physics* **230**, 8247–8255 (2011).

[47] Chan, F. and Vese, L., "Active contours without edges," *IEEE Transactions on image processing* **10**(2), 266–277 (2001).

[48] Chan, F. and Vese, L., "An active contour model without edges," *Scale-Space Theories in Computer Vision* **1682**, 141–151 (1999).

[49] Rockafellar, R. T., "Augmented lagrange multiplier functions and duality in nonconvex programming," *SIAM Journal on Control* **12**(2), 268–285 (1974).

[50] Rockafellar, R., "Augmented lagrangians and applications of the proximal point algorithm in convex programming," *Mathematics of Operations Research* , 97–116 (1976).

[51] Bertsekas, D., "On penalty and multiplier methods for constrained minimization," *SIAM Journal on Control and Optimization* **14**, 216 (1976).

[52] Glowinski, R. and Le Tallec, P., [*Augmented Lagrangian and operator-splitting methods in nonlinear mechanics*], Society for Industrial and Applied Mathematics (1989).

[53] Gabay, D. and Mercier, B., "A dual algorithm for the solution of nonlinear variational problems via finite element approximation," *Computers & Mathematics with Applications* **2**(1), 17–40 (1976).

[54] Glowinski, R. and Oden, J., "Numerical methods for nonlinear variational problems," *Journal of Applied Mechanics* **52**, 739 (1985).

[55] Goldstein, T. and Osher, S., "The split Bregman algorithm for L1 regularized problems," *SIAM Journal on Imaging Sciences* **2**(2), 323–343 (2009).

[56] Alliney, S., "Digital filters as absolute norm regularizers," *IEEE Transactions on Signal Processing* **40**(6), 1548–1562 (1992).

[57] Alliney, S., "Recursive median filters of increasing order: a variational approach," *IEEE Transactions on Signal Processing* **44**(6), 1346–1354 (1996).

[58] Alliney, S., "A property of the minimum vectors of a regularizing functional defined by means of the absolute norm," *Signal Processing, IEEE Transactions on* **45**(4), 913–917 (1997).

[59] Nikolova, M., "Minimizers of cost-functions involving nonsmooth data-fidelity terms. application to the processing of outliers," *SIAM Journal on Numerical Analysis* **40**(3), 965–994 (2002).

[60] Nikolova, M., "A variational approach to remove outliers and impulse noise," *Journal of Mathematical Imaging and Vision* **20**(1), 99–120 (2004).

[61] Chan, T. F. and Esedoglu, S., "Aspects of total variation regularized l 1 function approximation," *SIAM Journal on Applied Mathematics* **65**(5), 1817–1837 (2005).

[62] Chan, T., Esedoglu, S., and Nikolova, M., "Algorithms for finding global minimizers of image segmentation and denoising models," *SIAM Journal on Applied Mathematics* **66**(5), 1632–1648 (2006).

[63] Chen, T., Yin, W., Zhou, X. S., Comaniciu, D., and Huang, T. S., "Total variation models for variable lighting face recognition," *Pattern Analysis and Machine Intelligence, IEEE Transactions on* **28**(9), 1519–1524 (2006).

[64] Yin, W., Goldfarb, D., and Osher, S., "The total variation regularized l^1 model for multiscale decomposition," *Multiscale Modeling & Simulation* **6**(1), 190–211 (2007).

[65] Yang, J., Zhang, Y., and Yin, W., "An efficient tvl1 algorithm for deblurring multichannel images corrupted by impulsive noise," *SIAM Journal on Scientific Computing* **31**(4), 2842–2865 (2009).

[66] Hong, M. and Luo, Z.-Q., "On the linear convergence of the alternating direction method of multipliers," *arXiv preprint arXiv:1208.3922* (2012).

[67] Deng, W. and Yin, W., "On the global and linear convergence of the generalized alternating direction method of multipliers," *Journal of Scientific Computing* , 1–28 (2012).

[68] Lin, T., Ma, S., and Zhang, S., "On the global linear convergence of the admm with multi-block variables," *arXiv preprint arXiv:1408.4266* (2014).

[69] Davis, D. and Yin, W., "Convergence rate analysis of several splitting schemes," *arXiv preprint arXiv:1406.4834* (2014).

[70] Nishihara, R., Lessard, L., Recht, B., Packard, A., and Jordan, M. I., "A general analysis of the convergence of admm," *arXiv preprint arXiv:1502.02009* (2015).

[71] LeCun, Y. and Cortes, C., "The mnist database of handwritten digits," *URL http://yann.lecun.com/exdb/mnist/* (1998).

[72] Bache, K. and Lichman, M., "Uci machine learning repository," *URL http://archive. ics. uci. edu/ml* **901** (2013).

[73] Tibshirani, R., Hastie, T., Narasimhan, B., and Chu, G., "Diagnosis of multiple cancer types by shrunken centroids of gene expression," *Proceedings of the National Academy of Sciences* **99**(10), 6567–6572 (2002).

[74] Bickel, P. and Levina, E., "Some theory for fisher's linear discriminant function,naive bayes', and some alternatives when there are many more variables than observations," *Bernoulli* **10**(6), 989–1010 (2004).

[75] Fan, J. and Fan, Y., "High dimensional classification using features annealed independence rules," *Annals of statistics* **36**(6), 2605 (2008).

[76] Fan, J., Feng, Y., and Tong, X., "A road to classification in high dimensional space," *Journal of the Royal Statistical Society: Series B, to appear* (2012).

[77] Cai, T. and Liu, W., "A direct estimation approach to sparse linear discriminant analysis," *Journal of the American Statistical Association* **106**(496), 1566–1577 (2011).

[78] Hao, N., Dong, B., and Fan, J., "Sparsifying the fisher linear discriminant by rotation," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* (2014). http://dx.doi.org/10.1111/rssb.12092.

[79] Golub, T., Slonim, D., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J., Coller, H., Loh, M., Downing, J., Caligiuri, M., et al., "Molecular classification of cancer: class discovery and class prediction by gene expression monitoring," *Science* **286**(5439), 531–537 (1999).

[80] Gordon, G., Jensen, R., Hsiao, L., Gullans, S., Blumenstock, J., Ramaswamy, S., Richards, W., Sugarbaker, D., and Bueno, R., "Translation of microarray data into clinically relevant cancer diagnostic tests using gene expression ratios in lung cancer and mesothelioma," *Cancer research* **62**(17), 4963 (2002).

[81] Van der Maaten, L. and Hinton, G., "Visualizing data using t-sne," *Journal of Machine Learning Research* **9**(2579-2605), 85 (2008).

[82] Person, K., "On lines and planes of closest fit to system of points in space," *Philiosophical Magazine* **2**, 559–572 (1901).

[83] Torgerson, W. S., "Multidimensional scaling: I. theory and method," *Psychometrika* **17**(4), 401–419 (1952).

[84] Belkin, M. and Niyogi, P., "Laplacian eigenmaps and spectral techniques for embedding and clustering.," in [*NIPS*], **14**, 585–591 (2001).

[85] van der Maaten, L. J., Postma, E. O., and van den Herik, H. J., "Dimensionality reduction: A comparative review," *Journal of Machine Learning Research* **10**(1-41), 66–71 (2009).