TOPICS ON EUCLIDEAN DISTANCE MATRIX AND UNSUPERVISED ENSEMBLE LEARNING

by

Luwan Zhang

A dissertation submitted in partial fulfillment of
the requirements for the degree of

Doctor of Philosophy

(Statistics)

at the

UNIVERSITY OF WISCONSIN–MADISON

2017

Date of final oral examination: 05/12/2017

The dissertation is approved by the following members of the Final Oral Committee:

Ming Yuan, Professor, Statistics
Grace Wahba, IJ Schoenberg-Hilldale Professor, Statistics
Karl Rohe, Assistant Professor, Statistics
Xiaojin Zhu, Sheldon & Marianne Lubar Professor, Computer Sciences
Anru Zhang, Assistant Professor, Statistics

ACKNOWLEDGMENTS

Time flies. Now that I am finally about to make myself to the completion of my thesis and the ensuing Ph.D, I can't help recalling the moment when my flight landed at the Dane county regional airport. It was a late afternoon in the summer of 2012, bright and breezy. In front of me was a brand new world full of unknowns that await me to explore. I was curious, passionate and optimistic. From that moment, I set my foot on the journey of my PhD pursuit. This journey was not always smooth and filled with lessons, hardships, and setbacks. I am proud of myself that those roadblocks never weaken my determination and make me a stronger person to fight for bigger goals in the future.

Writing acknowledgments sounds like a breeze after wrestling with all the other chapters, but to me it is the hardest and most emotional part. The several that I am about to mention are the most influential people that make a big difference in my life. You are the people I could always turn to when I seek guidance in this crazy but exciting world, and for that I am forever grateful.

I would like to express my deepest gratitude to my advisor, Professor Ming Yuan, for his continuous support, patience, caring and motivation through the entire course of my Ph.D study. No one can be as flexible as Ming is to give me freedom to select topics I like to work on, thanks to his immense knowledge and broad research interests. He could always enlighten me whenever I got stuck. His sharp insights can always save me a lot of energy discerning the right direction for my research to go. Last year, at the time when I felt the most disoriented, Ming introduced the field of computational biology to me and gave me a great opportunity to collaborate with one of the leading biology groups. I never feel fortunate enough for having such an advisor who keeps helping me explore new territories. Besides the research, Ming is constantly caring about me and standing on my side like my own old brother. He always listened closely despite being incredibly busy. I can not count how many times he articulated concerns directly and provided clear perspectives on my tough situations. I can never be more grateful on how he has inspired me to persistently hold myself to a higher standard in such a way that I can make the most progress

possible on the journey to the success. Thank you Ming for being my advisor and mentor!

I would like to give special thanks to Professor Grace Wahba. It is my great honor and privilege to work with such a famous statistician. Grace is so approachable and nice to speak with. She is always happy to know my research progress and instruct me whenever I need her. The long-standing Thursday group meeting that she has been chairing provides me a wonderful platform to improve my presentation skills and to brainstorm with other brilliant minds. Grace always comes to the meeting with the most captivating smile and she radiates a lot of positive energies by sharing her colorful and lively life. I admire Grace not only because of her great accomplishments and contributions in both Statistics and Computer Sciences, but for her endless curiosity and engagement in the cutting-edge research as well. Thank you Grace for being a role model that I wish to become!

My sincere thanks also go to the rest of my thesis committee members: Professor Xiaojin Zhu, Professor Karl Rohe and Professor Anru Zhang. My thesis can not be that well-shaped without your insightful comments and hard questions. In particular, I want to thank Professor Xiaojin Zhu for your impressive teaching that spurred my interests in the area of Machine Learning and Artificial Intelligence. I also want to thank Professor Karl Rohe and Professor Anru Zhang for showing me good examples on how to become a successful young researcher after graduation. I will continue to learn from you!

My time at Madison was made more enjoyable in large part due to the friends surrounding me and the sports I am fond of. I would like to thank all student members in the Thursday group meeting, including Zhigeng Geng, Tai Qin, Jing Kong, Han Chen, Shulei Wang, Xiaowu Dai, Hao Zhou, Yilin Zhang, Yuan Li, Cuize Han and Duzhe Wang, for presenting your beautiful works and sharing new ideas with me. Your hard working motivates me to keep fueling! Research becomes more fun with you together! As an active person, I have been swimming for over 20 years. Swimming has been my irreplaceable outlet for stress and I truly appreciate the post-workout freshness and invigoration. I also feel lucky that I discovered yoga two years ago. Since practicing yoga, I have been more confident, powerful and

concentrated.

Last but not the least, I would like to thank my parents who give me endless support and unconditional love all the time. You are bright, open-minded and always create the best environment to help me develop critical thinking and teach me to make independent judgements. Thank you for chatting with me till I fell asleep in countless lonely weekend nights. Thank you for tolerating my bad temper and frustrating words at times. Thank you for cheering me up!

CONTENTS

LIST OF TABLES

LIST OF FIGURES

ABSTRACT

This thesis is devoted to the study of Euclidean distance matrix and unsupervised ensemble learning under the high-dimensional setting. It consists of three pieces of work, focusing on proposing a shrinkage estimator of Euclidean distance matrix, distance-based ANOVA models with application to HIV virus sequence variation study, and developing a new method for unsupervised ensemble learning. A brief description is given in the following for each part of the thesis.

In the first part of thesis, we discuss the problem of recovering an Euclidean distance matrix from noisy or imperfect observations of pairwise dissimilarity scores between a set of objects. This problem naturally arises in many different contexts as it allows us to map objects from an arbitrary domain to Euclidean spaces, and therefore facilitates subsequent statistical analyses. It also provides tools for visualization. In this chapter, we introduce a novel yet simple estimator of an Euclidean distance matrix based on the so called regularized kernel estimator. We show that such an estimator can be elegantly characterized as applying a constant amount of shrinkage to all observed pairwise distances. This fact allows us to establish risk bounds for this estimator, implying that the true distances can be estimated consistently in an average sense as the number of objects increases. In addition, such a characterization suggests an efficient algorithm to compute the distance matrix estimator, as an alternative to the usual second order cone programming known not to scale well for large problems. Numerical experiments and an application in visualizing the diversity of Vpu protein sequences from a recent HIV-1 study further demonstrate the practical merits of the proposed method.

As a sequel of Chapter 1, the second part pays attention to conducting statistical analyses after mapping a set of objects from an arbitrary domain to the Euclidean space. In this chapter, we specifically consider the generalization of ANOVA models to non-Euclidean data by proposing an Euclidean distance-based ANOVA framework(DANOVA). This framework allows us to investigate statistically significant variance components for a given set of objects from an arbitrary domain. The essence of DANOVA relies on the one-to-one correspondence between an Euclidean

distance matrix and its minimum trace kernel established in Chapter 1. Given a general dissimilarity matrix X based on a collection of objects $\{O_i\}_{i=1}^n$ from domain $\mathcal{O}$, a set of minimal Euclidean embeddings $\{P_i\}_{i=1}^n$ can be obtained through the projection of X onto the Euclidean distance matrix cone. This set of embeddings, as a set of messengers, carries all information regarding the original objects to the Euclidean space. All (M)ANOVA-typed test statistics can be constructed in some form of inner/outer products among these embeddings. The test significance can be further investigated using a resampling method. The efficacy of the DANOVA framework is further demonstrated though extensive simulation studies. Finally, the DANOVA framework is applied to the aforementioned Vpu protein sequences to examine the effect of HIV infection progression on the sequence diversity.

The third part mainly concerns developing a new ensemble method for classification problems when the true class labels are not available(a.k.a unsupervised setting). The motivation arises from an intrinsic drawback of crowdsourcing, in which an classifer's quality is not measurable. Therefore, a crowdsourcing-based prediction is very sensitive and unstable because it heavily depends on each constituent. In this work, we propose a two-step procedure including a pruning step to get rid of classifiers with poor performance and a subsequent predicting step to estimate the well-known Bayes classifier using the pruned ensemble. We model the joint distribution of all available classifiers together with the underlying true classifier as an Ising model. The presence of an edge between an available classifier and the true classifier indicates their dependence relationship whereas the absence suggests such a classifier provides no additional information on the truth and should be eliminated from the ensemble. The key in the pruning step is to estimate the neighbourhood of each available classifier by performing ℓ_1-regularized logistic regression based on the Ising model approximation. We could show the existence and uniqueness of an Ising model that minimizes the Kullback-Leibler divergence to the marginal distribution of the ensemble. The neighbourhood of the root can be further recovered successfully with exponentially decaying error under the high-dimensional setting.

Finally, we conclude the thesis in Chapter 4.

1 DISTANCE SHRINKAGE AND EUCLIDEAN EMBEDDING VIA REGULARIZED KERNEL ESTIMATION

1.1 Introduction

The problem of recovering an Euclidean distance matrix from noisy or imperfect observations of pairwise dissimilarity scores between a set of objects arises naturally in many different contexts. It allows us to map objects from an arbitrary domain to Euclidean spaces, and therefore makes them amenable for subsequent statistical analyses, and also provides tools for visualization. Consider, for example, evaluating (dis)similarity between molecular sequences. A standard approach is through sequence alignment and measuring the (dis)similarity between a pair of sequences using their corresponding alignment score (see, Durbin et al., 1998 [12]). Although encoding invaluable insights into the relationship between sequences, it is well known that these scores do not correspond directly to a distance metric in the respective sequence space and therefore cannot be employed in kernel based learning methods. Similarly, there are also numerous other instances where it is possible to derive similarity or dissimilarity scores for pairs of objects from expert knowledge or other information, which, if successfully converted into positive semi-definite kernels or Euclidean distances, could allow themselves to play an important role in a myriads of statistical and computational analyses (e.g., Schölkopf and Smola, 2002 [39]; Székely, Rizzo and Bakirov, 2007 [42]).

A canonical example where this type of problem occurs is multidimensional scaling which aims to place each object in a low dimensional Euclidean space such that the between-object distances are preserved as well as possible. As such it also forms the basis for several other more recent approaches to nonlinear dimension reduction and manifold learning. See, Schölkopf (1998) [41] , Tenenbaum, De Silva and Langford (2000) [43], Lu et al. (2005) [24], Venna and Kaski (2006) [46], Chen and Buja (2009, 2013) [5, 6] among others. Despite the popularity of multidimensional scaling, very little is known about to what extent the distances among the embedded

points could faithfully reflect the true pairwise distances when observed with noises; and it is largely used only as an exploratory data analysis tool.

Another example where it is of interest to reconstruct an Euclidean distance matrix is the determination of molecular structures using nuclear magnetic resonance (NMR, for short) spectroscopy, a technique pioneered by Nobel laureate Kurt Wüthrich (see, e.g., Wüthrich, 1986 [49]). As demonstrated by Wüthrich, distances between atoms could be inferred from chemical shifts measured by NMR spectroscopy. These distances obviously need to conform to a three dimensional Euclidean space yet experimental data on distances are inevitably noisy and as a result, the observed distances may not translate directly into locations of these atoms in a stable structure. Therefore, this becomes a problem of recovering an Euclidean distance matrix in 3D from noisy observations of pairwise distances. Similar problems also occur in graph realization and Euclidean representation of graphs where the goal is to embed the vertex set of a graph in an Euclidean space in such a fashion that the distance between two embedded vertexes matches their corresponding edge weight (see, e.g., Pouzet, 1979 [33]). While an exact embedding of a graph is typically of very high dimension, it is useful in some applications to instead seek approximate yet low dimensional embeddings (see, e.g., Roy, 2010 [37]).

More specifically, let $\{O_i : i = 1, 2, \ldots, n\}$ be a collection of objects from domain $\mathcal{O}$ which could be the coordinates of atoms in the case of molecular structure determination using NMR spectroscopy, or the vertex set of a graph in the case of graph realization. Let $\{x_{ij} : 1 \leqslant i < j \leqslant n\}$ be the observed dissimilarity scores between them such that

$$x_{ij} = d_{ij} + \varepsilon_{ij}, \qquad 1 \leqslant i < j \leqslant n,$$

where ε_{ij}s are the measurement errors and $D = (d_{ij})_{1 \leqslant i,j \leqslant n}$ is a so-called Euclidean distance matrix in that there exist points $p_1, \ldots, p_n \in \mathbb{R}^k$ for some $k \in \mathbb{N}$ such that

$$d_{ij} = \|p_i - p_j\|^2, \qquad 1 \leqslant i < j \leqslant n; \tag{1.1}$$

see, e.g., (Darrotto, 2013 [8]). Here $\|\cdot\|$ stands for the usual Euclidean distance . Our goal is to estimate the Euclidean distance matrix D from the observed matrix $X = (x_{ij})_{1\leqslant i,j\leqslant n}$ where we adopt the convention that $x_{ji} = x_{ij}$ and $x_{ii} = 0$. In the light of (1.1), D can be identified with the points p_is, which suggests an embedding of O_is in $\mathbb{R}^k$. Obviously, if O_is can be embedded in the Euclidean space of a particular dimension, then it is also possible to embed them in a higher dimensional Euclidean space. We refer to the smallest k in which such an embedding is possible as the embedding dimension of D, denoted by $\dim(D)$. As is clear from the aforementioned examples, oftentimes, either the true Euclidean distance matrix D itself is of low embedding dimension; or we are interested in an approximation of D that allows for a low dimensional embedding. Such is the case, for example, for molecular structure determination where the the embedding dimension of the true distance matrix D is necessarily three. Similarly, for multidimensional scaling or graph realization, we typically are interested in mapping objects in two or three dimensions.

Recall that

$$d_{ij} = p_i^\top p_i + p_j^\top p_j - 2p_i^\top p_j,$$

which relates D to the so-called kernel (or Gram) matrix $K = (p_i^\top p_j)_{1\leqslant i,j\leqslant n}$. Furthermore, it is also clear that the embedding dimension $\dim(D)$ equals to $\mathsf{rank}(K)$. Motivated by this correspondence between an Euclidean distance matrix and a kernel matrix, we consider estimating D by $\hat{D} = (\hat{d}_{ij})_{1\leqslant i,j\leqslant n}$ where

$$\widehat{d}_{ij} = \left\langle \hat{K}, (e_i - e_j)(e_i - e_j)^\top \right\rangle = \hat{k}_{ii} + \hat{k}_{jj} - 2\hat{k}_{ij}. \tag{1.2}$$

Here $\langle A, B\rangle = \mathsf{trace}(A^\top B)$, e_i is the ith column vector of the identity matrix, and $\hat{K} = (\hat{k}_{ij})_{1\leqslant i,j\leqslant n}$ is the the so-called regularized kernel estimate; see, e.g., Lu et al. (2005) [24]. More specifically,

$$\widehat{K} = \underset{M\succeq 0}{\mathrm{argmin}} \left\{ \sum_{1\leqslant i<j\leqslant n} \left(x_{ij} - \left\langle M, (e_i - e_j)(e_i - e_j)^\top \right\rangle \right)^2 + \lambda_n \mathsf{trace}(M) \right\}, \tag{1.3}$$

where $\lambda_n \geqslant 0$ is a tuning parameter that balances the tradeoff between goodness-of-fit and the preference towards an estimate with smaller trace norm. Hereafter, we write $M \succeq 0$ to indicate that a matrix M is positive semi-definite. The trace norm penalty used in defining $\hat{K}$ encourages low-rankness of the estimated kernel matrix and hence low embedding dimension of $\hat{D}$. See, e.g., Lu et al. (2005) [24], Yuan et al. (2007) [51], Negahban and Wainwright (2011) [29], Rohde and Tsybakov (2011) [36], and Lu, Monteiro and Yuan (2012) [25] among many others for similar use of this type of penalty. The goal of the current article is to study the operating characteristics and statistical performance of the estimate $\hat{D}$ defined by (1.2).

A fundamental difficulty in understanding the behavior of the proposed distance matrix estimate $\hat{D}$ comes from the simple observation that a kernel is not identifiable given pairwise distances alone, even without noise, as the latter is preserved under translation while the former is not. Therefore, it is not clear what exactly $\hat{K}$ is estimating, and subsequently what the relationship between $\hat{D}$ and D is. To address this challenge, we introduce a notion of minimum trace kernel to resolve the ambiguity associated with kernel estimation. Understanding of this concept allows us to more directly and explicitly characterize $\hat{D}$ as first applying a constant amount of shrinkage to all observed distances; and then projecting the shrunken distances to an Euclidean distance matrix. Because the distance between a pair of points shrinks when they are projected onto a linear subspace, this characterization offers a geometrical explanation to the ability of $\hat{D}$ to induce low dimensional embeddings. In addition, this direct characterization of $\hat{D}$ also suggests an efficient way to compute it using a version of Dykstra's alternating projection algorithm thanks to the special geometric structure of $\mathcal{D}_n$, the set of $n \times n$ distance matrices. See, e.g., Glunt et al. (1990) [17]. Obviation of semidefinite programming, and more generally second order cone programming's computational expense is the principal advantage of this alternating projection technique. Furthermore, based on this explicit characterization, we establish statistical risk bounds for the discrepancy $\hat{D} - D$ and show that the true distances can be recovered consistently in average if D allows for (approximate) low dimensional embeddings.

The rest of the chapter is organized as follows. In Section 1.2, we discuss in

details the shrinkage effect of the estimate $\hat{D}$ by exploiting the duality between a kernel matrix and an Euclidean distance matrix. Taking advantage of our explicit characterization of $\hat{D}$ and the geometry of the convex cone of Euclidean distance matrices, Section 1.3 establishes risk bounds for $\hat{D}$ and Section 1.4 describes how $\hat{D}$ can be computed using an efficient alternating projection algorithm. The merits of $\hat{D}$ is further illustrated via numerical examples, both simulated and real, in Section 1.5. All proofs are relegated to Appendix.

1.2 Distance Shrinkage

In this section, we show that there is a one-to-one correspondence between an Euclidean distance matrix and a so-called minimum trace kernel; and exploit this duality explicitly to characterize $\hat{D}$.

Minimum Trace Kernels

Despite the popularity of regularized kernel estimate $\hat{K}$, rather little is known about its statistical performance. This is perhaps in a certain sense inevitable because a kernel is not identifiable given pairwise distances alone. To resolve this ambiguity, we introduce the concept of minimum trace kernel, and show that $\hat{K}$ is targeting at the unique minimum trace kernel associated with the true Euclidean distance matrix.

Recall that any $n \times n$ positive semidefinite matrix K can be identified with a set of points $p_1, \ldots, p_n \in \mathbb{R}^k$ for some $k \in \mathbb{N}$ such that $K = PP^\top$ where $P = (p_1, \ldots, p_n)^\top$. At the same time, these points can also be associated with an $n \times n$ Euclidean distance matrix $D = (d_{ij})_{1 \leqslant i,j \leqslant n}$ where

$$d_{ij} = \|p_i - p_j\|^2, \qquad 1 \leqslant i < j \leqslant n.$$

Obviously,

$$d_{ij} = \langle K, B_{ij} \rangle,$$

where

$$B_{ij} = (e_i - e_j)(e_i - e_j)^\top.$$

It is clear that any positive semi-definite matrix M can be a kernel matrix and therefore translated uniquely into a distance matrix. In other words,

$$\mathscr{T}(M) = \mathsf{diag}(M)\mathbf{1}^\top + \mathbf{1}\mathsf{diag}(M)^\top - 2M = (m_{ii} + m_{jj} - 2m_{ij})_{1 \leqslant i,j \leqslant n}$$

is a surjective map from the set $\mathcal{S}_n$ of $n \times n$ positive semi-definite matrices to $\mathcal{D}_n$. Hereafter, we write $\mathbf{1}$ as a vector of ones of conformable dimension. The map $\mathscr{T}$, however, is not injective because, geometrically, translation of the embedding points results in different kernel matrix yet the distance matrix remains unchanged. As a result, it may not be meaningful, in general, to consider reconstruction of a kernel matrix from dissimilarity scores alone.

It turns out that one can easily avoid such an ambiguity by requiring the embeddings to be centered in that $P^\top \mathbf{1} = \mathbf{0}$ where $\mathbf{0}$ is a vector of zeros of conformable dimension. We note that even with the centering, the embeddings as represented by P for any given Euclidean distance matrix still may not be unique as distances are invariant to rigid motions. However, their corresponding kernel matrix, as the following result shows, is indeed uniquely defined. Moreover the kernel matrix can be characterized as having the smallest trace among all kernels that correspond to the same distance matrix, hence will be referred to as the minimum trace kernel.

Theorem 1.1. *Let D be an $n \times n$ distance matrix. Then the preimage of of D under $\mathscr{T}$*

$$\mathcal{M}(D) = \{M \in \mathcal{S}_n : \mathscr{T}(M) = D\}$$

is convex; and $-JDJ/2$ is the unique solution to following convex program

$$\underset{M \in \mathcal{M}(D)}{\mathit{argmin}}\ \mathsf{trace}(M),$$

where $J = I - (\mathbf{1}\mathbf{1}^\top/n)$. In addition, if $p_1, \ldots, p_n \in \mathbb{R}^n$ is an embedding of D such that $p_1 + \ldots + p_n = \mathbf{0}$, then $PP^\top = -JDJ/2$, where $P = (p_1, \ldots, p_n)^\top$.

In the light of Theorem 1.1, $\mathscr{T}$ is bijective when restricted to the set of minimum trace kernels:

$$\mathcal{K} = \{M \succeq 0 : \mathsf{trace}(M) \leqslant \mathsf{trace}(A), \quad \forall A \in \mathcal{M}(\mathscr{T}(M))\}.$$

and its inverse is $\mathscr{R}(M) = -JMJ/2$ as a map from distance matrices to kernels with minimum trace. From this viewpoint, the regularized kernel estimate $\hat{K}$ intends to estimate $\mathscr{R}(D)$ instead of the original data-generating kernel. In addition, it is clear that

Proposition 1.2. *For any $\lambda_n > 0$, the regularized kernel estimate $\widehat{K}$ as defined in (1.3) is a minimum trace kernel. In addition, any embedding $\hat{P}$ of $\hat{K}$, that is $\hat{K} = \hat{P}\hat{P}^\top$, is necessarily centered so that $\hat{P}^\top \mathbf{1} = \mathbf{0}$.*

The relationships among the data-generating kernel K, D, $\mathscr{R}(D)$, regularized kernel estimate $\hat{K}$ as defined by (1.3), and the distance matrix estimate $\hat{D}$ as defined by (1.2) can be described by Figure 1.1.

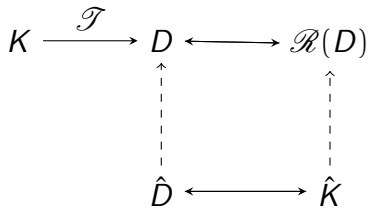


Figure 1.1: Relationships among K, D, $\mathscr{R}(D)$, $\hat{K}$ and $\hat{D}$: the true distance matrix D is determined by the data-generating kernel K; there is a one-to-one correspondence between D and the minimum trace kernel $\mathscr{R}(D)$. Similarly, there is a one-to-one correspondence between $\hat{D}$ and $\hat{K}$ which are estimate of D and $\mathscr{R}(D)$ respectively.

Distance Shrinkage

We now study the properties of the proposed distance matrix estimate given by (1.2). Following Theorem 1.1, $\widehat{D}$ can be equivalently expressed as

$$\widehat{D} = \underset{M \in \mathcal{D}_n}{\mathrm{argmin}} \left\{ \frac{1}{2} \|X - M\|_{\mathrm{F}}^2 + \lambda_n \mathrm{trace}\left(-\frac{1}{2} JMJ \right) \right\}, \tag{1.4}$$

where $\|\cdot\|_{\mathrm{F}}$ stands for the usual matrix Frobenius norm. It turns out that $\hat{D}$ actually allows for a more explicit expression.

To this end, observe that the set $\mathcal{D}_n$ of $n \times n$ Euclidean distance matrices is a closed convex cone (Schönberg, 1935 [38]; Young and Householder, 1938 [50]). Let $\mathscr{P}_{\mathcal{D}_n}$ denote the projection to $\mathcal{D}_n$ in that

$$\mathscr{P}_{\mathcal{D}_n}(A) = \underset{M \in \mathcal{D}_n}{\mathrm{argmin}} \|A - M\|_{\mathrm{F}}^2.$$

for $A \in \mathbb{R}^{n \times n}$. Then

Theorem 1.3. *Let $\widehat{D}$ be defined by (1.2). Then*

$$\widehat{D} = \mathscr{P}_{\mathcal{D}_n}\left(X - \frac{\lambda_n}{2n} D_0 \right)$$

where D_0 is an Euclidean distance matrix whose diagonal elements are zero and off-diagonal entries are ones.

Theorem 1.3 characterizes $\hat{D}$ as the projection of $X - (\lambda_n/2n)D_0$ to an Euclidean distance matrix. Therefore, it can be computed as soon as we can evaluate the projection onto the closed convex set $\mathcal{D}_n$. As shown in Section 1.4, this could be done efficiently using an alternating projection algorithm thanks to the geometric structure of $\mathcal{D}_n$. In addition, subtraction of $(\lambda_n/2n)D_0$ from X amounts to applying a constant shrinkage to all observed pairwise distances. Geometrically, distance shrinkage can be the result of projecting points in an Euclidean space onto a lower dimensional linear subspace, and therefore encourages low dimensional embed-

dings. We now look at the specific example when $n = 3$ to further illustrate such an effect.

In the special case of $n = 3$ points, the projection to Euclidean distance matrices can be computed analytically. Let

$$X = \begin{bmatrix} 0 & x_{12} & x_{13} \\ x_{12} & 0 & x_{23} \\ x_{13} & x_{23} & 0 \end{bmatrix}$$

be the observed distance matrix. We now determine the embedding dimension of $\mathscr{P}_{\mathbb{D}_3}(X - \eta D_0)$.

Let

$$Q = \frac{1}{3+\sqrt{3}} \begin{bmatrix} 2+\sqrt{3} & -1 & -(1+\sqrt{3}) \\ -1 & 2+\sqrt{3} & -(1+\sqrt{3}) \\ -(1+\sqrt{3}) & -(1+\sqrt{3}) & -(1+\sqrt{3}) \end{bmatrix}$$

be a 3×3 Householder matrix. Then, for a 3×3 symmetric hollow matrix X,

$$QXQ = \begin{bmatrix} -\frac{1}{3}x_{12} - \frac{1+\sqrt{3}}{3}x_{13} + \frac{1+\sqrt{3}}{6+3\sqrt{3}}x_{23} & \frac{2}{3}x_{12} - \frac{1}{3}x_{13} - \frac{1}{3}x_{23} & * \\ \frac{2}{3}x_{12} - \frac{1}{3}x_{13} - \frac{1}{3}x_{23} & -\frac{1}{3}x_{12} + \frac{1+\sqrt{3}}{6+3\sqrt{3}}x_{13} - \frac{1+\sqrt{3}}{3}x_{23} & * \\ * & * & * \end{bmatrix},$$

where we only give the 2×2 leading principle matrix of QXQ and leave the other entries unspecified. As shown by Hayden and Wells (1988), the minimal embedding dimension of $\mathscr{P}_{\mathbb{D}_3}(X)$ can be determined by the eigenvalues of the principle matrix.

More specifically, denote by

$$\tilde{D}(X) = \begin{bmatrix} \frac{1}{3}x_{12} + \frac{1+\sqrt{3}}{3}x_{13} - \frac{1+\sqrt{3}}{6+3\sqrt{3}}x_{23} & -\frac{2}{3}x_{12} + \frac{1}{3}x_{13} + \frac{1}{3}x_{23} \\ -\frac{2}{3}x_{12} + \frac{1}{3}x_{13} + \frac{1}{3}x_{23} & \frac{1}{3}x_{12} - \frac{1+\sqrt{3}}{6+3\sqrt{3}}x_{13} + \frac{1+\sqrt{3}}{3}x_{23} \end{bmatrix},$$

and

$$\tilde{D}(X) = U \begin{bmatrix} \alpha_1 & 0 \\ 0 & \alpha_2 \end{bmatrix} U^\top$$

its eigenvalue decomposition. Write

$$\Delta_x := \sqrt{2[(x_{12}-x_{13})^2+(x_{12}-x_{23})^2+(x_{13}-x_{23})^2]}. \tag{1.5}$$

Then, it can be calculated that

$$\alpha_1 = \frac{(x_{12}+x_{13}+x_{23})+\Delta_x}{3}, \qquad \text{and} \qquad \alpha_2 = \frac{(x_{12}+x_{13}+x_{23})-\Delta_x}{3}. \tag{1.6}$$

In the light of Theorem 6.1 of Glunt et al. (1990) [17], we have

Proposition 1.4.

$$\dim(\mathscr{P}_{\mathcal{D}_3}(X)) = \begin{cases} 2 & \textit{if } x_{12}+x_{13}+x_{23} > \Delta_x \\ 1 & \textit{if } -\frac{1}{2}\Delta_x < x_{12}+x_{13}+x_{23} \leqslant \Delta_x \\ 0 & \textit{otherwise} \end{cases},$$

where Δ_x is given by (1.5), and $\dim(\mathscr{P}_{\mathcal{D}_3}(X)) = 0$ *means* $\mathscr{P}_{\mathcal{D}_3}(X) = \mathbf{0}$.

To appreciate the effect of distance shrinkage, consider the case when $\mathscr{P}_{\mathcal{D}_3}(X)$ has a minimum embedding dimension of two. By Proposition 1.4, this is equivalent to assuming $\alpha_2 > 0$. Observe that

$$\tilde{D}(X-\eta D_0) = \tilde{D}(X) - \eta I_2.$$

The eigenvalues of $\tilde{D}(X-\eta D_0)$ are therefore $\alpha_1-\eta$ and $\alpha_2-\eta$ where $\alpha_1 \geqslant \alpha_2$ are the eigenvalues of $\tilde{D}(X)$ as given by (2.3). This indicates that, by applying sufficient amount of distance shrinkage, we can reduce the minimum embedding dimension as illustrated in Figure 1.2.

More specifically,

- If
$$\frac{1}{3}(x_{12}+x_{13}+x_{23}) - \frac{\Delta_x}{3} \leqslant \eta < \frac{1}{3}(x_{12}+x_{13}+x_{23}) + \frac{2\Delta_x}{3},$$
then the minimum embedding dimension of $\mathscr{P}_{\mathcal{D}_3}(X-\eta D_0)$ is one.

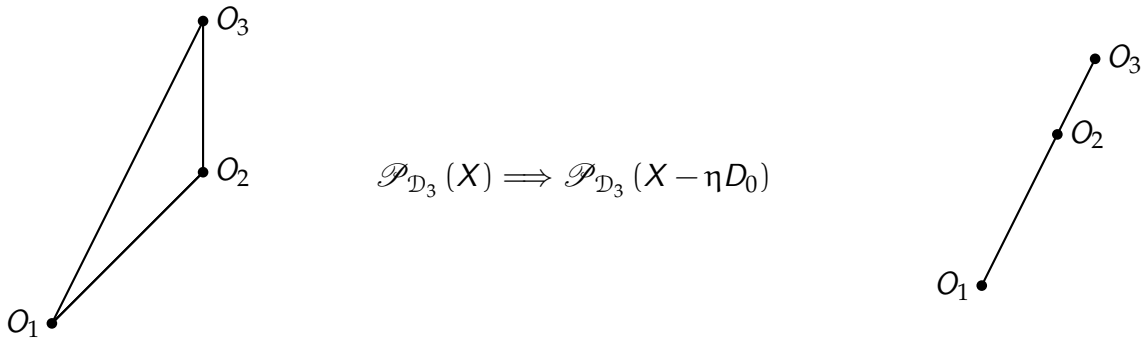


Figure 1.2: Effect of distance shrinkage when $n = 3$.

- If
$$\eta \geqslant \frac{1}{3}(x_{12} + x_{13} + x_{23}) + \frac{2\Delta_x}{3},$$
then the minimum embedding dimension of $\mathscr{P}_{\mathbb{D}_3}(X - \eta D_0)$ is zero;

1.3 Estimation Risk

The previous section provides an explicit characterization of the proposed distance matrix estimate $\hat{D}$ as a distance shrinkage estimator. We now take advantage this characterization to establish statistical risk bounds for $\widehat{D}$.

Estimation Error for Distance Matrix

A natural measure of the quality of a distance matrix estimate $\tilde{D}$ is the averaged squared error of all pairwise distances:

$$L(\tilde{D}, D) := \frac{2}{n(n-1)} \sum_{1 \leqslant i < j \leqslant n} \left(\tilde{d}_{ij} - d_{ij}\right)^2.$$

It is clear that when both $\tilde{D}$ and D are $n \times n$ Euclidean distance matrices,

$$L(\tilde{D}, D) = \frac{1}{n(n-1)} \|\tilde{D} - D\|_{\mathrm{F}}^2.$$

For convenience, we shall now consider bounding $\|\hat{D} - D\|_{\mathrm{F}}^2$. Taking advantage of the characterization of $\hat{D}$ as a projection onto the set of $n \times n$ Euclidean distance matrices, we can derive the following oracle inequality.

Theorem 1.5. *Let $\widehat{D}$ be defined by (1.2). Then for any λ_n such that $\lambda_n \geqslant 2\|X - D\|$,*

$$\|\widehat{D} - D\|_F^2 \leqslant \inf_{M \in \mathcal{D}_n} \left\{ \|M - D\|_F^2 + \frac{9}{4}\lambda_n^2(\dim(M) + 1) \right\},$$

where $\|\cdot\|$ stands for the matrix spectral norm.

Theorem 1.5 gives a deterministic upper bound for the error of $\hat{D}$, $\|\hat{D} - D\|_{\mathrm{F}}^2$ in comparison with that of an arbitrary approximation to D. More specifically, let $\tilde{D}$ be the closest Euclidean distance matrix with embedding dimension r to D, in terms of Frobenius norm. Then Theorem 1.5 implies that with sufficiently large tuning parameter λ_n,

$$L(\hat{D}, D) \leqslant L(\tilde{D}, D) + \frac{Cr\lambda_n^2}{n^2},$$

for some constant $C > 0$. In particular, if D itself is embedding dimension r, then

$$L(\hat{D}, D) \leqslant \frac{Cr\lambda_n^2}{n^2}.$$

More explicit bounds for the estimation error can be derived from this general result. Consider, for example, the case when the observed pairwise distances are the true distances subject to additive noise:

$$x_{ij} = d_{ij} + \varepsilon_{ij}, \qquad 1 \leqslant i < j \leqslant n, \tag{1.7}$$

where the measurement errors ε_{ij}s are independent with mean $\mathbb{E}(\varepsilon_{ij}) = 0$ and variance $\mathrm{var}(\varepsilon_{ij}) = \sigma^2$. Assume that the distributions of measurement errors have light tails such that

$$\mathbb{E}(\varepsilon_{ij})^{2m} \leqslant (c_0 m)^m, \qquad \forall m \in \mathbb{N} \tag{1.8}$$

for some constant $c_0 > 0$. Then the spectral norm of $X - D$ satisfies

$$\|X - D\| = 2\sigma\left(\sqrt{n} + O_p(n^{-1/6})\right).$$

See, e.g., Sinai and Soshnikov (1998) [40]. Thus,

Corollary 1.6. *Let $\widehat{D}$ be defined by (1.2). Under the model given by (1.7) and (1.8), if $\lambda_n = 4\sigma(n^{1/2} + 1)$, then with probability tending to one,*

$$\|\widehat{D} - D\|_F^2 \leqslant \inf_{M \in \mathcal{D}_n} \left\{\|M - D\|_F^2 + 36n\sigma^2(\dim(M) + 1)\right\},$$

as $n \to \infty$. In particular, if $\dim(D) = r$, then with probability tending to one,

$$\|\widehat{D} - D\|_F^2 \leqslant 36n\sigma^2(r + 1).$$

In other words, under the model given by (1.7) and (1.8),

$$L(\hat{D}, D) \leqslant L(\tilde{D}, D) + \frac{Cr\sigma^2}{n},$$

for some constant $C > 0$, where as before, $\tilde{D}$ is the closest Euclidean distance matrix to D with embedding dimension r. In particular, if D itself is embedding dimension r, then

$$L(\hat{D}, D) \leqslant \frac{Cr\sigma^2}{n}.$$

Low Dimensional Approximation

As mentioned before, in some applications, the chief goal may not be to recover D itself but rather its embedding in a prescribed dimension. This is true, in particular, for multidimensional scaling and graph realization where we are often interested in embedding a distance matrix in $\mathbb{R}^2$ or $\mathbb{R}^3$. Following the classical multidimensional

scaling, a parameter of interest in these cases is

$$D_r := \underset{M \in \mathcal{D}_n(r)}{\operatorname{argmin}} \|J(D - M)J\|_{\mathrm{F}}^2,$$

where $\mathcal{D}_n(r)$ is the set of all $n \times n$ Euclidean distance matrices of embedding dimension at most r. An obvious estimate of D_r can be derived by replacing D with $\hat{D}$:

$$\hat{D}_r := \underset{M \in \mathcal{D}_n(r)}{\operatorname{argmin}} \|J(\hat{D} - M)J\|_{\mathrm{F}}^2. \tag{1.9}$$

Similar to the classical multidimensional scaling, the estimate $\hat{D}_r$ can be computed more explicitly as follows. Let $\hat{K}$ be the regularized kernel estimate corresponding to $\hat{D}$, and $\hat{K} = U\Gamma U^\top$ be its eigenvalue decomposition with $\Gamma = \mathrm{diag}(\gamma_1, \gamma_2, \ldots)$ and $\gamma_1 \geqslant \gamma_2 \geqslant \ldots$. Then $\hat{D}_r = \mathscr{T}(\hat{K}_r)$ where $\hat{K}_r = U\mathrm{diag}(\gamma_1, \ldots, \gamma_r, 0, \ldots)U^\top$.

The risk bounds we derived for $\hat{D}$ can also be translated into that for $\hat{D}_r$. More specifically,

Corollary 1.7. *Let $\hat{D}_r$ be defined by (1.9) where $\hat{D}$ is given by (1.2) with $\lambda_n \geqslant 2\|X - D\|$. Then there exists a numerical constant $C > 0$ such that*

$$\|J(\hat{D}_r - D)J\|_F^2 \leqslant C\left(\min_{M \in \mathcal{D}_n(r)} \|J(D - M)J\|_F^2 + \lambda_n^2 r\right),$$

In particular, under the model given by (1.7) and (1.8), if $\lambda_n = 4\sigma(n^{1/2} + 1)$, then with probability tending to one,

$$\|J(\hat{D}_r - D)J\|_F^2 \leqslant C\left(\min_{M \in \mathcal{D}_n(r)} \|J(D - M)J\|_F^2 + nr\sigma^2\right).$$

1.4 Computation

It is not hard to see that the optimization problem involved in defining the regularized kernel estimate can be formulated as a second order cone program (see, e.g.,

Lu et al. 2005 [24]; Yuan et al., 2007 [51]). This class of optimization problems can be readily solved using generic solvers such as SDPT3 (Toh, Todd and Tutuncu, 1999 [44]; Tutuncu, Toh and Todd, 2003 [45]). Although in principle, these problems can be solved in polynomial time, on the practical side, the solvers are known not to scale well to large problems. Instead of starting from the regularized kernel estimate, as shown in Section 1.3, $\hat{D}$ can be directly computed as a projection onto the set of Euclidean distance matrices. Taking advantage of this direct characterization and the particular geometric structure of the closed convex cone $\mathcal{D}_n$, we can devise a more efficient algorithm to compute $\hat{D}$.

We shall adopt, in particular, an alternating projection algorithm introduced by Dykstra (1983) [13]. Dykstra's algorithm is a refinement of the von Neumann alternating projection algorithm specifically designed to compute projection onto the intersection of two closed convex sets by constructing a sequence of projections to the two sets alternatively.

Data: x.
Result: Projection of x onto the intersection of two closed convex set $\mathcal{C}_1$ and $\mathcal{C}_2$.
Initialization: $x_0 = x$, $p_0 = 0$, $q_0 = 0$, $k = 0$;
repeat
  $s_k \leftarrow \mathscr{P}_{\mathcal{C}_1}(x_k + p_k)$;
  $p_{k+1} \leftarrow x_k + p_k - s_k$;
  $x_{k+1} \leftarrow \mathscr{P}_{\mathcal{C}_2}(s_k + q_k)$;
  $q_{k+1} \leftarrow s_k + q_k - x_{k+1}$;
  $k \leftarrow k + 1$;
until a certain convergence criterion is met;
return x_{k+1} ;

Algorithm 1: Dykstra's alternating projection algorithm: $\mathscr{P}_{\mathcal{C}_1}$ and $\mathscr{P}_{\mathcal{C}_2}$ are the projections onto $\mathcal{C}_1$ and $\mathcal{C}_2$ respectively.

The idea can also be illustrated by Figure 1.3 where the projection of a point onto the intersection of two half-planes is computed.

Now consider evaluating $\hat{D}$ which is the projection of $X - \eta_n D_0$ onto $\mathcal{D}_n$. Observe

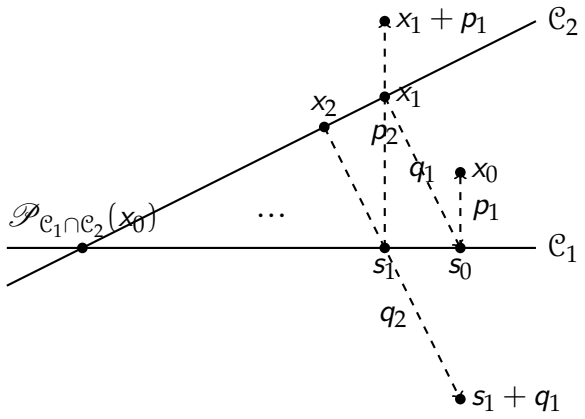


Figure 1.3: Illustration of alternating projection algorithm.

that $\mathcal{D}_n$ is the intersection of two closed convex cones:

$$\mathcal{C}_1 = \{M \in \mathbb{R}^{n\times n} : JMJ \preceq 0\},$$

and

$$\mathcal{C}_2 = \{M \in \mathbb{R}^{n\times n} : \mathsf{diag}(M) = \mathbf{0}\}.$$

Dykstra's alternating projection algorithm can then be readily applied with input $X - \eta_n D_0$. The use of alternating projection algorithms is motivated by the fact that although $\mathscr{P}_{\mathcal{C}_1\cap\mathcal{C}_2}$ is difficult to evaluate, projections to $\mathcal{C}_1$ and $\mathcal{C}_2$ actually have explicit form and are easy to compute.

More specifically, for any symmetric matrix $A \in \mathbb{R}^{n\times n}$, let $\bar{A}_{11}$ be the $(n-1)$th leading principle submatrix of its Householder transform QAQ where $Q = I - vv^\top/n$ and $v = [1, \ldots, 1, 1+\sqrt{n}]^\top$. In other words,

$$A = Q \begin{bmatrix} \bar{A}_{11} & \bar{A}_{12} \\ \bar{A}_{21} & \bar{A}_{22} \end{bmatrix} Q$$

Let $\bar{A}_{11} = U\Gamma U^\top$ be its eigenvalue decomposition. Then

$$\mathscr{P}_{\mathcal{C}_1}(A) = Q \begin{bmatrix} U\Gamma^+ U^\top & \bar{A}_{12} \\ \bar{A}_{21} & \bar{A}_{22} \end{bmatrix} Q$$

where $\Gamma^+ = \mathsf{diag}(\max\{\gamma_{ii}, 0\})$. See Hayden et al. (1988) [20]. On the other hand, it is clear that $\mathscr{P}_{\mathcal{C}_2}(A)$ simply replaces all diagonal entries of A with zeros.

Dealing with Missing Data

We have thus far focused on the case when all pairwise distances are observable. Although this is true in many applications, there are also situations where some of the distances may not be available. Missing data can be conveniently handled within our framework through a combination of the alternating projection and EM algorithm. More specifically, recall that $\Omega \in \{(i,j) : 1 \leqslant i < j \leqslant n\}$ is the set of entries observed in X. As the complete data case, we proceed to estimate D by $\hat{D}^\Omega = \mathscr{T}(\hat{K}^\Omega)$ where

$$\hat{K}^\Omega = \underset{M \succeq 0}{\mathsf{argmin}} \left\{ \sum_{(i,j)\in\Omega} \left(x_{ij} - \left\langle M, (e_i - e_j)(e_i - e_j)^\top \right\rangle \right)^2 + \lambda_n \mathsf{trace}(M) \right\}.$$

Here we use the superscript Ω to signify the dependence on the set Ω of the observed entries. Unlike the case without missing data, Dbâ„¦ in general can not be characterized as a projection of $X_\Omega = (x_{ij})_{(i,j)\in\Omega}$ onto the set of Euclidean distance matrices. To address this difficulty, we iterate between an E step where the missing observations are imputed using the current estimate of the pairwise distances, and an M step where we can appeal to the alternating projection algorithm on the observed distances along with those imputed in the E step.

Data: $X_\Omega = (x_{ij})_{(i,j)\in\Omega}$, $\eta_n \geqslant 0$
Result: $\hat{D}$
Initialization: initialize x_{ij} for $i < j$ and $(i,j) \notin \Omega$, and let
$X = X^\top = (x_{ij})_{1\leqslant i,j\leqslant n}$ where $x_{ii} = 0$; $k = 0$, and $X^{(0)} = X$;
repeat
  M Step – $D^{(k+1)} = \mathscr{P}_{\mathbb{D}_n}(X^{(k)} - \eta_n D_0)$;
  E Step – $x_{ij}^{(k+1)} = x_{ij}$ if $(i,j) \in \Omega$, 0 if $i = j$, and $d_{ij}^{(k+1)}$ otherwise ;
until a certain convergence criterion is met;
$\hat{D} \leftarrow D^{(k+1)}$;
return $\hat{D}$;

Algorithm 2: EM algorithm to handle missing data.

Tuning

The ability to handle missing data also facilitates the tuning of λ_n or equivalently η_n. Clearly,the performance of the proposed method depends on the choice of the tuning parameter. In some cases, we want to embed data into an Euclidean space of a fixed dimensionality. For example, the atoms of a protein have to live in a three dimensional space. To this end, we can experiment with different values of the tuning parameter and use the one corresponding to the desired embedding dimension. Our experience suggests this strategy works fairly well in numerical experiments and the performance of the resulting estimate is also fairly stable for a broad range of tuning parameter choices. In many other situations, however, a more objective choice of tuning parameter may become desirable. A common strategy to address this is through cross-validation, which can be done effectively using the algorithm presented before.

To do cross-validation, we first randomly divide the entries of X into T mutually exclusive subsets: $\Omega^{(1)}, \Omega^{(2)}, \ldots, \Omega^{(T)}$, for some fixed T, so that

$$\Omega^{(1)} \cup \Omega^{(2)} \cup \ldots \cup \Omega^{(T)} = \{(i,j) : 1 \leqslant i < j \leqslant n\}$$

In particular, the choice of $T = 5$ or 10 is often advocated in practice (see, e.g.,

Hastie, Tibshirani and Friedman, 2009 [19]). For each $t = 1, \ldots, T$, we can then apply the algorithm given in the previous subsection to compute the distance shrinkage estimate with a given tuning parameter η_n based on partial observations:

$$X_{-\Omega^{(t)}} := \{X_{ij} : 1 \leqslant i < j \leqslant n, (i,j) \notin \Omega^{(T)}\}$$

Denote by $\hat{D}^{(t),\eta_n}(t = 1, \ldots, T)$ the resulting estimates. We evaluate the suitability of a tuning parameter η_n by its cross validation score:

$$CV(\eta_n) = \frac{1}{T}\sum_{t=1}^{T}\left[\sum_{(i,j)\in\Omega^{(t)}}(X_{ij} - \hat{D}_{ij}^{(t),\eta_n})^2\right]$$

The same procedure can be repeated for a sequence of different values of η_n, and the one associated with the smallest cross valuation score will be selected to the final choice. The distance shrinkage estimate based on this choice of the tuning parameter is then computed based on all observations to yield the final estimate.

1.5 Numerical Examples

To illustrate the practical merits of the proposed methods and the efficacy of the algorithm, we conducted several numerical experiments.

Sequence Variation of Vpu Protein Sequences

The current work was motivated in part by a recent study on the variation of Vpu (HIV-1 virus protein U) protein sequences and their relationship to preservation of tetherin and CD4 counter-activities (Pickering et al., 2014 [32]). Viruses are known for their fast mutation and therefore an important task is to understand the diversity within a viral population. Of particular interest in this study is a Vpu sequence repertoire derived from actively replicating plasma virus from 14 HIV-1-infected individuals. Following standard MACS criteria, five of these individuals can be classified as Long-term nonprogressors, five as rapid progressors, and four as

normal progressors, according to how long the progression from seroconversion to AIDS takes. A total of 304 unique amino acid sequences were obtained from this study.

We first performed pairwise alignment between these amino acid sequences using various BLOSUM substitution matrices. The results using different substitution matrices are fairly similar; and to fix ideas, we shall report here analysis based on the BLOSUM62 matrix. These pairwise similarity scores $\{s_{ij} : 1 \leqslant i \leqslant j \leqslant n\}$ are converted into dissimilarity scores:

$$x_{ij} = s_{ii} + s_{jj} - 2s_{ij}, \qquad \forall 1 \leqslant i < j \leqslant n.$$

As mentioned earlier, $X = (x_{ij})_{1 \leqslant i,j \leqslant n}$ is not an Euclidean distance matrix. To this end, we first applied the classical multidimensional scaling to X. The three dimensional embedding is given in the top left panel of Figure 1.4. The amino acid sequences derived from the same individuals are represented by the same symbol and color. Different colors correspond to the three different classes of disease progression: long-term nonprogressors are represented in red, normal in green, and rapid progressors in purple. For comparison, we also computed $\hat{D}$ with various choices of the tuning parameters. Similar to the observations made by Lu et al. (2005) [24], the corresponding embeddings are qualitatively similar for a wide range of choices of λ_n. A typical one is given in the top right panel of Figure 1.4. It is clear that both embeddings share a lot of similarities. For example, sequences derived from the same individual are more similar as they tend to cluster together. The key difference, however, is that the embedding corresponding to $\hat{D}$ suggests an outlying sequence. We went back to the original pairwise dissimilarity scores and identified the sequence as derived from a rapid progressor. It is fairly clear from the original scores that this sequence is different from the others. The minimum dissimilarity score from the particular sequence to any other sequence is 245 whereas the largest score between any other pair of sequences is 215. The histogram of the scores between the sequence and other sequences, or among other sequences are given in the bottom panel of Figure 1.4.

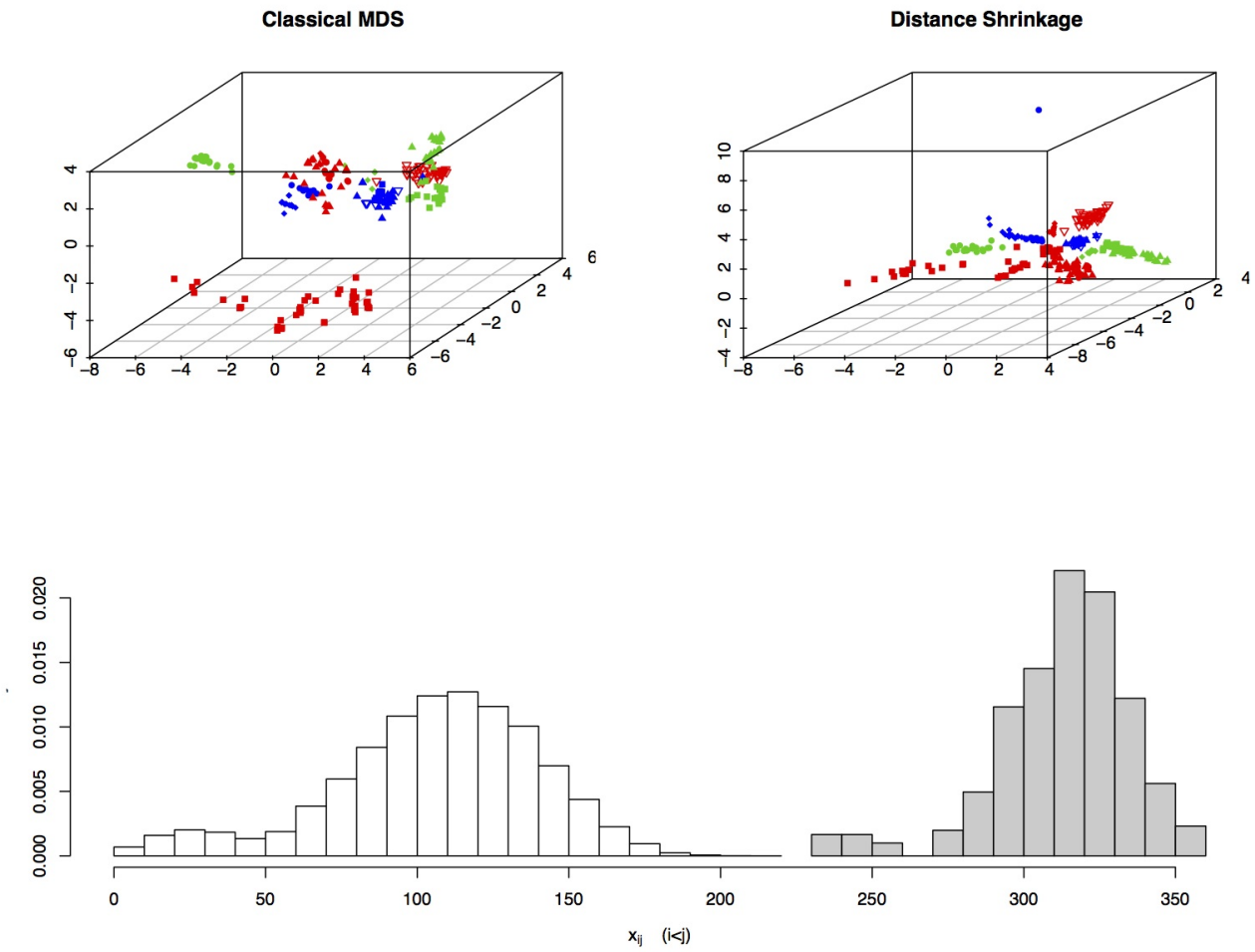


Figure 1.4: Three dimensional embedding for 304 amino acid sequences: the top panels are embeddings from classical multidimensional scaling and distance shrinkage respectively. The histogram of the pairwise dissimilarity scores is given in the bottom panel. The shaded histogram corresponds to those scores between the outlying sequence and the other sequences.

Given these observations, we now consider the analysis with the outlying sequence removed. To gain insight, we consider different choices of λ_n to visually inspect the Euclidean embeddings given by the proposed distance shrinkage. The embeddings given in Figure 1.5 correspond to λ_n equals 4000, 8000, 12000 and 16000 respectively. These embedding are qualitatively similar.

Simulated Examples

To further compare the proposed distance shrinkage approach with the classical multidimensional scaling, we carried out several sets of simulation studies. For illustration purposes, we took the setup of the molecular conformation problem discussed earlier. In particular, we considered the problem of protein folding, a process of a random coil conformed to a physically stable three-dimensional structure equipped with some unique characteristics and functions.

We started by extracting the existing data on the 3D structure of the channel-forming trans-membrane domain of Vpu protein from HIV-1 mentioned before. The data obtained from protein data bank (symbol: 1PJE) contains the 3D coordinates of a total of $n = 91$ atoms. The exact Euclidean distance matrix D was then calculated from these coordinates. We note that in this case the embedding dimension is known to be three. We generated observations x_{ij} by adding an measurement error $\varepsilon_{ij} \sim N(0, \sigma^2)$ for $1 \leqslant i < j \leqslant n$. We considered three different values of σ^2 – 0.05, 0.25 and 0.5 respectively, representing relatively high, medium and low signal to noise ratio. For each value of σ^2, we simulated one hundred datasets and computed for each dataset the Euclidean distance matrix corresponding to the classical multidimensional scaling and the distance shrinkage. We evaluated the performance of each method by the Kruskal's stress defined as $\|\hat{D} - D\|_{\mathrm{F}}/\|D\|_{\mathrm{F}}$. The results are summarized by Table 1.1.

To better appreciate the difference between the two methods, Figure 1.6 gives the ribbon plot of the protein backbone structure corresponding to the true Euclidean distance matrix and the estimated ones from a typical simulation run with different signal to noise ratios. It is noteworthy that the improvement of the distance

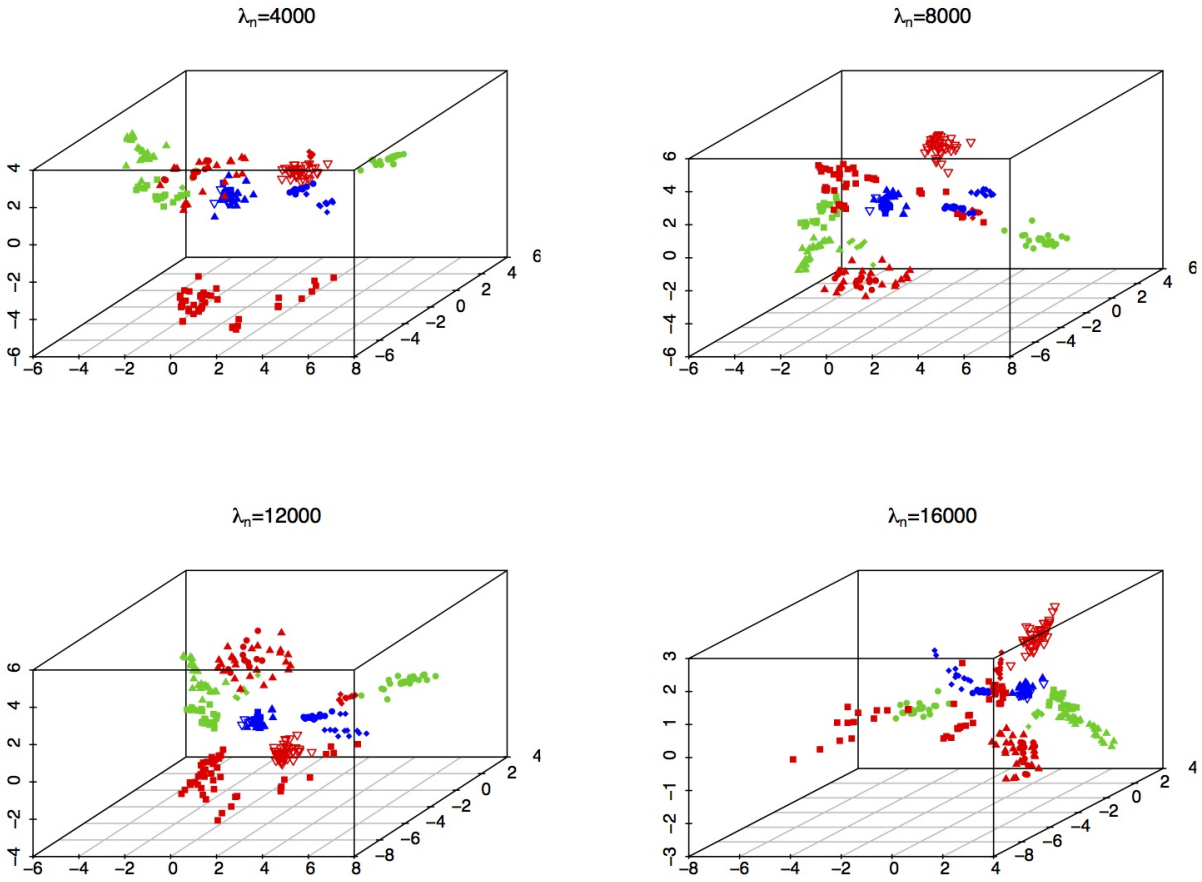


Figure 1.5: Euclidean embedding of 303 amino acid sequences via distance shrinkage: the outlying sequence was removed from the original data and each panel corresponds to different choice of λ_n.

Signal-to-Noise Ratio	Method	Mean	Standard error
High	Distance Shrinkage	0.010	2.0e-04
	Classical MDS	0.078	9.3e-04
Medium	Distance Shrinkage	0.024	4.8e-04
	Classical MDS	0.185	2.5e-03
Low	Distance Shrinkage	0.035	8.4e-04
	Classical MDS	0.301	3.9e-03

Table 1.1: Kruskal's stress for 1PJE data with measurement error.

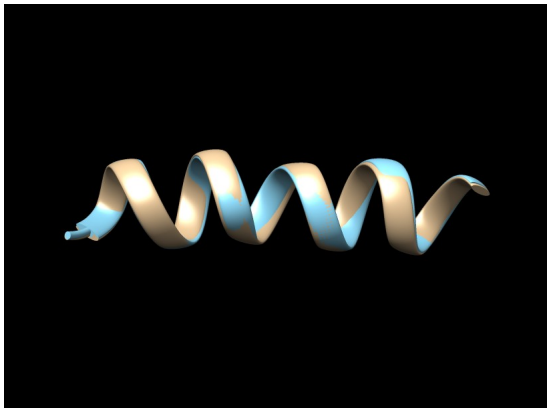
shrinkage over the classical multidimensional scaling becomes more evident with higher level of noise.

Our theoretical analysis suggests better performances for larger number of atoms. To further illustrate this effect of n, we repeated the previous experiment for HIV-1 virus protein U cytoplasmic domain (protein data bank symbol: 2K7Y) consisting of $n = 671$ atoms. We simulated data in the same fashion as before and the Kruskal stress, based on one hundred simulated dataset for each value of σ^2, is reported in Table 1.2. The performance compares favorable with that for 1PJE.

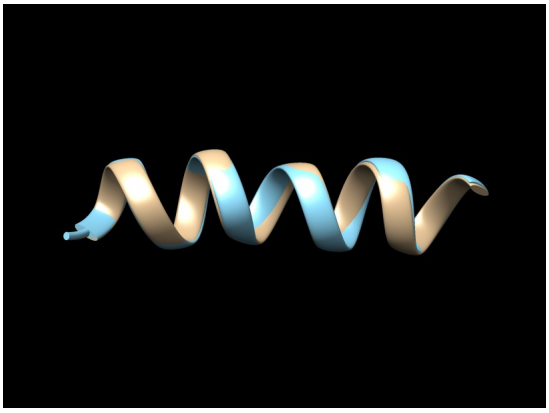
Signal-to-Noise Ratio	Method	mean	standard error
High	Distance Shrinkage	1.66e-04	2.70e-07
	Classical MDS	3.2e-03	4.84e-06
Medium	Distance Shrinkage	8.32e-04	1.48e-06
	Classical MDS	1.61e-02	2.45e-05
Low	Distance Shrinkage	1.7e-03	3.05e-06
	Classical MDS	3.22e-02	5.28e-05

Table 1.2: Kruskal's stress for 2K7Y data with measurement error.

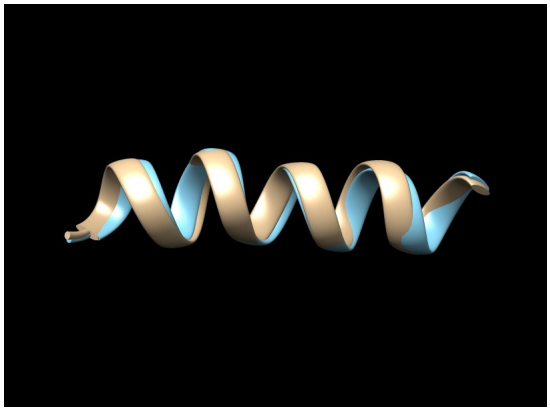
To demonstrate the efficacy of cross-validation as a tuning method, we give in Figure 1.7 the true Kruskal stress as a function of the tuning parameter λ along with the five fold cross validation scores for a typical simulated dataset under each of the three levels of signalto-noise ratio. These plots were generated by computing the distance matrix estimate for a series of values for the tuning parameter. It is



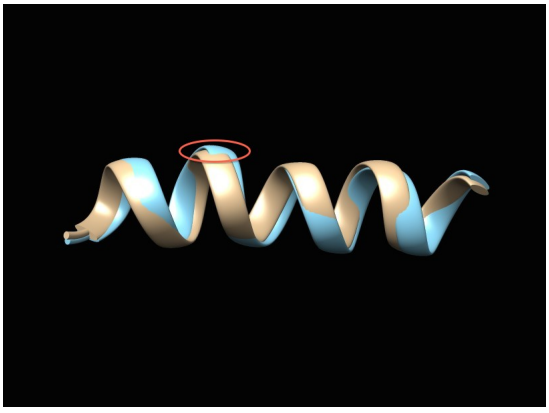
(a) Distance Shrinkage, High signal-to-noise ratio



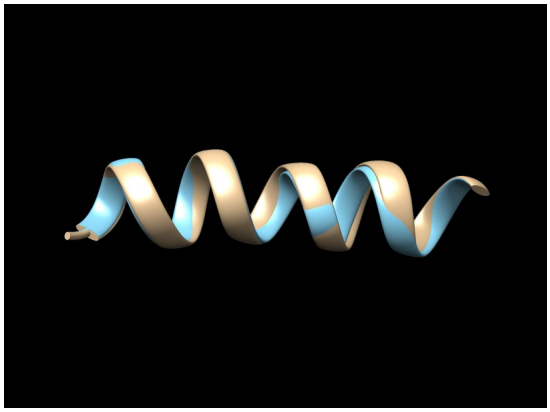
(b) Classical MDS, High signal-to-noise ratio



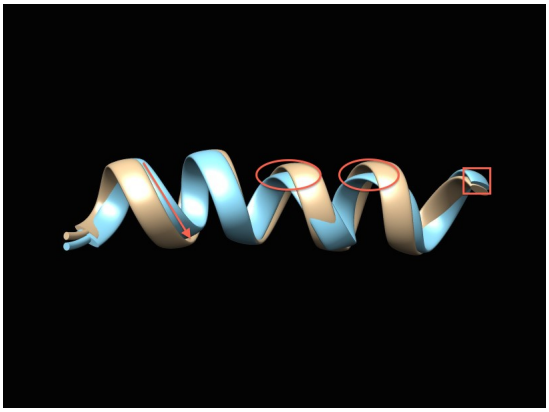
(c) Distance Shrinkage, Medium signal-to-noise ratio



(d) Classical MDS, Medium signal-to-noise ratio



(e) Distance Shrinkage, Low signal-to-noise ratio



(f) Classical MDS, Low signal-to-noise ratio

Figure 1.6: Ribbon plot of 1PJE protein back structure: the true structure is represented in gold whereas the structured corresponding to the estimated Euclidean distance matrix is given in blue. The left panels are for the distance shrinkage estimate whereas the right panels are for the the classical multidimensional scaling. Particular regions where the distance shrinkage shows visible improvement is circled out in red in the right panels.

clear from these plots that the tuning parameter selected by the cross validation is fairly close to optimal choice that minimizes the true Kruskal stress.

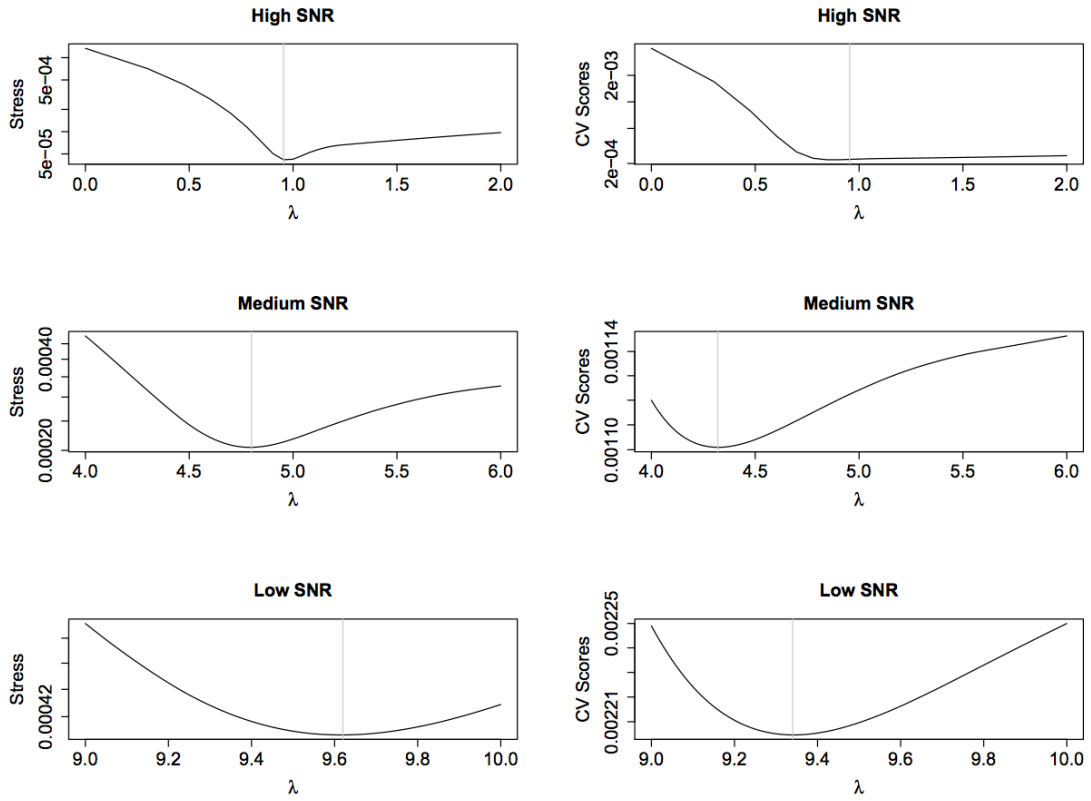


Figure 1.7: Comparison of Kruskal stress and cross-validation scores for simulated 2K7Y data. The left column gives plots of the Krusal stress as a function of the tuning parameter λ for different signal-to-noise ratios, and the right column gives plots of the cross-validation scores. In each panel, the minimizing tuning parameter is marked with the grey vertical line.

In the next set of simulation, we assess the effect of missing data for the proposed distance shrinkage estimate. Similar to before, we take the 3D coordinates data from protein data bank for five different proteins with different number of atoms. Pairwise distances were first computed for each of the protein. To mimic the typical

NMR experiments, we assume that the larger distances are missing. In particular, we consider cases where the top 50%, 25% or 10% of the distances are unobservable. For those observed distances, independent Gaussian measurement errors with mean 0 and variance 0.5 were added. We ran the proposed distance shrinkage estimate on the simulated data. We experimented a range of tuning parameter choices and the performance is fairly similar. The results are summarized in Table 1.3. As expected, the method performs better with the amount of missing data reduces. The distance shrinkage estimate works reasonably well even with 10% of missing data.

PDB ID	# of Atoms	Kruskal's Stress		
		50% Missing	25% Missing	10% Missing
1PTQ	402	.57	.35	.18
1HOE	558	.56	.33	.15
1PHT	811	.56	.34	.17
1AX8	1003	.57	.36	.18

Table 1.3: Effect of Missing Data.

Finally, to further demonstrate the robustness of the approach to non-Gaussian measurement error, we generated pairwise distance scores between the 671 atoms following Gamma distributions:

$$x_{ij} \sim \mathrm{Ga}(d_{ij}, 1), \qquad \forall 1 \leqslant i < j \leqslant 671,$$

so that both the mean and variance of x_{ij} are d_{ij}, where d_{ij} is the true squared distance between the ith and jth atoms. We again applied both classical multidimensional scaling and distance shrinkage to estimate the true distance matrix and reconstruct the 3D folding structure. The result from a typical simulated dataset is given in Figure 1.8.

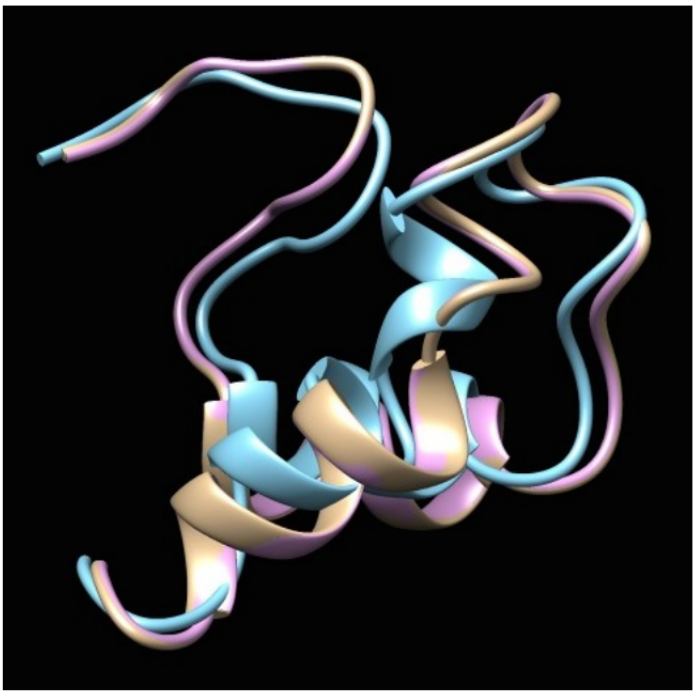


Figure 1.8: Ribbon plot of 2K7Y protein back structure: the true structure, and the structures corresponding to the classical multidimensional scaling and the distance shrinkage estimate are represented in gold, blue and pink respectively.

2 DISTANCE-BASED VARIANCE COMPONENT ANALYSIS WITH APPLICATION TO HIV VIRUS SEQUENCE VARIATION STUDY

2.1 Introduction

HIV, human immunodeficiency virus, is the virus that causes AIDS (Acquired Immunodeficiency Syndrome). Since the epidemic began in the early 1980s, AIDS has been notorious for its incurableness, though numerous effort has been devoted to conduct research in all areas of HIV infections including developing and testing preventive HIV vaccines. The Vpu protein, encoded by HIV-1, is known to be the culprit to attack the human immune system by destroying CD4 positive (CD4+) T cells, a type of white blood cell that is vital to fight against infection. Pickering et. al in [32] studied the preservation of Vpu functions thoughtout the course of HIV indfection. In the paper, they reported an interesting phenomenon that they observed a set of highly diverse Vpu amino acid sequences sampled from 14 individuals with different progression rates, yet no solid evidence has been provided to show whether such sequence variation is due to pure randomness or can be indeed attributed to certain meaningful factors, for instance, different progression rates and/or individual differences. Since the functional variation of a protein largely hinges on its sequence variation, it is vital to discover key factors that could well explain such sequence variation. These factors, once confirmed, may potentially give some insights on developing niche or even personalized HIV preventions and therapies. Despite many potential medical benefits, the confirmation itself is very challenging and not convincing without a rigorous statistical testing procedure. Traditionally, analysis of variance(ANOVA) can be used to analyze differences among several groups of data. However, it can not be readily applied in non-numeric case. To bridge this gap, in this chapter, we would like to expand its territory by introducing a new variant of ANOVA that can be suitable for any type of data.

ANOVA models have witnessed a long successful history of studying variability

of data in many applications. The fundamental technique of ANOVA is a partitioning of the total sum of squares SS into components distributed to different factors and errors. Under the normality assumption, an F-test can be conducted to investigate the significance of each factor. The F test turns out to be a uniformly most powerful(UMP) test. MANOVA, as a natural extension of ANOVA, aims to perform similar tests in the multivariate setting. The partitioning of SS is replaced by a decomposition into several positive definite matrices. However, in the multivariate setting, no UMP test statistic exists unless only two groups of data are under comparison in which case Hotelling's T^2 is the UMP test and invariant to affine transformations. Four most commonly used ones are Wilk's Lambda, Hotelling-Lawley trace, Pillai-Bartlett trace, and Roy's largest root, see for example [2, 23] and references therein. Olson in [30] via extensive simulations concluded that the first three tests have similar power, while the Roy's test has the most power only when the trace of the non-centrality matrix is concentrated in its largest eigenvalue. However, their applicability is largely hampered by the limitation that the working space has to be the Euclidean space where data is intrinsically numeric. Unfortunately, in practice, data oftentimes comes in various formats, to name a few, may it be sequences of amino acids from a large family in molecular biology, text data such as comments from social media, or images in different resolutions from engineering. In all these scenarios, original data points can not be directly fed into any ANOVA/MANOVA model, which makes subsequent statistical inferences infeasible. Thus, a natural question worth of asking is that how to convert the original data through possible transformations or projections into Euclidean space?

Zhang et. al in [52] proposed a distance shrinkage method on a given dissimilarity matrix to get a unique set of Euclidean embeddings in a desired dimension. This set of embeddings, as a set of messengers, carries all information on the original objects to the Euclidean space and makes statistical inferences amenable. In practice, there are many well-defined similarity metrics designed for specific domains to get a dissimilarity matrix. For example, pairwise similarity for any two amino acid sequences can be calculated using a BLOSUM matrix. The likeness of any two sentences can be measured lexically by their cosine similarity or semantically using

the Perl package supplied in WordNet. For any two images, their similarity can be easily calculated based on their rgb-coded matrix.

Along this line, in this chapter, we propose a Euclidean distance-based ANOVA framework(DANOVA) to investigate statistically significant variance components for a given set of objects from an arbitrary domain. Before presenting full details, here we only outline the major idea. Given a collection of objects $\{O_i\}_{i=1}^n$ from domain $\mathcal{O}$, and their dissimilarity matrix using a pre-specified similarity metric denoted by X, the first step is to project these objects to their corresponding minimal Euclidean embeddings $\{P_i\}_{i=1}^n$. Then a test statistic can be constructed in some form of inner/outer products among these embeddings. The test significance can be further investigated using a resampling method. In fact, the idea of using a dissimilarity matrix to do statistical inferences is not new. Anderson in [1] actually proposed a pseudo F-ratio statistic as a surrogate of the usual F statistic used in ANOVA. However, this pseudo F-ratio test statistic only depends on the inner products as opposed the outer products. As the data exihbits more complex variability, this test tends to lose power.

The rest of the chapter is organized as follows. In Section 2.2 we first give a brief review on the Euclidean distance matrix as well as its embeddings. The construction of a distance-based ANOVA table will be elaborated. We further discuss the test statistics and use a permutation test to get the p-value. Some numerical experiments including both simulations and real data analysis are provided in Section 2.3.

2.2 Method

Euclidean Distance Matrix

For a given collection of data points $\{p_i\}_{i=1}^n$ in $\mathbb{R}^k$, the associated Euclidean distance matrix $D = (d_{ij})_{1\leqslant i,j\leqslant n}$ is given by

$$d_{ij} = \|p_i - p_j\|^2 = p_i^\top p_i + p_j^\top p_j - 2p_i^\top p_j, \qquad 1 \leqslant i, j \leqslant n \tag{2.1}$$

Conversely, given an Euclidean distance matrix D, a set of points $\{p_i\}_{i=1}^n$ in $\mathbb{R}^k$ satisfying (2.1) is a set of Euclidean embeddings of dimension k with respect to D. It is easy to see that embeddings of different dimensions could result in the same D because Euclidean distance is invariant to any rigid motions, such as translations, rotations and reflections. The embedding dimension of D, denoted by $\dim(D)$, refers to the smallest k such that there is a set of points in $\mathbb{R}^k$ satisfying (2.1). To determine $\dim(D)$, Zhang et al. in [52] established a one-to-one correspondence by introducing a notion of **minimum trace kernel**. More specifically, for an Euclidean distance matrix D, its minimum trace kernel K is $-JDJ/2$, where $J = I - \mathbf{1}\mathbf{1}^\top/n$ is the centering matrix, and $\dim(D)$ equals to $\mathsf{rank}(K)$.

As we mentioned earlier, in practice, a dissimilarity matrix X, based on a collection of objects $\{O_i\}_{i=1}^n$ from domain $\mathcal{O}$, may not be necessarily an Euclidean distance matrix, therefore we shall project it to the Euclidean space. Let $\mathcal{S}_n$ denote the set of all $n \times n$ symmetric matrices and $\mathcal{D}_n$ denote the set of all $n \times n$ Euclidean distance matrices, then $\mathcal{D}_n$ is an intersection of two closed convex cones, see [38, 50],

$$\mathcal{C}_1 = \{M \in \mathcal{S}_n : JMJ \preceq 0\},$$

and

$$\mathcal{C}_2 = \{M \in \mathcal{S}_n : \mathsf{diag}(M) = \mathbf{0}\}.$$

Since the set $\mathcal{D}_N$ is a closed convex cone, the projection of X to $\mathcal{D}_N$ is uniquely defined denoted by $D = \mathscr{P}_{\mathcal{D}_n}(X)$, where $\mathscr{P}_{\mathcal{D}_n}$ denotes the projection to $\mathcal{D}_N$,

$$\mathscr{P}_{\mathcal{D}_n}(A) = \underset{M \in \mathcal{D}_n}{\mathsf{argmin}} \, \|A - M\|_{\mathrm{F}}^2 \quad A \in \mathbb{R}^{n \times n} \tag{2.2}$$

Now consider the eigen-decomposition of the minimum trace kernel K

$$K = U^\top \Lambda U = P^\top P \tag{2.3}$$

in which $\Lambda = \mathrm{diag}(\lambda_1, \cdots, \lambda_d)$ is the diagonal matrix generated by all nonzero eigenvalues of K and $\lambda_1 \geqslant \cdots \geqslant \lambda_d > 0$. $P = \sqrt{\Lambda} U$ represents a set of Euclidean

embeddings satisfying $P\mathbf{1} = \mathbf{0}$, meaning that all embeddings are centered around the origin. Hereafter, we call such P matrix as a set of **minimal embeddings** of $\{O_i\}_{i=1}^n$ in the Euclidean space. Each column of P denoted as $p_i \in \mathbb{R}^d$ is the minimal embedding of object O_i. In some applications, the minimal embeddings themseleves may not be of primary interest, but rather their approximations in some prescribed dimension $r < d$, in which case only the first r rows in P are needed.

Distance ANOVA(DANOVA)

Now we elaborate our proposed DANOVA procedure and establish the connection to the well-known MANOVA. In MANOVA, there are multiple settings that give rise to a similar hypothesis testing problem. For the ease of illustration, we would like to use perhaps the most common one, which is one-way fixed effect MANOVA model, to convey our core methodology.

One-way design: for a factor of interest A with I levels, n_i random objects at level i are collected from domain $\mathcal{O}$ denoted as $\{O_{ij}, i = 1, \cdots, I, j = 1, \cdots, n_i\}$. In light of (2.2) and (2.3), a set of minimal embeddings $\{p_{ij} \in \mathbb{R}^d, i = 1, \cdots, I, j = 1, \cdots, n_i\}$ can be obtained. According to a classical MANOVA setting, the data generating model follows

$$p_{ij} = \mu + \alpha_i + \varepsilon_{ij}, \quad i = 1, \cdots, I, \quad j = 1, \cdots, n_i \tag{2.4}$$

in which μ stands for the overall centroid, α_i is the i-th drift from μ satisfying $\sum_{i=1}^I n_i \alpha_i = \mathbf{0}$, and $\{\varepsilon_{ij}\}_{i=1,\ldots,I, j=1,\ldots,n_i}$ are random error terms. Let $N = \sum_{i=1}^I n_i$ denote the total number of observations and consider a series of group centroids

$$\bar{p}_{i\cdot} = \sum_{j=1}^{n_i} p_{ij}/n_i, \quad i = 1 \cdots I$$

and the grand centroid

$$\bar{p}_{\cdot\cdot} = \sum_{i=1}^{I} \sum_{j=1}^{n_i} p_{ij}/N$$

Thus, p_{ij} can be decomposed as suggested by model given in (2.4),

$$p_{ij} = \bar{p}_{\cdot\cdot} + (\bar{p}_{i\cdot} - \bar{p}_{\cdot\cdot}) + (p_{ij} - \bar{p}_{i\cdot}) \tag{2.5}$$

The decomposition in (2.5) leads to the sum of outer product matrix partitioning

$$\begin{aligned} &\text{Total:} \quad \sum_{i=1}^{I} \sum_{j=1}^{n_i} (p_{ij} - \bar{p}_{\cdot\cdot})(p_{ij} - \bar{p}_{\cdot\cdot})^\top \\ &\text{Factor A :} \quad \sum_{i=1}^{I} n_i (\bar{p}_{i\cdot} - \bar{p}_{\cdot\cdot})(\bar{p}_{i\cdot} - \bar{p}_{\cdot\cdot})^\top \end{aligned} \tag{2.6}$$

Now denote by B the block diagonal matrix $\begin{bmatrix} \mathbf{1}_{n_1} & & \\ & \mathbf{1}_{n_2} & \\ \cdots & \cdots & \cdots \\ & & \mathbf{1}_{n_I} \end{bmatrix}$ and W the weight matrix $\text{diag}(n_1^{-1}, n_2^{-1}, \cdots, n_I^{-1})$, then the Distance ANOVA table can be constructed given in Table 2.1.

Source	df	Sum of outer product matrix
A	$I-1$	$H := PBWB^\top P^\top$
Error	$N-I$	$E := \Lambda - H$
Total	$N-1$	Λ

Table 2.1: One-way DANOVA Table

Test Statistics

Now we wish to investigate the effect of factor A. It is essential to conduct the following hypothesis test

$$H_0 : \alpha_1 = \cdots = \alpha_I = \mathbf{0} \leftrightarrow H_a : \text{some } \alpha_i \text{ are not } \mathbf{0}$$

Here we specifically consider five test statistics, including Wilk's Lambda, Hotelling-Lawley trace, Pillai-Bartlett trace, Roy's largest root, and pseudo F-ratio. Let $\{\tau_j\}_{j=1}^d$ be the eigenvalues of matrix $H\Lambda^{-1}$, then

$$
\begin{aligned}
&\text{(1) Wilks' Lambda:} \quad \frac{\det(\Lambda - H)}{\det(\Lambda)} = \det(I_d - H\Lambda^{-1}) = \prod_{j=1}^{d}(1-\tau_j) \\
&\text{(2) Lawley-Hotelling trace:} \quad \mathsf{trace}(H(\Lambda - H)^{-1}) = \sum_{j=1}^{d} \frac{\tau_j}{1-\tau_j} \\
&\text{(3) Pillai-Bartlett trace:} \quad \mathsf{trace}(H\Lambda^{-1}) = \sum_{j=1}^{d} \tau_j = \sum_{j=1}^{d} h_{jj}/\lambda_j \\
&\text{(4) Roy's largest root:} \quad \text{maximum eigenvalue of } H(\Lambda - H)^{-1} = \max_{j=1,\cdots,d} \frac{\tau_j}{1-\tau_j} \\
&\text{(5) pseudo F-ratio:} \quad \frac{N-I}{I-1}\left(\frac{-\sum_{i=1}^{I} n_i^{-1}\|D_i\|_1}{\sum_{i=1}^{I} n_i^{-1}\|D_i\|_1 - \mathsf{trace}(D)}\right)
\end{aligned}
\tag{2.7}
$$

It is worth of mentioning that the above test statistics (1)–(4) are invariant of any rigid motions. From the expressions given in (2.7), it is easy to see that these test statistics are fuctions of the eigenvalues of $H\Lambda^{-1}$. Given two sets of minimal embeddings P_1, P_2 w.r.t. D, since they must be related by $P_2 = QP_1 + c$, where Q is an orthogonal matrix in $\mathbb{R}^d$ and $c \in R^d$ is a constant vector, the resulting matrix $H_i\Lambda_i^{-1}, i = 1, 2$ should share the same eigenvalues.

Permutation Test

Theories for the classical MANOVA models usually have to assume two conditions: (i) error terms are iid mutivariate normal random variables (ii) a balanced design. Even under these two conditions, exact distributions of aforementioned test statistics can not be derived unless for a few special cases. The asymptotic distributions will sometimes be used instead. However, in practice, the normality assumption is not easy to check especially when the dimension is high. With cost effectiveness in mind, a balanced design may be too ideal to make for some real projects. Therefore, for

practical concerns, we would like to take an nonparametric approach to approximate the distribution of the test statistic. More specifically, we consider a permutation test, see, for example, Edgington [14], Manly [26].

The idea of a permutation test is as follows. If the factor under the examination did not matter, then with equal probability we can obtain the observations in any order by reference to the factor level. So, another possible value of the test statistic can be calculated by a random shuffling on the observed labelling assignment. This random shuffling can be repeated a large number of times indexed by $\mathcal{I}$. Each time a new value of the test statistic T_i can be calculated. A distibution $\hat{P}_T$ can be derived to approximate the true distribution P_T, based on $\{T_i, i \in \mathcal{I}\}$. When $|\mathcal{I}|$ is sufficiently large, $\hat{P}_T$ will resemble P_T. Thus, the p-value can be calculated by

$$\frac{1 + \sum_{i \in \mathcal{I}} \mathbb{I}(T_i > T_{obs})}{1 + |\mathcal{I}|} \tag{2.8}$$

where T_{obs} represents the value of the test statistic using the actual observations. Note that 1 is added in (2.8) because T_{obs} is one of possible realizations under $\hat{P}_T$.

2.3 Numerical Experiments

Simulations

Protein sequence evolution: in this simulation, we would like to examine the efficacy of the proposed DANOVA procedure in testing the significant divergence along the evolutionary process. More specifically, we construct the transition probability matrix P by the so-called Jukes-Cantor model in which all possible substitutions are equally likely to occur. Consider the map from 20 amino acids as well as a deletion/insertion to integers 1-21, then $P(\pi) = (p_{ij})_{1 \leqslant i,j \leqslant 21}$ describes the

substitution probability in the next generation given the current state

$$p_{ij} := \Pr(X_{t+1} = j | X_t = i) = \begin{cases} 1-\pi & 1 \leqslant i = j \leqslant 20 \\ (\pi-\beta)/19 & 1 \leqslant i \neq j \leqslant 20 \\ 1-20\beta & i \text{ or } j = 21 \\ \beta & i = j = 21 \end{cases}$$

where π measures the overall mutation rate, and β measures the deletion/insertion rate. Obviously, a larger π indicates that more divergence can be expected at a faster rate. We consider two protein sequence generating schemes. The first scheme is region based in which only amino acids within a particular region of the full length are allowed to mutate and the rest shall stay intact. The purpose of this scheme is to mimic the fact that a family of homologous proteins tend to share some highly conserved regions to preserve similar functions. On the contrary, the second scheme allows the whole sequence to mutate and the purpose is to investigate the sensitivty of DANOVA to different mutation rates. Two examples are given in Figure 2.1.

(I) Region based generating scheme

1. Initialize a root sequence of length L, and specify the number of branches g to grow.

2. Specify g non-overlapping regions R_i such that $R_i \cap R_j = \emptyset$ and $\cup_{i=1}^{g} R_i \subseteq \{1, 2, \cdots, L\}$.

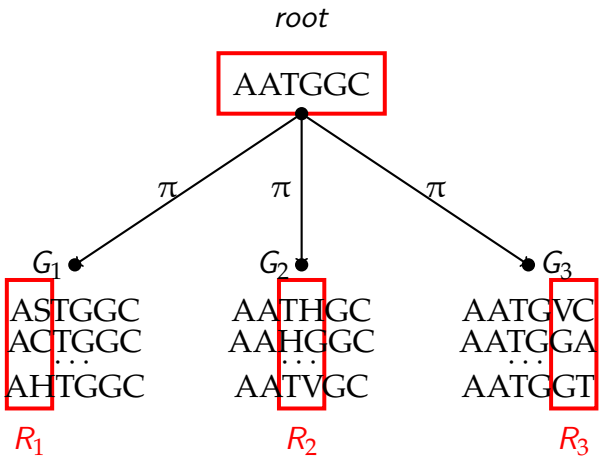
3. Given a mutation rate π, for each branch i, generate n sequences independently in which the amino acid at each position $j \in R_i$ in the root sequence is allowed to mutate independently according to $P(\pi)$.

(II) Rate based generating scheme

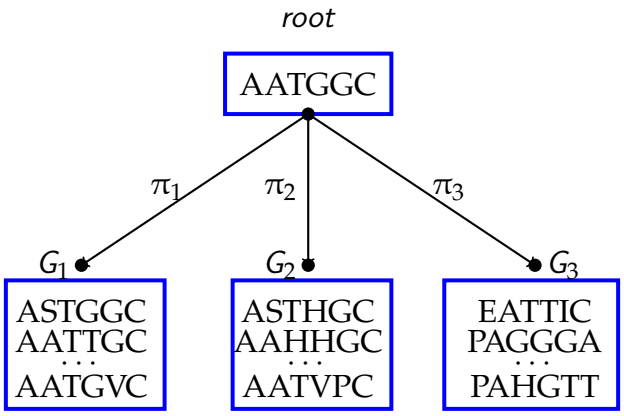
1. Initialize a root sequence of length L, and specify the number of branches g to grow.

2. Specify g different mutation rates $\pi_1, \pi_2, \cdots, \pi_g$. The mutation region is the whole length $\{1, 2, \cdots, L\}$.

3. For each branch i, generate n sequences of length L independently in which the amino acid at every position in the root sequence is allowed to mutate independently according to $P(\pi_i)$.



(a) An example of the region based generating scheme



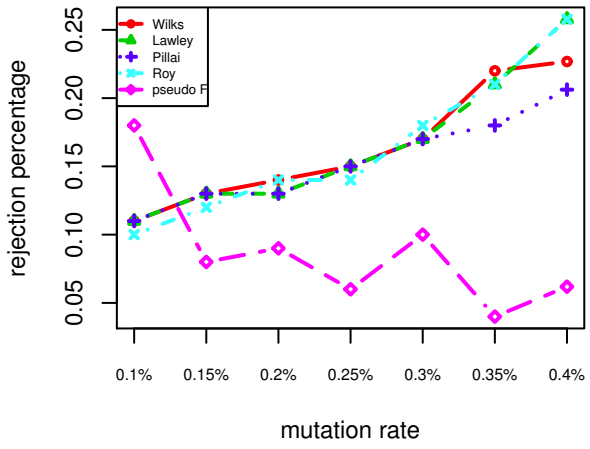
(b) An example of the rate based generating scheme

Figure 2.1: Examples of sequence generating schemes

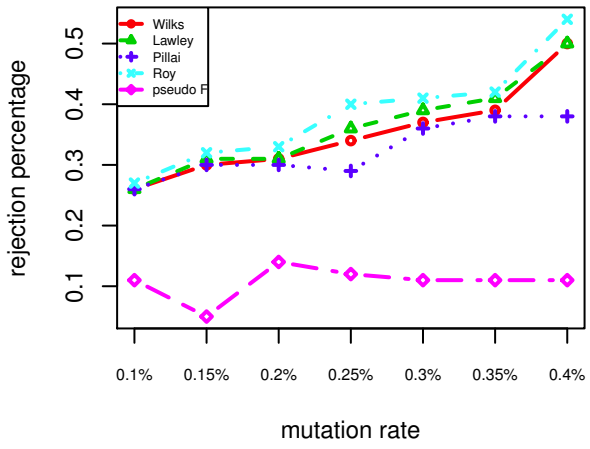
To fully examine the performance of the porposed DANOVA, extensive experiments were performed under various scalings of (L, g, n, π, β), where L is the full length of a sequence, g is the number of branches, n is the number of sequences generated under each branch, π is the mutation rate, and β is the deletion/insertion rate. With $L = 100, g = 3, \beta = 0.00001$ pre-specified, three nonoverlapping regions of length 30 were considered in the first generating scheme. $R_1 = 1:30, R_2 = 35:64, R_3 = 69:98$. To investigate the sensitivity of the proposed DANOVA to small mutations, we intentionally choose low mutation rates ranging from 0.1% to 0.4%. In Table2.2 – 2.3, Figure2.2 – 2.3, rejection percentages under the significance level 5% were displayed based on 100 indepedent experiments in which 5000 permutations were performed. Results given in (Table2.2, Figure2.2) and (Table2.3, Figure 2.3) show the effect of mutation rate π and sample size n respectively on rejection percentages for five test statistics. In all settings, the four MNANOVA-typed test statistics significantly outperform the ANOVA-typed test statistic pseudo F-ratio. This leading advantage can be widened by the increase of dimensions as the mutation rate increases shown in Table2.2 – 2.3. Figure2.2 – 2.3 suggest that the pseudo F-ratio tends to be more conservative as the mutation rate π and sample size n increase compared to the other four test statistics.

Real Data Analysis

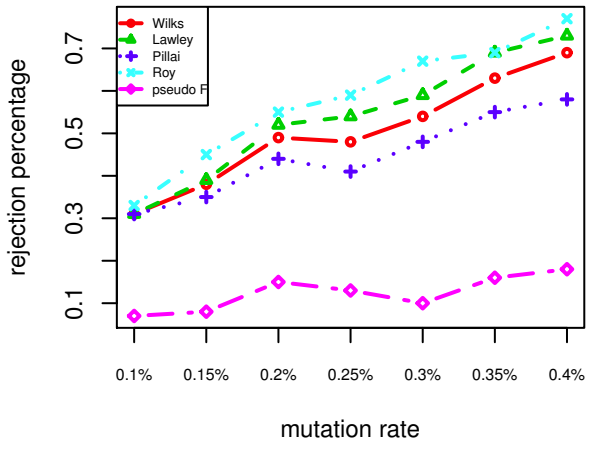
Data description: in the study of preservation of Vpu function conducted by Pickering et. al in [32], 304 unique HIV-1 Vpu amino acid sequences of length 81 were obtained from 14 HIV-1-infected individuals with different progression rates. Infectors were classified according to standard MACS criteria: individuals that progress from seroconversion to AIDS in less than 5 years are designated rapid progressors(RP); 5-10 years as normal progressors(NP); greater than 10 years as long-term non-progressors(LTNP). Among these 14 patients, 5 are LTNPs, 5 are RPs, and 4 are NPs. To fully represent Vpu repertoire, plasma samples from each individual were obtained up to three different time points, ranging from serconversion(0 years) to 10.4 years when possible. A summary on this Vpu data



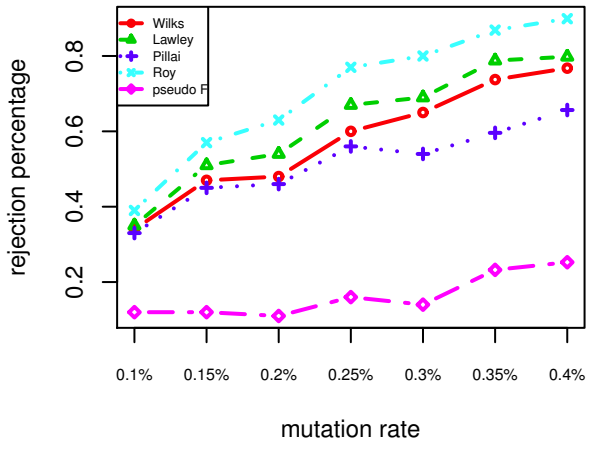
(a) $n = 25$



(b) $n = 50$



(c) $n = 75$



(d) $n = 100$

Figure 2.2: Plots of rejection percentage versus mutation rate π under different sample sizes $n = 25, 50, 75, 100$ respectively given significance level 5%. The results are based on three groups of sequences that are generated under the region based scheme.

n	π	d	Rejection percentage				
			Wilks lambda	Lawley-Hotelling	Pillai-Barlett	Roy's largest root	pseudo F
25	0.1%	9	0.11	0.11	0.11	0.10	0.18
	0.15%	14	0.13	0.13	0.13	0.12	0.08
	0.2%	18	0.14	0.13	0.13	0.14	0.09
	0.2%	18	0.15	0.15	0.15	0.14	0.06
	0.3%	24	0.17	0.17	0.17	0.18	0.1
	0.35%	27	0.22	0.21	0.18	0.21	0.04
	0.4%	30	0.23	0.26	0.21	0.26	0.06
50	0.1%	19	0.356	0.353	0.356	0.344	0.436
	0.15%	27	0.301	0.301	0.300	0.294	0.456
	0.2%	35	0.265	0.264	0.264	0.253	0.473
	0.25%	38	0.234	0.228	0.237	0.208	0.442
	0.3%	48	0.191	0.184	0.200	0.170	0.415
	0.35%	53	0.211	0.196	0.221	0.159	0.402
	0.4%	59	0.50	0.50	0.38	0.54	0.11
75	0.1%	28	0.31	0.31	0.31	0.33	0.07
	0.15%	40	0.38	0.39	0.35	0.45	0.08
	0.2%	52	0.49	0.52	0.44	0.55	0.15
	0.25%	62	0.48	0.54	0.41	0.59	0.13
	0.3%	72	0.54	0.59	0.48	0.67	0.10
	0.35%	81	0.63	0.69	0.55	0.69	0.16
	0.4%	89	0.69	0.73	0.58	0.77	0.18
100	0.1%	37	0.34	0.35	0.33	0.39	0.12
	0.15%	53	0.47	0.51	0.45	0.57	0.12
	0.2%	68	0.48	0.54	0.46	0.63	0.11
	0.25%	82	0.60	0.67	0.56	0.77	0.16
	0.3%	95	0.65	0.69	0.54	0.80	0.14
	0.35%	106	0.74	0.79	0.60	0.87	0.23
	0.4%	117	0.78	0.80	0.66	0.90	0.25

Table 2.2: Rejection percentage table based on three groups of sequences that are generated under the region based scheme. The rejection percentage was obtained by conducting 100 independent experiments in which 5000 permutations are performed.

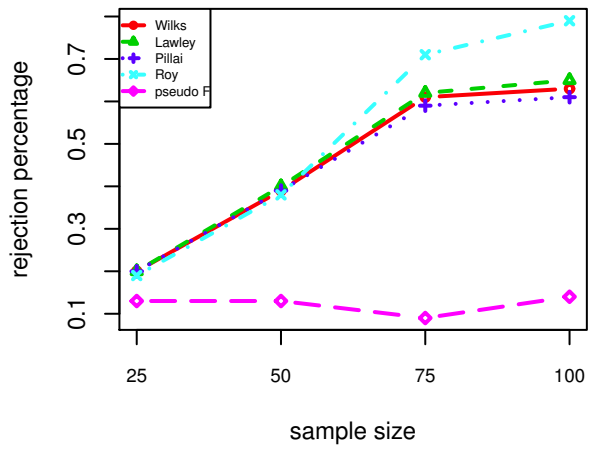


Figure 2.3: A plot of rejection percentage versus sample size n. The result is based on three groups of sequences that are generated with different mutation rates $\pi_1 = 0.1\%, \pi_2 = 0.2\%, \pi_3 = 0.3\%$ respectively.

n	d	Rejection percentage				
		Wilks lambda	Lawley-Hotelling	Pillai-Barlett	Roy's largest root	pseudo F
25	14	0.20	0.20	0.20	0.19	0.13
50	28	0.39	0.40	0.39	0.38	0.13
75	41	0.61	0.62	0.59	0.71	0.09
100	53	0.63	0.65	0.61	0.79	0.14

Table 2.3: Rejection percentage table based on three groups of sequences that are generated with different mutation rates $\pi_1 = 0.1\%, \pi_2 = 0.2\%, \pi_3 = 0.3\%$ respectively.The rejection percentage was obtained by conducting 100 independent experiments in which 5000 permutations are performed.

can be referred to Table2.4. In particular, mean AA changes in Table2.4 measures the average number of amino acid(AA) mutations compared to a given ancestor sequence across all sequences collected from each individual. From Figure2.4, it is easy to observe such measure varies significantly among these 14 HIV-1-infected individuals as well as among three groups, which motivates us to conduct a formal test to examine the significance of these two sources of variation.

Individual	no. Vpu	mean AA changes	Individual	no. Vpu	mean AA changes	Individual	no. Vpu	mean AA changes
LTNP 1	45	12.6	NP 1	22	7.64	RP 1	14	3.07
LTNP 2	15	7.73	NP 2	32	9.38	RP 2	14	7.5
LTNP 3	25	8.4	NP 3	28	8.5	RP 3	22	4.36
LTNP 4	14	8.71	NP 4	8	5.38	RP 4	13	9.54
LTNP 5	42	7.83				RP 5	10	7.5
LTNP	141	9.54	NP	90	8.32	RP	73	6.07

Table 2.4: Vpu amino acid sequence data summary. No. Vpu specifies the total number of sequences obtained from each individual. Mean AA changes measures the average number of amino acid mutations across all sequences collected from each individual. The bottom row summarizes these two measures for each group.

Two-way nested model: since there are three progression rates and infected individuals were randomly chosen, it is natural to adopt a two-way nested mixed model in which the fixed effect is naturally the progrssion rate and the random effect accounts for the individual difference. More specifically, we refer LTNP, NP and RP as the first, second and third group respectively. Let n_{ij} denote the total number of sequences sampled from subject j in group i, and n_i denote the total number of sequences in group i. Let

$$P = \left[\underbrace{p_{1,1,1}, \cdots, p_{1,1,45}}_{P_{11}}, \cdots, p_{1,5,42}, \cdots, p_{2,4,8}, \cdots, \underbrace{p_{3,1,1}, \cdots, p_{3,5,10}}_{P_3} \right]$$

denote the minimal embeddings of all 304 Vpu amino acid sequences in which $p_{i,j,k}$ represents the minimal embedding corresponding to the k-th amino acid sequence from subject j in group i. Let P_{ij} denote all sequences from subject j in group i, and P_i denote all sequences in group i. Thus, the two-way nested model is

$$p_{i,j,k} = \mu + \alpha_i + \beta_{i,j} + \varepsilon_{i,j,k}$$

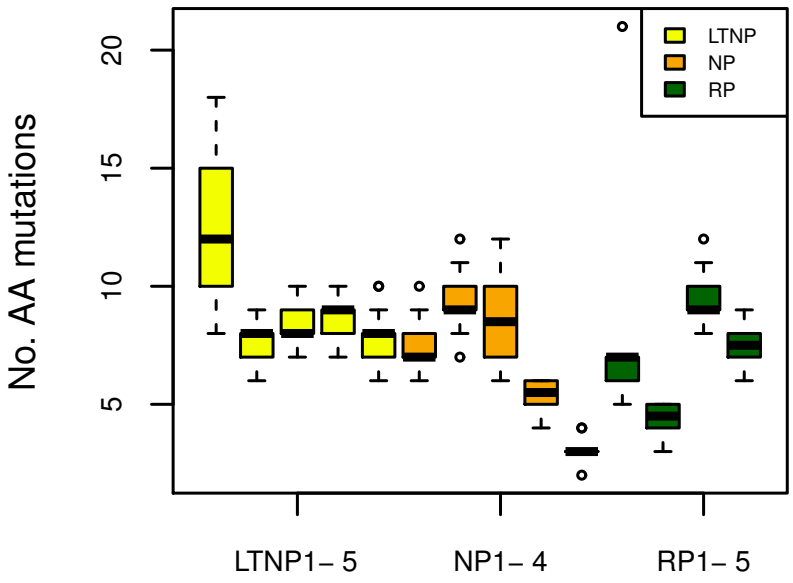


Figure 2.4: 14 boxplots of total number of amino acid mutations. Each represents one HIV-1-infected individual.

where μ can be viewed as the common ancestor or a consensus sequence, α_i represents the mutation for group i and $\beta_{i,j}$ represents the mutation for subject j in group i. Let D denote their corresponding doubly centered Euclidean distance matrix. $|D_i|_1$ denotes the componentwise summation restricted on the submatrix corresponding to sequences from all subjects in group i. Similarly, $|D_{ij}|_1$ denotes the componentwise summation restricted on the submatrix of D corresponding to sequences from subject j in group i. Now we are interested in testing the following

two hypotheses

$$H_{01} : \text{no progression rate effect}$$

$$H_{02} : \text{no individual effect}$$

To thoroughly examine the group effect and the individual effect, we consider in a full spectrum of Wilk's Lambda, Hotelling-Lawley trace, Pillai-Bartlett trace, Roy's largest root and pseudo F-ratio. Table 2.5–2.6 display the formula to calculate these test statistics and the resulting p-values. These five tests agree very well on the significance of the individual effect, while the pseudo F-ratio tends to be conservative regarding the group effect from Figure 2.5.

Source	df	Sum of outer product matrix formula
Group	2	$H_{grp} = \sum_{i=1}^{3} n_i^{-1} P_i \mathbf{1}_{n_i} \mathbf{1}_{n_i}^\top P_i^\top$
Subject	11	$H_{subj} = \sum_{i=1}^{3} \sum_{j=1}^{n_{ij}} n_{ij}^{-1} P_{ij} \mathbf{1}_{n_{ij}} \mathbf{1}_{n_{ij}}^\top P_{ij}^\top - H_{grp}$
Error	290	$E = \Lambda - \sum_{i=1}^{3} \sum_{j=1}^{n_{ij}} n_{ij}^{-1} P_{ij} \mathbf{1}_{n_{ij}} \mathbf{1}_{n_{ij}}^\top P_{ij}^\top$
Total	303	$\Lambda = \text{diag}(\lambda_1, \cdots, \lambda_d)$

Table 2.5: Matrices of sum of outer products in DANOVA for Vpu sequence variation study

Test Name	Source	Test Statistic	P value
Wilks Lambda	Group	$-\log(\det(E)/\det(H_{grp}+E))$	$< 1e-05$
	Subject	$-\log(\det(E)/\det(H_{subj}+E))$	$< 1e-05$
Lawley-Hotelling trace	Group	$\text{trace}(H_{grp}E^{-1})$	$< 1e-05$
	Subject	$\text{trace}(H_{subj}E^{-1})$	$< 1e-05$
Pillai-Bartlett trace	Group	$\text{trace}(H_{grp}(H_{grp}+E)^{-1})$	$< 1e-05$
	Subject	$\text{trace}(H_{subj}(H_{subj}+E)^{-1})$	$< 1e-05$
Roy's largest root	Group	$\lambda_{max}(H_{grp}E^{-1})$	$< 1e-05$
	Subject	$\lambda_{max}(H_{subj}E^{-1})$	$< 1e-05$
pseudo F-ratio	Group	$\frac{290}{2} \frac{-\sum_{i=1}^{3} n_i^{-1} \|(D)_i\|_1}{\sum_{i,j} n_{ij}^{-1} \|D_{ij}\|_1 - \text{trace}(D)}$	0.3935
	Subject	$\frac{290}{11} \frac{\sum_{i=1}^{3} n_i^{-1} \|D_i\|_1 - \sum_{i,j} n_{ij}^{-1} \|D_{ij}\|_1}{\sum_{i,j} n_{ij}^{-1} \|D_{ij}\|_1 - \text{trace}(D)}$	$< 1e-05$

Table 2.6: DANOVA table for Vpu sequence variation study

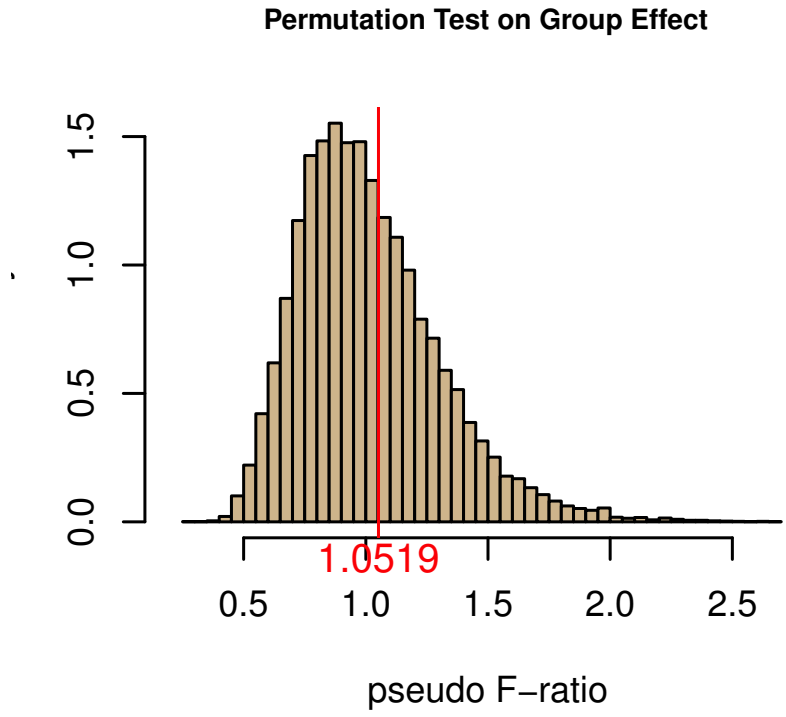


Figure 2.5: Distribution of pseudo F-ratio statistic based on 5000 permutations. The red solid line indicates the observed value.

3 UNSUPERVISED ENSEMBLE LEARNING VIA ISING MODEL APPROXIMATION

3.1 Introduction

In the area of Machine Learning, driven by the persistent quest for a better predictive performance on a given classification problem, numerous efforts have been devoted to propose new techniques to achieve this goal. The most notable development for classification is Ensemble Learning, which stems from the concept that combines multiple existing classifiers in order to outperform any individual constituent one. These base classifiers in the ensemble can be selected from a wide variety of sources, either from human advisers or machine-based algorithms. Many supervised Ensemble Learning methods have been developed so far and they have enjoyed great popularity and success in various applications, such as Random Forest(see [4]), Boosting(see [15, 16]), among others. Despite their long-lasting remarkable performance, all such supervised Ensemble Learning methods must live on the knowledge of true labels. However, nowadays, data often times come without labels due to many reasons, for example, limited budgets, confidential issues, privacy concerns, etc. In order to inherit the advantage of Ensemble Learning, it is necessary to study its counterpart under the unsupervised setting. Moreover, under the supervised setting, in order to make the classification error converge to its asymptotic value,it is necessary to combine a large number of base classifiers, which not only requires large memory but substantially slows down the computation speed. Martinez-Muoz et al. in [28] have investigated several pruning strategies to reduce the size of the ensemble using an idea of ordered aggregation in which the order is determined by the accuracy performance on the training set. Unfortunately, cursed by the lack of a training set, evaluations on the reliability of each base classifier becomes impossible in the unsupervised setting, which arises another interesting yet challenging question, that is, how to discern relevant base classifiers to reduce the size of the original ensemble?

As human become more and more obsessed with complex and difficult problems, unsupervised Ensemble Learning has gained unprecedented explorations and exploitations. Perhaps a good example that could show its increasing popularity and well explain the necessity of pruning would be Crowdsourcing. In a generic Crowdsourcing setting, a series of tasks is completed by soliciting contributions from a large number of people termed as workers. Each worker acts as a base classifier who provides his/her solution regarding to each task. The organizer will later collect all solutions from the crowd to help make decisions. Many Crowdsourcing platforms have been founded online geared toward to different purposes, to name a few, Kaggle serves data mining and forecasting, Challenge.gov helps the U.S. government seek innovative solutions from the public to solve mission-centric problems, and Topcoder allows enterprises to outsource their custom design and development tasks to anonymous coding competitors. Despite many substantial advantages of Crowdsourcing, let it be cost, time savings, or unlimited resources, the collected answers from the crowd can be very noisy since workers are not necessarily commissioned from a specific group or professional organizations. Most are indeed self-volunteered and as such their accuracies are not reliable. In such a scenario, it is essential to recognize experts among the crowd so as to approach to the true answer down the road.

In order to predict true labels on a given data set under the unsupervised setting, the oldest and most straightforward solution is Majority Voting, which in fact requires no additional information but following the rule of majority. Later Dawid and Skene in [9] pioneered a classical method built upon the conditional independence assumption given the true classifier. More specifically, the true classifier is associated with a prior probability parameter and each available classifier in the ensemble is bundled with a pair of probability parameters, representing the probability that the classifier predicts correctly given the truth. All parameters can be estimated based on EM algorithm and both E-step and M-step have simple closed-form solutions. The prediction can be subsequently given by the posterior probability. More recently, Parisi et al. in [31] proposed a spectral method and constructed a meta-learner which can be expressed in a linear form of all avail-

able classifiers. Their method essentially utilizes the rank-one structure on the covariance matrix of the ensemble when the conditional independence is assumed. Along this line, Jaffe et al. in [22] extended the spectral method into the dependence case in which all available classifiers are allowed to be dependent through some unobserved latent variables. As such, the covariance matrix can be represented as a convex combination of two tank-one matrices. However, the dependence structure in this scenario is still very restricted in the sense that no direct dependence structure is allowed among these available classifiers. In addition, the number of latent variables is unknown and numerical experiments in the paper simply rely on the result given by the spectral clustering on the sample covariance matrix. Besides, they did not discuss the behaviour of their proposed algorithm in the high-dimensional setting. To handle the dependence structure, Donmez et al. in [11] utilized the mechanism of hierarchical log-linear models for categorical data(See [3]). In the paper, they considered a second-order log-linear model that captures pairwise interactions between any two available classifiers given the true classifier. However, the parameter space in their model expands quadratically as the size of the ensemble increases. As a consequence, the estimation task is extensive and the resulting prediction performance can not improve much.

In this chapter, inspired by Donmez et al. [11], we focus on the binary classification problem and model the joint distribution of all available classifiers in the ensemble together with the true classifier as an Ising model. As a classical undirected graphical model, Ising model has witnessed rich applications in a variety of domains, including statistical physics [21], natural language processing [27], image analysis [48] [18] [7], and spatial statistics [35], among others. Our main motivation to use Ising model is that it helps explain how individual elements in a community modify their behaviour so as to conform to the behaviour of other individuals in their vicinity. Moreover, the strength of an edge potential can reflect the dependence intensity between two connected nodes. In our analysis, the associated Ising graph has three types of nodes, including a unique hidden node, expert nodes, and non-expert nodes. The hidden node corresponds to the true classifier due to its unsupervised nature. An expert node is defined as a node directly connected to the

hidden node whereas an non-expert node is a node not connected to the hidden node. Given the hidden node, all expert nodes are separable, meaning that their corresponding classifiers are conditionally independent. As no edge is allowed between the hidden node and an non-expert node, such an non-expert classifier provides no additional information regarding to the truth. By identifying all expert nodes and eliminating all non-expert nodes, the original ensemble can be pruned, and the resulting parameter space can be substantially reduced, which ensures the estimation consistency in the later stage. In the subsequent predicting step, the Bayes classifier can be estimated via EM algorithm. Furthermore, we are able to show our proposed method can be well suited in the high-dimensional setting both theoretically and numerically.

We organize this chapter as follows. In Section 3.2, we describe some useful facts on exponential families and in particular Ising model that serve as a fundamental element in later analysis. In Section 3.3, we give a formal formulation to the problem of our interest and break our goal into two steps. The first step is to select expert nodes, also known as the pruning step, and the second step is to build the Bayes classifier on top of these selected ones. In Section 3.4 we narrate in full details a procedure to do neighbourhood selection based on an Ising model approximation. The Bayes classifier estimation would be elaborated in Section 3.5 using the EM algorithm. In Section 3.6, we describe a method of partitioning on the expert set and then introduce an alternative to the Bayes classifier estimation termed as augmented majority vote. In Section 3.7, we discuss some numerical results to show the performance of our proposed two-step unsupervised learning method.

3.2 Exponential Families and Ising Model

Basics of Exponential Families

Many graphical models can be naturally viewed as exponential families, a broad class of distributions that have been extensively studied in the statistics literature. Given a random vector $(X_1, \ldots, X_m)$ taking values in some space $\mathcal{X}^m = \otimes_{s=1}^m \mathcal{X}_s$,

let $\phi := (\phi_\alpha, \alpha \in \mathcal{I})$ be a collection of functions $\phi_\alpha : \mathcal{X}^m \mapsto \mathbb{R}$, known as *potential functions* or *sufficient statistics*, where $\mathcal{I}$ is an index set with $d := |\mathcal{I}|$ elements to be further specified. Thus ϕ can be viewed as a vector-valued mapping from $\mathcal{X}^m$ to $\mathbb{R}^d$. Let $\theta = (\theta_\alpha, \alpha \in \mathcal{I})$ be its associated vector of canonical parameters. The density at a point $(x_1, \ldots, x_m) \in \mathcal{X}^m$ has the exponential form of

$$P_\theta(x_1, \ldots, x_m) = \exp\{\langle \theta, \phi(x) \rangle - A(\theta)\}$$

where $A(\theta) = \log \int_{\mathcal{X}^m} \exp\{\langle \theta, \phi(x) \rangle\} dx$ is known as *log partition function*. The canonical parameters θ of interest belong to the set

$$\Omega := \{\theta \in \mathbb{R}^d \mid A(\theta) < +\infty\}$$

$P_\theta(\cdot)$ is known as a *regular* family if the domain Ω is an open set and it is *minimal* if there is no nonzero vector $\beta \in \mathbb{R}^d$ such that the linear combination $\langle \beta, \phi(x) \rangle = \sum_{\alpha \in \mathcal{I}} \beta_\alpha \phi_\alpha(x)$ is equal to some constant.

Let P be any density(not necessarily in exponential families), given sufficient statistics ϕ, the mean parameters $\mu = (\mu_1, \ldots, \mu_d)^\top$ is defined to be

$$\mu_\alpha = \mathbb{E}_P[\phi_\alpha] = \int_{\mathcal{X}^m} \phi_\alpha(x) P(x) dx, \forall \alpha \in \mathcal{I}$$

Consider the set of mean parameters that can be realized by any distribution P not limited to exponential families:

$$\mathcal{M} := \left\{ \mu \in \mathbb{R}^d \;\middle|\; \exists P \text{ s.t. } \mathbb{E}_P[\phi_\alpha] = \mu_\alpha, \forall \alpha \in \mathcal{I} \right\}$$

Wainwright and Jordan in [47] showed that if an exponential family is regular and minimal,it could enjoy the following two properties:

(a) *moment matching conditions*:

$$\left. \frac{\partial A(\theta)}{\partial \theta} \right|_\theta = \mathbb{E}_{P_\theta}[\phi] \tag{3.1}$$

(b) *forward mapping*:

$$\nabla A : \Omega \mapsto \mathcal{M} \text{ is one-to-one}$$

Ising Model

The *Ising Model* is a classical example of an undirected graphical model in an exponential form. Consider a binary-valued undirected graphical model $G = (V, E)$, with the vertex set $V = \{0, 1, \ldots, p\}$ and the edge set $E \subset V \times V$ being the collection of all pairwise interactions. Each node $s \in V$ is associated with a random variable $X_s \in \{-1, +1\}$. X_s, X_t are allowed to interact directly only if $(s, t) \in E$. Thus, the Ising model setup leads to a density at any realization $(x_0, x_1, \ldots, x_p) \in \{-1, +1\}^{p+1}$ to be

$$P_\theta(x_0, x_1, \ldots, x_p) = \exp\left\{ \sum_{s \in V} \theta_s x_s + \sum_{(s,t) \in E} \theta_{st} x_s x_t - A(\theta) \right\} \tag{3.2}$$

where $\theta_{st} \in \mathbb{R}$ is a potential of edge (s, t), and $\theta_s \in \mathbb{R}$ is a potential for node s including all uncertainty that can not be explained by pairwise interactions with other nodes.

The log partition function $A(\theta)$ is naturally given by the sum

$$\log\left(\sum_{x \in \{-1,+1\}^{p+1}} \exp\left\{ \sum_{s \in V} \theta_s x_s + \sum_{(s,t) \in E} \theta_{st} x_s x_t \right\} \right)$$

Since this sum is finite for all choices of $\theta \in \mathbb{R}^d$, where $d = |V| + |E|$, and the domain Ω is the full space $\mathbb{R}^d$, the family is thus regular. Moreover, it is a minimal representation, in the sense that there is no non-trivial linear combination of all potentials equal to a constant.

3.3 Problem Formulation

In this chapter, we consider the following binary classification problem. Suppose we have p binary classifiers of unknown reliability denoted as $\{f_s\}_{s=1}^{p}$ defined on a common domain $\mathcal{X}$, i.e., $\forall s = 1, \ldots, p$

$$f_s : \mathcal{X} \mapsto \{-1, +1\}$$

with each providing a prediction on a set of i.i.d. instances $D = \{x_i\}_{i=1}^{n} \subset \mathcal{X}^n$. We use $f_s^{(i)}$ to denote the prediction given by f_s on the i-th instance x_i. Let f_0 denote the underlying true classification rule on the domain $\mathcal{X}$. Thus, $(f_0(X), f_1(X), \ldots, f_p(X))^\top$ forms a random vector on $\mathcal{Y} = \{-1, +1\}^{p+1}$, and the joint distribution of such a random vector on $\mathcal{Y}$ takes the form

$$P_{\theta^*}(f_0(x), f_1(x), \ldots, f_p(x)) = \exp\left\{ \theta_0^* f_0(x) + \sum_{0 \leqslant s < t \leqslant p} \theta_{st}^* f_s(x) f_t(x) - A(\theta^*) \right\}, \forall x \in \mathcal{X} \tag{3.3}$$

where $\theta_{st}^* \in \mathbb{R}$ encodes the strength of dependence between f_s and f_t, and $\theta_0^* \in \mathbb{R}$ determines the underlying true classification rule since $P_{\theta^*}(f_0(x) = +1) = \frac{e^{2\theta_0^*}}{1+e^{2\theta_0^*}}, \forall x \in \mathcal{X}$. In the following, we simply write f_s for the random variable $f_s(X)$, where the randomness should be understood implicitly. In the associated graph $G = (V, E)$, $\forall s \in V$, define its neighbourhood:

$$\mathcal{N}_s := \{t \in V : |\theta_{st}^*| > 0\}$$

as well as its degree:

$$d_s := |\mathcal{N}_s|$$

Let $f_{\mathcal{N}_s}$ denote all classifiers corresponding to its neighbourhood $\mathcal{N}_s$. Obviously, $f_s \perp f_{V \setminus \mathcal{N}_s}$ given $f_{\mathcal{N}_s}$. Furthermore, the true classifier f_0 corresponds to the hidden node, whereas each available classifier corresponds to either an expert node with $\theta_{0s}^* \neq 0$ or an non-expert node with $\theta_{0s}^* = 0$.

Without loss of generality, we assume $f_1, \ldots, f_{d_0}$ are directly connected to f_0. and these classifiers are viewed as experts and $\mathcal{N}_0$ is naturally the expert set. Furthermore, the graph G of interest associated with the distribution given in (3.3) has the following properties:

($G1$) (**Identifiability of f_0**) There is a unique hidden node corresponding to f_0 and it is most densely connected in the sense that consider the degree sequence $\{d_{(s)}\}_{s=0}^{p}$ sorted in a descending order, then $d_0 = d_{(0)}$ and $d_0 \geqslant d_{(1)} + 2$.

($G2$) (**Non-informativeness of an non-expert**) Any non-expert node is only allowed to access to the hidden node through at least one expert node in $\mathcal{N}_0$. This indicates that such an non-expert classifier is simply redundant and can be removed from the ensemble.

($G3$) (**Separability of experts**) No edge is allowed between any pair among expert nodes, i.e. $\theta^*_{st} = 0, \forall (s,t) \subseteq \mathcal{N}_0$. This indicates that after removing all non-expert classifiers, the remaining classifiers are conditionally independent given f_0.

Therefore, the underlying joint distribution given in (3.3) can be further decomposed into:

$$P_{\theta^*}(f_0, f_1, \ldots, f_p) \propto \exp\left\{\theta^*_0 + \sum_{s=1}^{d_0} \theta^*_{0s} f_0 f_s\right\} \exp\left\{\sum_{1 \leqslant s < t \leqslant p} \theta^*_{st} f_s f_t\right\} \tag{3.4}$$

Since the prediction performance of each individual available classifier on the domain $\mathcal{X}$ is unknown, there are two natural questions to ask:

(1) Among all available classifiers, who are experts?

(2) How to rely solely on experts in the ensemble to deliver a more accurate prediction rule?

To answer the first question, one should see that it essentially amounts to estimate $\mathcal{N}_0$, in other words, $\forall s = 1, \cdots, p$, test if $\theta^*_{0s} = 0$. With regards to the second question,

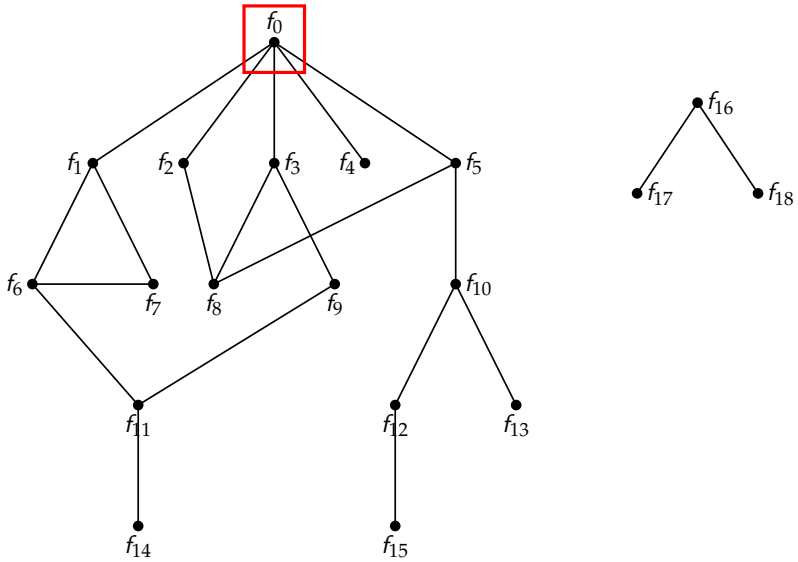


Figure 3.1: An illustrative example of a valid graph structure satisfying properties$(G1)$–$(G3)$. f_0 in the red square is unknown. The expert set $\mathcal{N}_0 = \{1, 2, 3, 4, 5\}$. All the rest are non-expert nodes.

it is in fact tightly related to the first one in the sense that the Bayes classifier, which minimizes the 0-1 misclassification rate among all possible classifiers, is fully dependent on those experts. More explicitly, the Bayes classifier denoted as f_B is

$$f_B(x) = sgn\left\{\sum_{s\in\mathcal{N}_0} \theta^*_{0s} f_s(x) + \theta^*_0\right\}, \forall x \in \mathcal{X} \tag{3.5}$$

To solve the aforementioned two tasks, we propose the following two-step procedure:

(1) Pruning step – estimate the expert set denoted as $\hat{\mathcal{N}}_0$.

(2) Predicting step – estimate $\theta^*_{0s}, \forall s \in \hat{\mathcal{N}}_0$ denoted as $\hat{\theta}^*_{0s}$, and plug in (3.5) to estimate the Bayes classifier denoted as $\hat{f}_B$.

3.4 Neighbourhood Selection

In this section, we will elaborate in full details the pruning step to estimate $\mathcal{N}_0$ using ℓ_1-regularized Logistic regression based on the Ising model approximation.

Ising Model Approximation

For any node $s \in V$, let $f_{\setminus s} := f_{V\setminus\{0,s\}}$. $\forall s \in V\setminus\{0, \mathcal{N}_0\}$,

$$\frac{P_{\theta^*}(f_s = +1|f_{\setminus s})}{P_{\theta^*}(f_s = -1|f_{\setminus s})} = \exp\left\{2\sum_{r\in\mathcal{N}_s}\theta^*_{rs}f_r\right\} \tag{3.6}$$

which indicates that the conditional distribution of f_s given all other available classifiers follows a logistic model. With this observation, we can estimate $\mathcal{N}_s$ by performing an ℓ_1-regularized logistic regression of f_s on the other variables $f_{\setminus s}$ proposed by Ravikumar et al. in [34]. In fact, the nice logistic form given in (3.6) is essentially blessed by the Ising model structure imposed on the underlying joint distribution

$$P_{\theta^*}(f_s, f_{\mathcal{N}_s}) \propto \exp\{\sum_{r\in\mathcal{N}_s}\theta^*_{rs}f_r f_s\}$$

However, for any node $s \in \mathcal{N}_0$, its conditional distribution $P_{\theta^*}(f_s|f_{\setminus s})$ no longer follows a logistic model:

$$\frac{P_{\theta^*}(f_s = +1|f_{\setminus s})}{P_{\theta^*}(f_s = -1|f_{\setminus s})} = \exp\left\{2\sum_{r=d_0+1}^{p}\theta^*_{rs}f_r\right\}\frac{e^{\theta^*_{0s}+L^*_s} + e^{-\theta^*_{0s}-L^*_s}}{e^{-\theta^*_{0s}+L^*_s} + e^{\theta^*_{0s}-L^*_s}} \tag{3.7}$$

where $L^*_s := \theta^*_0 + \sum_{\substack{t\in\mathcal{N}_0\\ t\neq s}}\theta^*_{0t}f_t$.

Despite the non-eligibility of a logistic form in (3.7), Lemma 3.1 can show there always exists a unique Ising model that can best approximates the true distribution $P_{\theta^*}(f_{\mathcal{N}_0\cup\mathcal{N}_s\setminus 0})$ in terms of their Kullback-Leibler divergence. In light of Lemma 3.1, it is straightforward to see there is a unique Ising model than can best approximate

the marginal distribution of $f_1, \ldots, f_p$, which will be presented in Theorem 3.2.

Lemma 3.1. *For each node s from $\mathcal{N}_0$, there always exists a unique Ising model $Q^s_{\tilde{\theta}}(\cdot)$ that can best approximates the true distribution $P_{\theta^*}(f_{\mathcal{N}_0 \cup \mathcal{N}_s \setminus 0})$ in terms of their Kullback-Leibler divergence. Furthermore, $Q^s_{\tilde{\theta}}(\cdot)$ takes the form*

$$\exp\left\{ \sum_{\substack{r,t \in \mathcal{N}_0 \cup \mathcal{N}_s \setminus 0 \\ r \neq t}} \tilde{\theta}_{rt} f_r f_t - A(\tilde{\theta}) \right\}$$

in which

$$\tilde{\theta}_{rt} = \theta^*_{rt}, \ \forall (r,t) \nsubseteq \mathcal{N}_0$$

$$\tilde{\theta}_{rt} = \frac{1}{2} \log \left(\frac{e^{\theta^*_{0r}+\theta^*_{0t}+\theta^*_0} + e^{-\theta^*_{0r}-\theta^*_{0t}-\theta^*_0} + e^{-\theta^*_{0r}-\theta^*_{0t}+\theta^*_0} + e^{\theta^*_{0r}+\theta^*_{0t}-\theta^*_0}}{e^{\theta^*_{0r}-\theta^*_{0t}+\theta^*_0} + e^{-\theta^*_{0r}+\theta^*_{0t}-\theta^*_0} + e^{-\theta^*_{0r}+\theta^*_{0t}+\theta^*_0} + e^{\theta^*_{0r}-\theta^*_{0t}-\theta^*_0}} \right), \ \forall (r,t) \subseteq \mathcal{N}_0$$

Theorem 3.2. *Suppose a random vector $(f_0, f_1, \ldots, f_p)^\top \in \mathcal{Y}$ follows the distribution given in (3.3), then there exists a unique Ising model $Q_{\tilde{\theta}}(\cdot)$ that best approximates $P_{\theta^*}(f_1, \ldots, f_p)$ marginalizing over f_0 in terms of their Kullback-Leibler divergence. Furthermore,*

$$Q_{\tilde{\theta}}(f_1, \ldots, f_p) = \exp\left\{ \sum_{1 \leqslant s < t \leqslant p} \tilde{\theta}_{st} f_s f_t - A(\tilde{\theta}) \right\} \tag{3.8}$$

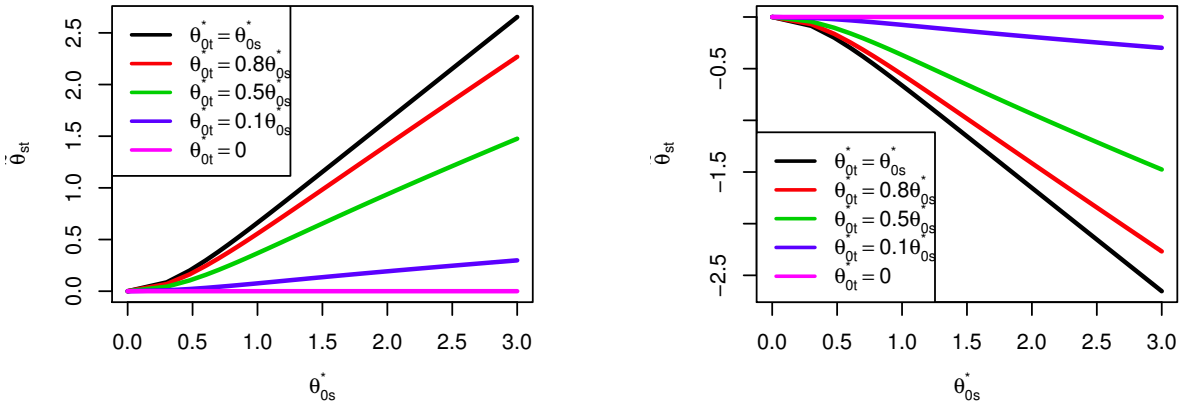
in which

$$\tilde{\theta}_{st} = \begin{cases} \theta^*_{st} & \text{if} \quad \{s,t\} \nsubseteq \mathcal{N}_0 \\ \frac{1}{2} \log \left(\frac{e^{\theta^*_{0s}+\theta^*_{0t}+\theta^*_0} + e^{-\theta^*_{0s}-\theta^*_{0t}-\theta^*_0} + e^{-\theta^*_{0s}-\theta^*_{0t}+\theta^*_0} + e^{\theta^*_{0s}+\theta^*_{0t}-\theta^*_0}}{e^{\theta^*_{0s}-\theta^*_{0t}+\theta^*_0} + e^{-\theta^*_{0s}+\theta^*_{0t}-\theta^*_0} + e^{-\theta^*_{0s}+\theta^*_{0t}+\theta^*_0} + e^{\theta^*_{0s}-\theta^*_{0t}-\theta^*_0}} \right) & \text{if} \quad \{s,t\} \subseteq \mathcal{N}_0 \end{cases}$$

Proposition 3.3. *Consider an Ising model approximation $Q_{\tilde{\theta}}(\cdot)$ given in Theorem 3.2, $\forall \{s,t\} \subseteq \mathcal{N}_0$, $\tilde{\theta}_{st} \neq 0$. Furthermore, $\tilde{\theta}_{st} > 0$ if and only if $\theta^*_{0s}\theta^*_{0t} > 0$ and $\tilde{\theta}_{st} < 0$ otherwise.*

In light of Proposition 3.3, for each node $s = 1, \ldots, p$, it is natural to introduce a neighbourhood $\tilde{\mathcal{N}}_s$ with respect to $Q_{\tilde{\theta}}(\cdot)$. Clearly, for all non-expert nodes, their

neighbourhoods remain the same, i.e., $\tilde{\mathcal{N}}_s = \mathcal{N}_s$. $\forall s \in \mathcal{N}_0, \tilde{\mathcal{N}}_s = \mathcal{N}_s \cup \mathcal{N}_0 \backslash 0$, which includes not only its original neighbourhood $\mathcal{N}_s$ but its siblings in $\mathcal{N}_0$ as well. Figure 3.2 shows how the value of $\tilde{\theta}_{st}$ varies as the strengths in two edges change. Intuitively, the more signals appear on both edges simultaneously, the more signal



(a) same (b) opposite

Figure 3.2: Two examples of the value of $\tilde{\theta}_{st}$ as a function of $(\theta^*_{0s}, \theta^*_{0t}, \theta^*_0)$ given in Theorem 3.2. The left panel corresponds to $\theta^*_{0s}\theta^*_{0t} > 0$ and the right corresponds to $\theta^*_{0s}\theta^*_{0t} < 0$. In both scenarios, set $\theta^*_0 = 0$.

$\tilde{\theta}_{st}$ can be able to preserve. And the sign of $\tilde{\theta}_{st}$ can reflect whether the two involved edge potentials have the same sign. For example, in the graph given in Figure 3.1, $\tilde{\mathcal{N}}_1 = \{2, 3, 4, 5\}$. And the graph w.r.t. its Ising approximation $Q_{\tilde{\theta}}(\cdot)$ can be referred to Figure 3.3.

Therefore, an immediate consequence from Theorem 3.2 is, $\forall s = 1, \ldots, p$, the best logistic model that can approximate its conditional distribution $P_{\theta^*}(f_s|f_{\backslash s})$ can be written in the form

$$Q_{\tilde{\theta}}(f_s|f_{\backslash s}) = \frac{\exp\{2f_s \sum_{t\in\tilde{\mathcal{N}}_s} \tilde{\theta}_{st} f_t\}}{\exp\{2f_s \sum_{t\in\tilde{\mathcal{N}}_s} \tilde{\theta}_{st} f_t\} + 1} \tag{3.9}$$

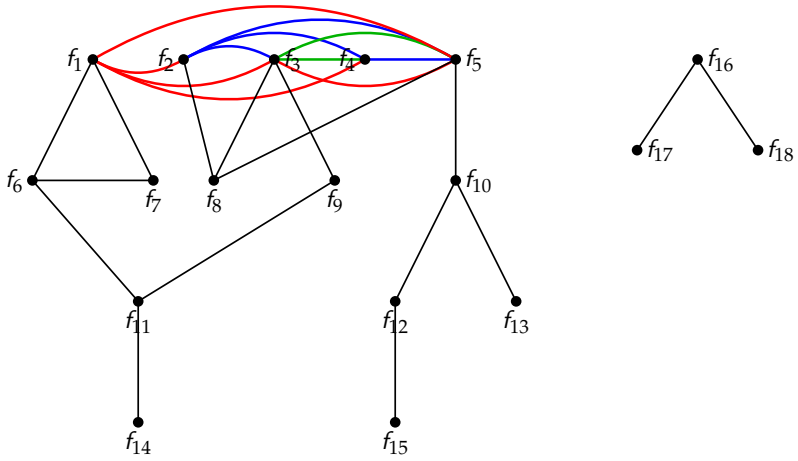


Figure 3.3: The Ising approximation graph structure given in Figure 3.1. All expert nodes are fully connected.

Nodewise Logistic Regression

Now that Theorem 3.2 showed that there is a unique Ising model that best approximates the marginal distribution of all p available classifiers, for each node $s \in V \backslash \{0\}$, we can perform ℓ_1-regularized logistic regression to select its neighbourhood. The ℓ_1-regularized regression is of the form

$$\min_{\theta \in \mathbb{R}^{p-1}} \left\{ -\frac{1}{n} \sum_{i=1}^{n} \log Q_\theta(f_s^{(i)} | f_{\backslash s}^{(i)}) + \lambda_n \|\theta\|_1 \right\}$$

which amounts to

$$\min_{\theta \in \mathbb{R}^{p-1}} \left\{ -\frac{1}{n} \sum_{i=1}^{n} \log \left\{ \exp\left(\sum_{t \neq s} \theta_{st} f_t^{(i)} \right) + \exp\left(-\sum_{t \neq s} \theta_{st} f_t^{(i)} \right) \right\} - \sum_{t \neq s} \theta_{st} \hat{\mu}_{st} + \lambda_n \|\theta\|_1 \right\} \tag{3.10}$$

where $\hat{\mu}_{st} = \frac{1}{n} \sum_{i=1}^{n} f_s^{(i)} f_t^{(i)}$. Note that the objective function (3.10) is convex but not differentiable, due to the ℓ_1-penalty term. However, Ravikumar et al in [34] showed that the minimizer $\hat{\theta}$ is always achievable and unique under the regime of interest. Now we can use the edge potential estimate $\hat{\theta}$ to reconstruct the corresponding

neighbourhood

$$\hat{\mathcal{N}}_s := \{t \in \{1, 2, \cdots, p\} \backslash s \mid |\hat{\theta}_{st}| > 0\}$$

Neighbourhood Selection Correctness

In this section, we mainly discuss the correctness of $\hat{\mathcal{N}}_s$ by using ℓ_1-regularized logistic regression of f_s on all other available classifiers. To establish some theoretical results, we start off by introducing some necessary assumptions, most of which are imposed on the Ising model approximation $Q_{\tilde{\theta}}(\cdot)$.

Assumptions: $\forall s = 1, \ldots, p$, consider its Fisher information matrix w.r.t the Ising model approximation $Q_{\tilde{\theta}}(\cdot)$

$$I_s := -\mathbb{E}_{\tilde{\theta}}\left[\nabla^2 \log Q_{\tilde{\theta}}(f_s | f_{\backslash s})\right] = \mathbb{E}_{\tilde{\theta}}\left[h_{\tilde{\theta}}(f) f_{\backslash s} f_{\backslash s}^\top\right]$$

where $h_{\tilde{\theta}}(f) = \frac{4\exp(2f_s \sum_{t \neq s} \tilde{\theta}_{st} f_t)}{(\exp(2f_s \sum_{t \neq s} \tilde{\theta}_{st} f_t)+1)^2}$. For notation convenience, we temporarily drop the subscript denoting the node index and use $\mathcal{N}$ to stand for the corresponding neighbourhood and denote by $\mathcal{N}^c$ its complement. The following assumptions hold for each $s = 1, \ldots, p$.

($A1$) There exists a pair of positive constants (C_{min}, D_{max}) such that the Fisher information matrix I restricted on its neighbourhood $\mathcal{N}$ denoted as $I_{\mathcal{N}\mathcal{N}}$ satisfies

$$\lambda_{min}(I_{\mathcal{N}\mathcal{N}}) \geqslant C_{min}$$

and

$$\lambda_{max}(\mathbb{E}_{\tilde{\theta}}[f_{\backslash s} f_{\backslash s}^\top]) \leqslant D_{max}$$

($A2$) There exists an $\alpha \in (0, 1]$ such that $\|I_{\mathcal{N}^c\mathcal{N}}(I_{\mathcal{N}\mathcal{N}})^{-1}\|_\infty \leqslant 1 - \alpha$.

We further denote the edge set w.r.t. $Q_{\tilde{\theta}}$ by $\tilde{E}$ and let $\tilde{d}_s$ denote the degree for each node s. Clearly, $\tilde{d}_s = d_s + d_0 - 1, \forall s \in \mathcal{N}_0$ and $\tilde{d}_s = d_s$ for the rest. And let $d_{max} := \max_{s \in \{1 \ldots p\}} \tilde{d}_s$, and $\tilde{\theta}_{min} := \min_{(s,t) \in \tilde{E}} |\tilde{\theta}_{st}|$

Theorem 3.4. *Consider an Ising model $Q_{\tilde{\theta}}$ given in (3.8) such that $(A1)$ and $(A2)$ are satisfied, let the regularization parameter $\lambda_n \geqslant \frac{16\alpha}{1-\alpha}\sqrt{\frac{\log p}{n}}$, $n > Ld_{max}^3 \log p$ for some positive constants L and K, independent of (n, p, d_{max}), and $\tilde{\theta}_{min} = \Omega\left(\sqrt{\frac{d_{max}\log p}{n}}\right)$, with probability at least $1 - 2\exp(-K\lambda_n^2 n)$,*

$$\forall s \in \mathcal{N}_0,\ \tilde{\mathcal{N}}_s \text{ can be correctly recovered}$$

$$\forall s \in V\backslash\{0, \mathcal{N}_0\},\ \mathcal{N}_s \text{ can be correctly recovered}$$

In addition, $\forall (s,t) \in \tilde{E}$, the sign of $\tilde{\theta}_{st}$ can be correctly recovered.

Theorem 3.4 is a direct result following the proof given in Ravikumar et al in [34]. More details can refer to [34].

Reconstruction of $\mathcal{N}_0$

Now that all neighbourhoods $\hat{\mathcal{N}}_s, \forall s = 1, \ldots, p$ have been recovered, a natural question worth asking is how to use these neighbourhoods to reconstruct $\mathcal{N}_0$.

Theorem 3.5. *Consider a graph G satisfying the aforementioned properties $(G1)$–$(G3)$, and $\{\tilde{\mathcal{N}}_s\}_{s=1}^p$ denote a sequence of neighbourhoods w.r.t $Q_{\tilde{\theta}}$. Furthermore, a node s is said to be a knot if s is the single intersection of all its neighbour's neighbourhoods, that is, $A_s = s$, where $A_s := \cap_{r\in\tilde{\mathcal{N}}_s}\tilde{\mathcal{N}}_r, s = 1, \ldots, p$. Consider the collection of such knots denoted by $\mathcal{A}$, in which each knot s is also associated with an index i_s storing the position of s in the ordered sequence $|\tilde{\mathcal{N}}_{(1)}| \geqslant |\tilde{\mathcal{N}}_{(2)}| \geqslant \ldots \geqslant |\tilde{\mathcal{N}}_{(|\mathcal{A}|)}|$, then*

$$\mathcal{N}_0 = \{s \in \mathcal{A} : |\tilde{\mathcal{N}}_s| \geqslant i_s - 1\}$$

Theorem 3.5 immediately suggests a procedure to recover $\mathcal{N}_0$ which has been detailed in Algorithm 3.

Data: $\{(f_1^{(i)}, \cdots, f_p^{(i)})^\top\}_{i=1}^n$
Result: $\hat{\mathcal{N}}_0$
for $s = 1 \dots p$ do

- Perform ℓ_1-regularized regression given in (3.10).
- Output its estimated neighbourhood $\hat{\mathcal{N}}_s$

Initialize $\mathcal{A}, \hat{\mathcal{N}}_0 = \emptyset$
for $s = 1 \dots p$:

- $A_s := \cap_{r \in \hat{\mathcal{N}}_s} \hat{\mathcal{N}}_r$
- If $A_s = s$, then $\mathcal{A} \leftarrow \mathcal{A} \cup s$

Sort the sequence $\{|\hat{\mathcal{N}}_s|\}_{s \in \mathcal{A}}$ in a descending order and associate each $s \in \mathcal{A}$ with an index i_s defined in Theorem 3.5, then

$$\hat{\mathcal{N}}_0 = \{s \in \mathcal{A} : |\hat{\mathcal{N}}_s| \geqslant i_s - 1\}$$

return $\hat{\mathcal{N}}_0$

Algorithm 3: Pruning step – Reconstruction of $\mathcal{N}_0$

3.5 Bayes Classifier Estimation

Once we figure out $\mathcal{N}_0$, we only need to keep all experts in the original ensemble and as such the graph can be reduced into a tree structure with a single hidden node f_0 along with all expert nodes $f_{\mathcal{N}_0}$ shown in Figure 3.4.

As we mentioned in the beginning of this paper, our goal is to deliver a more accurate classifier based on these p existing ones. The Bayes classifier, among all possible classifiers, is best known for its ability to minimize the 0-1 misclassification

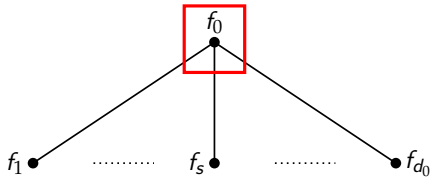


Figure 3.4: The reduced graph structure

rate. And under our formulation,

$$f_B(\cdot) = sgn\left\{\sum_{s\in\mathcal{N}_0}\theta^*_{0s}f_s(\cdot) + \theta^*_0\right\}$$

In this section, we discuss a method to get a consistent estimate of $\theta^*_{0s}, s \in \mathcal{N}_0$. Let $\psi_s := P_{\theta^*}(f_s f_0 = +1), s \in \mathcal{N}_0$ and $\pi := P_{\theta^*}(f_0 = +1)$.

According to the joint probability given in (3.3), some algebraic manipulations reveal the relationship between $(\theta^*_{0s}, \psi_s), s \in \mathcal{N}_0$:

$$\theta^*_{0s} = \frac{1}{2}\log\frac{\psi_s}{1-\psi_s} \tag{3.11}$$

Moreover,

$$\theta^*_0 = \frac{1}{2}\log\frac{\pi}{1-\pi} \tag{3.12}$$

Denote the hidden random variable indicating the underlying label for each instance x_i by $Z^{(i)} = \mathbb{I}(f_0^{(i)} = +1), i = 1, \ldots, n$, the complete data likelihood $L(\Theta)$ is

$$\prod_{i=1}^{n}\left\{Z^{(i)}\pi\prod_{s\in\mathcal{N}_0}\psi_j^{\frac{f_s^{(i)}+1}{2}}(1-\psi_s)^{\frac{1-f_s^{(i)}}{2}} + (1-Z^{(i)})(1-\pi)\prod_{s\in\mathcal{N}_0}(1-\psi_s)^{\frac{f_s^{(i)}+1}{2}}\psi_s^{\frac{1-f_s^{(i)}}{2}}\right\} \tag{3.13}$$

The maximization on the log-likelihood $\log L(\Theta)$, in general, turns out to be very difficult due to its non-concavity. However, the EM algorithm introduced in [10] by Dempster et al.(1977) can be used in this context and it essentially maximizes alternatively in the coordinate ascent fashion over the lower bound $Q(\Theta)$ defined

as:

$$Q(\Theta) = \sum_{i=1}^{n} \sum_{s \in \mathcal{N}_0} \left\{ Z^{(i)} \log \left[\psi_s^{\frac{f_s^{(i)}+1}{2}} (1-\psi_s)^{\frac{1-f_s^{(i)}}{2}} \right] + (1-Z^{(i)}) \log \left[(1-\psi_s)^{\frac{f_s^{(i)}+1}{2}} \psi_s^{\frac{1-f_s^{(i)}}{2}} \right] \right\}$$
$$+ \sum_{i=1}^{n} \left(Z^{(i)} \log \pi + (1-Z^{(i)}) \log(1-\pi) \right) \tag{3.14}$$

Define the posterior probability of being +1 as

$$\tau_i := P(f_0^{(i)} = +1 | f_{\mathcal{N}_0}^{(i)}) = \frac{\exp\left\{2\left(\theta_0^* + \sum_{s \in \mathcal{N}_0} \theta_{0s}^* f_s^{(i)}\right)\right\}}{1 + \exp\left\{2\left(\theta_0^* + \sum_{s \in \mathcal{N}_0} \theta_{0s}^* f_s^{(i)}\right)\right\}} \tag{3.15}$$

In E-step, each $Z^{(i)}$ is updated by its expectation τ_i. In the subsequent M-step, the MLE of ψ_s and π are given by

$$\begin{aligned} \hat{\psi}_s &= \frac{1}{2} + \frac{1}{n} \sum_{i=1}^{n} (\tau_i - \frac{1}{2}) f_s^{(i)} \\ \hat{\pi} &= \frac{\sum_{j=1}^{n} \tau_i}{n} \end{aligned} \tag{3.16}$$

Based on (3.11) and (3.12), we can get the estimation on θ_{0s}^* and θ_0^*. The whole procedure is summarized in Algorithm 4.

Data: $\{(f_{\mathcal{N}_0}^{(i)})^\top\}_{i=1}^n$

Result: $\{\hat{f}_B^{(i)}\}_{i=1}^n$

Initialization: $\theta_0^*, \theta_{0s}^*, s \in \mathcal{N}_0$ **repeat**

E Step –

- update τ_i by (3.15)

M Step –

- $\pi \leftarrow \frac{\sum_{j=1}^n \tau_i}{n}$
- $\hat{\psi}_s \leftarrow \frac{1}{2} + \frac{1}{n}\sum_{i=1}^n (\tau_i - \frac{1}{2}) f_s^{(i)}$
- update $\hat{\theta}_{0s}$ by (3.11) and $\hat{\theta}_0$ by (3.12)

until (3.14) converges;

for $i = 1 : n$

$$\hat{f}_B^{(i)} \leftarrow sgn\left(\hat{\theta}_0 + \sum_{s\in\mathcal{N}_0} \hat{\theta}_{0s} f_s^{(i)}\right)$$

return $\{\hat{f}_B^{(i)}\}_{i=1}^n$

Algorithm 4: Predicting step – Bayes classifier estimation

3.6 $\mathcal{N}_0$ Partitioning and Augmented Majority Vote

$\mathcal{N}_0$ **partitioning**: so far, identification of an expert node is solely determined by if the corresponding edge potential θ_{0s}^* is 0. In fact, for each expert node $s \in \mathcal{N}_0$, the sign of θ_{0s}^* can further reveal whether node s is a positive expert or a negative expert, in which positive/negative is measured by $P_{\theta^*}(f_s f_0 = +1) = \frac{e^{2\theta_{0s}^*}}{1+e^{2\theta_{0s}^*}}$ is strictly greater/less than 0.5. Naturally, a positive expert means its action shows more compliance with the truth whereas a negative expert tends to make the opposite movement. Once these two groups of experts can be differentiated, the negative group can be turned over to the positive side by reversing their labelings. Unfortunately, it is impossible to determine the tone of such an expert without any

access to the true labels. However, the good news is we can still partition the expert set $\mathcal{N}_0$ into two groups and either one of them corresponds to the positive group. In fact, Proposition 3.3 has shed light on the partitioning principle, that is, any two expert nodes with a positive $\tilde{\theta}_{st}$ would go to the same group and vice versa.
Augmented majority vote: as an alternative to estimate Bayes classifier, once the partitioning on $\mathcal{N}_0$ is complete, the group with a smaller size would be marked as the negative group and all labels within this group would be flipped. Then the majority vote policy can be executed on this twisted data set. However, one caveat is that the success of this method relies heavily on the assumption that more positive experts are available compared to their negative peers in the ensemble.

3.7 Numerical Experiments

Simulations

In this section, we describe some experimental results to evaluate the performance of our proposed two-step unsupervised ensemble learning method, which includes the pruning step given in Algorithm 3 and the subsequent predicting step shown in Algorithm 4.

More specifically, we considered a series of Ising models in which $\theta^*_{0s} = \pm 1, \forall s \in \mathcal{N}_0$ with probability 0.7/0.3 such that positive experts form the major party in $\mathcal{N}_0$, and $\theta^*_0 = 0$ such that $P_{\theta^*}(f_0(x) = +1) = 0.5, \forall x \in \mathcal{X}$. Moreover, by defining the signal-to-noise ratio(SNR) in this context as $|\theta^*_{0s}/\theta^*_{st}|, \forall s \in \mathcal{N}_0, t \in \mathcal{N}_s$, we considered three levels of signal-to-noise ratio: high, medium, low, in which $\theta^*_{st} = \pm 0.25, \pm 0.5, \pm 1$ with equal probability respectively. To fully investigate the performance of our method, experiments were performed under various scaling scenarios of (n, p, d_0). In particular, $p \in \{25, 36, 49, 64, 81, 100\}$, and $d_0 \in \{\log p, \sqrt{p}, p/4\}$. We set $n = 30 d_{max} \log p$ and the regularization parameter $\lambda_n = \sqrt{\frac{\log p}{n}}$. Given the distribution of the form in (3.3), we generated random data by Gibbs sampling. To improve the data quality, we collected every $2(p+1)$th sample after the first 1000 iterations. For each specific combination of (n, p, d_0), 200 independent trials were

performed and all results were averaged over these trials. To examine the performance of our proposed pruning step, two commonly used metrics ware evaluated, which are *Hit Rate* and *Precision*. *Hit Rate* is defined as the portion among d_0 elements in $\mathcal{N}_0$ that have been successfully recovered, whereas *Precision* corresponds to the portion among all selected nodes that truly lies in the expert set $\mathcal{N}_0$.

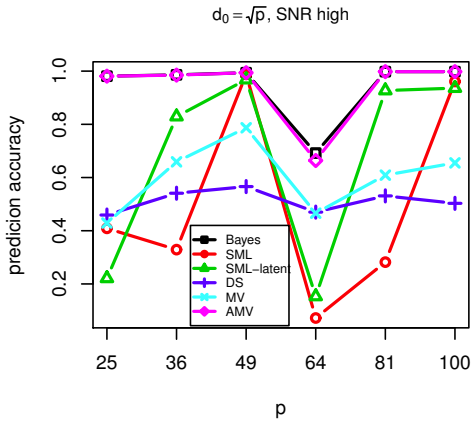
$$\textit{Hit Rate} := \frac{|\hat{\mathcal{N}}_0 \cap \mathcal{N}_0|}{|\mathcal{N}_0|}, \quad \textit{Precision} := \frac{|\hat{\mathcal{N}}_0 \cap \mathcal{N}_0|}{|\hat{\mathcal{N}}_0|} \tag{3.17}$$

To examine the subsequent prediction performance, our method was compared to several other widely used methods, including the majority vote, the classical Dawid-Skene estimator, and more recent spectral meta-learner(SML) proposed by Parisi et al. in [31] as well as its extension to the dependence case(SML-Latent) proposed by Jaffe et al. in [22]. The majority vote method serves as a baseline, while other methods are used to assess the efficacy of our method to handle the dependence case. Since both our predicting step and Dawid-Skene estimator are built upon the EM algorithm, to make the comparison fair, we took very careful initializations. For our method, after partitioning the expert set into two groups, we initialized all classifiers partitioned in the larger group with $\theta^*_{0s} = 1$, whereas all classifiers in the other group were assigned with $\theta^*_{0s} = -1$. In Dawid-Skene scenario, we randomly chose d_0 classifiers and initialized their corresponding $\theta^*_{0s} = \pm 1$. In both cases, initialize all the rest $\theta^*_{0s} = 0$. All results were presented in Table 3.1–Table 3.3. In particular, Figure 3.5 displays the predictive performance comparisons among all aforementioned methods under the scenario $d_0 = \sqrt{p}$ for all three levels of signal-to-noise ratio. From Figure 3.5, our predicting step including the Bayes classier estimation and the augmented majority vote wins over other methods and more noises can see more advantage. Based on Table 3.1–Table 3.3, as d_0 grows from being logarithmic to linear in p, both our pruning step and predicting step exhibit a consistently high performance despite the fact that there are occasional mediocre performances under the high noise level and d_0 is linear in p. In fact, the high performance of our pruning step plays a irreplaceable role in escorting the subsequent predicting step to the success. However, with no prior knowledge on

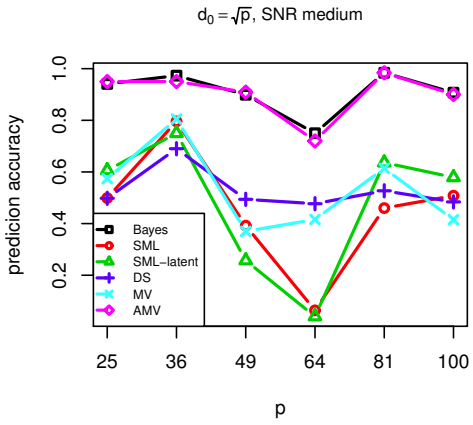
which subset of classifiers should be selected, the Dawid-Skene method is prone to failure due to uncontrollable randomness. As for the SML method as well as its latent version, despite a few high performance, its success is yet sporadic and very unstable. As the noise level increases, interactions between any two non-experts naturally become more significant. However, the latent SML method simply assumes the conditional independence structure between non-experts given their respective latent variables, as a consequence, its performance becomes worse with more noises. In addition, the time cost is also a big concern especially in the execution of the latent SML method. In order to find out the best total number of latent variables, an optimization problem which involves multiple rank-one matrix completions has to be solved p times. The workload would grow substantially as the dimensionality p increases.

p	d_0	Hit Rate	Precision	Bayes	SML	SML-Latent	DS	MV	AMV
25	$\log p$	0.986	0.987	0.942	0.647	0.581	0.513	0.607	0.946
	$\sqrt{p}$	0.998	1	0.981	0.409	0.221	0.459	0.429	0.981
	$p/4$	0.986	0.996	0.994	0.978	0.790	0.568	0.732	0.994
36	$\log p$	0.75	0.75	0.956	0.747	0.750	0.622	0.692	0.909
	$\sqrt{p}$	0.988	0.994	0.986	0.329	0.829	0.541	0.659	0.985
	$p/4$	0.963	0.973	0.998	0.994	0.952	0.604	0.867	0.998
49	$\log p$	0.984	0.991	0.960	0.728	0.731	0.509	0.599	0.958
	$\sqrt{p}$	0.984	0.989	0.994	0.988	0.968	0.566	0.787	0.994
	$p/4$	0.958	0.995	0.994	0.525	0.909	0.532	0.642	0.997
64	$\log p$	0.981	0.987	0.984	0.420	0.253	0.497	0.453	0.982
	$\sqrt{p}$	0.973	0.977	0.692	0.072	0.153	0.470	0.466	0.664
	$p/4$	0.907	0.933	1	0.998	0.940	0.742	0.945	1
81	$\log p$	0.984	0.992	0.984	0.692	0.674	0.509	0.617	0.984
	$\sqrt{p}$	0.991	0.998	0.998	0.282	0.927	0.531	0.609	0.998
	$p/4$	0.656	0.754	0.995	0.696	0.107	0.511	0.548	0.995
100	$\log p$	0.8	0.8	0.961	0.701	0.946	0.514	0.639	0.961
	$\sqrt{p}$	0.966	0.98	0.998	0.961	0.936	0.503	0.655	0.998
	$p/4$	0.892	0.979	1	0.957	0.869	0.688	0.912	1

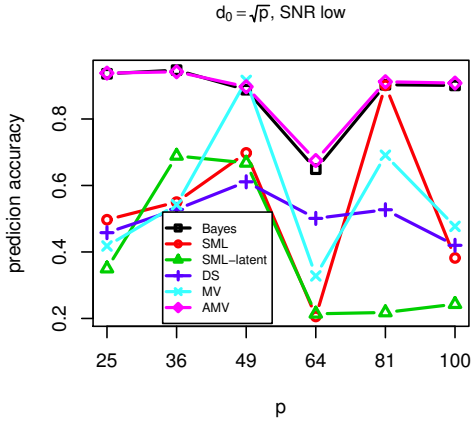
Table 3.1: Performance summary table under high SNR. The first two columns correspond to the two metrics defined in (3.17). The third to the last column is the prediction accuracy corresponding to our two-step method of estimating Bayes classifier, SML method, latent SML method, Dawid-Skene estimator, majority vote, and augmented majority vote respectively.



(a) high



(b) medium



(c) low

Figure 3.5: Predictive performance comparisons under the scenario $d_0 = \sqrt{p}$ for all three levels of signal-to-noise ratio among all 6 methods including Bayes classifier, SML, latent SML, Dawid-Skene, Majority Vote and Augmented Majority Vote.

p	d_0	Hit Rate	Precision	Bayes	SML	SML-Latent	DS	MV	AMV
25	$\log p$	0.984	0.989	0.934	0.596	0.365	0.519	0.584	0.939
	$\sqrt{p}$	0.971	0.993	0.941	0.499	0.606	0.498	0.574	0.950
	$p/4$	0.901	0.959	0.989	0.920	0.381	0.519	0.674	0.986
36	$\log p$	0.75	0.75	0.964	0.792	0.813	0.527	0.675	0.921
	$\sqrt{p}$	0.733	0.760	0.973	0.796	0.750	0.690	0.805	0.950
	$p/4$	0.904	0.949	0.993	0.518	0.557	0.523	0.699	0.994
49	$\log p$	0.982	0.995	0.959	0.728	0.577	0.500	0.612	0.959
	$\sqrt{p}$	0.967	0.984	0.899	0.392	0.257	0.494	0.370	0.909
	$p/4$	0.804	0.891	0.996	0.952	0.891	0.572	0.720	0.990
64	$\log p$	0.963	0.997	0.977	0.325	0.450	0.496	0.488	0.976
	$\sqrt{p}$	0.940	0.964	0.750	0.064	0.040	0.477	0.416	0.720
	$p/4$	0.758	0.876	0.991	0.859	0.755	0.629	0.900	0.989
81	$\log p$	0.985	0.997	0.984	0.460	0.636	0.527	0.614	0.984
	$\sqrt{p}$	0.906	0.929	0.995	0.064	0.824	0.492	0.468	0.995
	$p/4$	0.462	0.685	0.796	0.852	0.781	0.509	0.507	0.790
100	$\log p$	0.973	0.994	0.983	0.477	0.452	0.503	0.599	0.983
	$\sqrt{p}$	0.879	0.919	0.907	0.508	0.579	0.484	0.415	0.900
	$p/4$	0.291	0.480	0.589	0.840	0.567	0.593	0.778	0.596

Table 3.2: Performance summary table under medium SNR. The first two columns correspond to the two metrics defined in (3.17). The third to the last column is the prediction accuracy corresponding to our two-step method of estimating Bayes classifier, SML method, latent SML method, Dawid-Skene estimator, majority vote, and augmented majority vote respectively.

Real Data

We also evaluated the predictive performance of our algorithm using real data sets. In particular, we presented results based on three popular UCI datasets: Wisconsin Breast Cancer(WBC), MAGIC Gamma Telescope(MAGIC), Pen-Based Recognition of Handwritten Digits(DIGITS). We investigated a binary classification task on the DIGITS dataset: odd vs even digits. For each data set, the Crowdsourced ensemble was constructed using various machine learning algorithms including 10 Random Forests, 10 SVM, 10 Naive Bayes, 10 Logistic regressions, and 10 Decision trees. Each classifier was trained on a randomly selected subset of the data. In order to mimic the noisy nature of Crowdsourcing, we randomly corrupted a small portion in the ensemble by reversing their labelings. We also added labelings given by

p	d_0	Hit Rate	Precision	Bayes	SML	SML-Latent	DS	MV	AMV
25	$\log p$	0.978	0.986	0.946	0.772	0.548	0.533	0.665	0.940
	$\sqrt{p}$	0.921	0.985	0.936	0.497	0.350	0.458	0.418	0.938
	$p/4$	0.838	0.788	0.976	0.922	0.823	0.577	0.860	0.979
36	$\log p$	0.75	0.75	0.896	0.726	0.477	0.532	0.716	0.916
	$\sqrt{p}$	0.816	0.900	0.947	0.550	0.689	0.528	0.542	0.942
	$p/4$	0.254	0.291	0.994	0.443	0.158	0.596	0.966	0.993
49	$\log p$	0.942	0.980	0.955	0.529	0.413	0.506	0.647	0.957
	$\sqrt{p}$	0.475	0.487	0.888	0.698	0.668	0.611	0.916	0.897
	$p/4$	0.481	0.685	0.773	0.894	0.283	0.483	0.712	0.794
64	$\log p$	0.928	0.967	0.977	0.517	0.244	0.478	0.453	0.977
	$\sqrt{p}$	0.561	0.578	0.649	0.206	0.214	0.501	0.328	0.675
	$p/4$	0.144	0.190	0.993	0.038	0	0.676	0.999	0.993
81	$\log p$	0.963	0.988	0.981	0.461	0.361	0.514	0.557	0.981
	$\sqrt{p}$	0.498	0.492	0.903	0.901	0.218	0.527	0.691	0.912
	$p/4$	0.112	0.292	0.711	0	0.215	0.390	0.260	0.704
100	$\log p$	0.8	0.8	0.961	0.674	0.694	0.577	0.680	0.961
	$\sqrt{p}$	0.291	0.293	0.901	0.382	0.243	0.420	0.477	0.908
	$p/4$	0.17	0.293	0.273	0	0.006	0.482	0.046	0.263

Table 3.3: Performance summary table under low SNR. The first two columns correspond to the two metrics defined in (3.17). The third to the last column is the prediction accuracy corresponding to our two-step method of estimating Bayes classifier, SML method, latent SML method, Dawid-Skene estimator, majority vote, and augmented majority vote respectively.

random guessing with different probabilities to the ensemble. Table 3.4 summarizes the prediction performance results. The spectral method is very sensible to the corruption level, while our method behaves very stable.

Dataset	Data size	Corruption	MV	AMV	Bayes	SML	SML-Latent
WBC	699	0.5	0.364	0.852	0.953	0.053	0.051
MAGIC	19020	0.4	0.646	0.780	0.850	0.830	0.828
DIGITS	3498	0.3	0.855	0.911	0.910	0.904	0.903

Table 3.4: Prediction accuracy comparison on three UCI datasets. Corruption measures the portion of corrupted classifiers in the ensemble.

4 CONCLUDING REMARKS

In this thesis, two different lines of work have been discussed. Chapter 1 & 2 focus on the estimation of a Euclidean distance matrix and its induced inference. The work presented in Chapter 1 is, as far as to my knowledge, the first to study the performance of a low-rank estimator for an Euclidean distance matrix. More specifically, the contributions so far have been made in this chapter have three folds: (1) build a one-to-one map from an EDM to a kernel matrix by introducing a novel concept of minimum-trace-kernel (2) propose a simple estimator that encourages low rankness by applying a constant shrinkage to all observed pairwise distances. The effect of shrinkage determines the degree of dimension reduction. (3) provide an efficient algorithm to compute this estimator. This algorithm is also well adaptive to the missing case in which some pairwise distances are not available. (4) support visualization on a set of objects of interest from an arbitrary domain through its 1/2/3D embeddings. In Chapter 2, a DANOVA framework has been proposed to generalize (M)ANOVA models to non-Euclidean data. Explicit forms of four widely used test statistics, including Wilk's Lambda, Hotelling-Lawley trace, Pillai-Bartlett trace, and Roy's largest root, have been presented and we show that these test statistics are invariant under rigid motions, namely, unique to a given dissimilarity matrix. A permutation test is followed to calculate the p-value. Numerical experiments demonstrate that the power of these four test statistics is well-matched and superior to the pseudo-F ratio which is a natural extension of the F-statistic in the univariate case. Real data analysis further suggests that the HIV progression rate has a significant impact on the Vpu sequence diversity.

In Chapter 3, we study the binary classification problem under the unsupervised setting. Our motivation arises from an intrinsic drawback of crowdsourcing, in which a classifier's quality is not measurable. A new ensemble method consisting of a pruning step and a predicting step has been developed. We show that the existence and uniqueness of an Ising model that minimizes the Kullback-Leibler divergence to the marginal distribution of the ensemble. This fact allows us to

show the neighbourhood of the true classifier can be recovered successfully with exponentially decaying error under the high-dimensional setting. The Bayes classifier can be consistently estimated in the subsequent step. In both simulations and real data analysis, our method has exhibited great advantages over several other popular methods.

A APPENDIX

Proofs in Chapter 1

Proof of Theorem 1.1. Denote by $M_0 = -JDJ/2$. We first show that $M_0 \in \mathcal{M}(D)$. Note first that

$$J(e_i - e_j) = (e_i - e_j).$$

Therefore,

$$\langle M_0, B_{ij} \rangle = -\frac{1}{2}(e_i - e_j)^\top JDJ(e_i - e_j) = -\frac{1}{2}(e_i - e_j)^\top D(e_i - e_j) = d_{ij},$$

where in the last equality follows from the facts that D is symmetric and $\mathrm{diag}(D) = \mathbf{0}$. Together with the fact that $M_0 \succeq 0$ (Schönberg, 1935; Young and Householder, 1938), this implies that $M_0 \in \mathcal{M}(D)$.

Next, we show that for any $M \in \mathcal{M}(D)$, $\mathrm{trace}(M_0) \leqslant \mathrm{trace}(M)$. To this end, observe that

$$D = \mathscr{T}(M) = \mathrm{diag}(M)\mathbf{1}^\top + \mathbf{1}\mathrm{diag}(M)^\top - 2M.$$

Then

$$\begin{aligned}
\mathrm{trace}(M_0) &= \mathrm{trace}(-JDJ/2) \\
&= \frac{1}{2}\mathrm{trace}\left[\left(I - \frac{\mathbf{1}\mathbf{1}^\top}{n}\right)\left(2M - \mathrm{diag}(M)\mathbf{1}^\top - \mathbf{1}\mathrm{diag}(M)^\top\right)\left(I - \frac{\mathbf{1}\mathbf{1}^\top}{n}\right)\right] \\
&= \frac{1}{2}\mathrm{trace}\left(2M - \mathrm{diag}(M)\mathbf{1}^\top - \mathbf{1}\mathrm{diag}(M)^\top\right) \\
&\quad -\frac{1}{n}\mathbf{1}^\top\left(2M - \mathrm{diag}(M)\mathbf{1}^\top - \mathbf{1}\mathrm{diag}(M)^\top\right)\mathbf{1} \\
&\quad +\frac{1}{2n^2}\mathrm{trace}\left[\mathbf{1}\mathbf{1}^\top\left(2M - \mathrm{diag}(M)\mathbf{1}^\top - \mathbf{1}\mathrm{diag}(M)^\top\right)\mathbf{1}\mathbf{1}^\top\right] \\
&= -\frac{1}{2n}\mathbf{1}^\top\left(2M - \mathrm{diag}(M)\mathbf{1}^\top - \mathbf{1}\mathrm{diag}(M)^\top\right)\mathbf{1} \\
&= \mathrm{trace}(M) - \frac{1}{n}\mathbf{1}^\top M\mathbf{1}.
\end{aligned}$$

The positive semi-definiteness of M ensures that $\mathbf{1}^\top M\mathbf{1} \geqslant 0$, which implies that M_0 has the minimum trace in $\mathcal{M}(D)$. We now show it is also the only one.

Assume the contrary that there exists an $M \in \mathcal{M}(D)$ such that $M \neq M_0$ yet $\mathsf{trace}(M) = \mathsf{trace}(M_0)$. Following the previous calculation, we have $\mathbf{1}^\top M\mathbf{1} = 0$. Recall that $M \succeq 0$. The fact that $\mathbf{1}^\top M\mathbf{1} = 0$ necessarily implies that $\mathbf{1} \in \mathrm{k}er(M)$. As a result, $M = JMJ$, and

$$M - M_0 = J(M - M_0)J.$$

On the other hand,

$$\langle M, B_{ij}\rangle = \langle M_0, B_{ij}\rangle = d_{ij}, \quad \forall i < j.$$

Therefore,

$$\langle J(M - M_0)J, B_{ij}\rangle = \langle M - M_0, B_{ij}\rangle = 0, \quad \forall i < j.$$

It is not hard to see that

$$\{B_{ij} : i < j\} \cup \{e_i e_i^\top : 1 \leqslant i \leqslant n\}$$

forms a basis of the collection of $n \times n$ symmetric matrices. In other words, there exists α_{ij} $(1 \leqslant i \leqslant j)$ such that

$$M - M_0 = \sum_{1 \leqslant i < j \leqslant n} \alpha_{ij} B_{ij} + \sum_{i=1}^{n-1} \alpha_{ii} e_i e_i.$$

Recall that $\mathbf{1} \in \mathrm{k}er(M) \cap \mathrm{k}er(M_0)$. Hence

$$(M - M_0)\mathbf{1} = [\alpha_{11}, \ldots, \alpha_{nn}]^\top = \mathbf{0}.$$

In other words,

$$M - M_0 = \sum_{1 \leqslant i < j \leqslant n} \alpha_{ij} B_{ij}.$$

Thus

$$\|M - M_0\|_{\mathrm{F}}^2 = \|J(M - M_0)J\|_{\mathrm{F}}^2 = \sum_{1 \leqslant i < j \leqslant n} \alpha_{ij} \langle J(M - M_0)J, B_{ij} \rangle = 0.$$

This obviously contradicts with the assumption that $M \neq M_0$.

The second statement follows from the same argument. Note that $PP^\top \in \mathcal{M}(D)$. Because the embedding points are centered, we have $\mathbf{1}^\top PP^\top \mathbf{1} = 0$. The previous argument then suggests that $PP^\top = M_0$. □

Proof of Theorem 1.3. Recall that $J = I - (\mathbf{1}\mathbf{1}^\top/n)$. Observe that $D_0 = (n-1)I - nJ$. Therefore, for any $M \in \mathcal{D}_n$,

$$\begin{aligned}
\left\| \left(X - \frac{\lambda_n}{2n} D_0 \right) - M \right\|_{\mathrm{F}}^2 &= \|X - M\|_{\mathrm{F}}^2 + \frac{\lambda_n}{n} \langle M, D_0 \rangle + (\text{terms not involving } M) \\
&= \|X - M\|_{\mathrm{F}}^2 + \frac{\lambda_n}{n} \langle M, (n-1)I - nJ \rangle + (\text{terms not involving } M) \\
&= \|X - M\|_{\mathrm{F}}^2 - \lambda_n \langle M, J \rangle + (\text{terms not involving } M),
\end{aligned}$$

where the last equality follows from the fact that any distance matrix is hollow, e.g., its diagonals are zeros, hence $\langle M, I \rangle = 0$. Because J is idempotent,

$$\langle M, J \rangle = \langle M, J^2 \rangle = \mathrm{trace}(JMJ).$$

Therefore,

$$\begin{aligned}
\mathscr{P}_{\mathcal{D}_n} \left(X - \frac{\lambda_n}{2n} D_0 \right) &= \operatorname*{argmin}_{M \in \mathcal{D}_n} \left\{ \frac{1}{2} \|X - M\|_{\mathrm{F}}^2 - \frac{\lambda_n}{2} \mathrm{trace}(JMJ) \right\} \\
&= \operatorname*{argmin}_{M \in \mathcal{D}_n} \left\{ \frac{1}{2} \|X - M\|_{\mathrm{F}}^2 + \lambda_n \mathrm{trace} \left(-\frac{1}{2} JMJ \right) \right\},
\end{aligned}$$

which, in the light of (1.4), implies the desired statement. □

Proof of Theorem 1.5. By Theorem 1.3, $\hat{D} = \mathscr{P}_{\mathcal{D}_n}(X - (\lambda_n/2n)D_0)$. Write $\eta_n = \lambda_n/(2n)$ for simplicity. Recall that for any $M \in \mathbb{R}^{n\times n}$, its projection to the closed convex set $\mathcal{D}_n$, $\mathscr{P}_{\mathcal{D}_n}(M)$, can be characterized by the so-called Kolmogorov criterion:

$$\langle A - \mathscr{P}_{\mathcal{D}_n}(M), M - \mathscr{P}_{\mathcal{D}_n}(M)\rangle \leqslant 0, \qquad \forall A \in \mathcal{D}_n.$$

See, e.g., Escalante and Raydan (2011). In particular, taking $M = X - \eta_n D_0$ yields

$$\langle A - \hat{D}, D - \hat{D}\rangle \leqslant \langle X - D - \eta_n D_0, \hat{D} - A\rangle.$$

A classical result in distance geometry by Schönberg (1935) indicates that a distance matrix is conditionally negative semi-definite on the set

$$\mathcal{X}_n = \{x \in \mathbb{R}^n : x^\top \mathbf{1} = 0\},$$

that is, $x^\top Mx \leqslant 0$ for any $x \in \mathcal{X}_n$. See also Young and Householder (1938). In other words, if $M \in \mathcal{D}_n$, then the so-called Schönberg transform JMJ is negative semi-definite where, as before, $J = I - (\mathbf{1}\mathbf{1}^\top/n)$.

Let V be the eigenvectors of JAJ, and $V_\perp$ be an orthonormal basis of the orthogonal complement of the linear subspace spanned by $\{\mathbf{1}\}$ and V. Then $[\mathbf{1}/\sqrt{n}, V, V_\perp]$ forms an orthonormal basis of $\mathbb{R}^n$. Then for any symmetric matrix M, write

$$M = \mathscr{P}_0 M + \mathscr{P}_1 M,$$

where

$$\mathscr{P}_1 M = V_\perp V_\perp^\top M V_\perp V_\perp^\top$$

and

$$\begin{aligned}\mathscr{P}_0 M &= M - \mathscr{P}_1 M = [\mathbf{1}/\sqrt{n}, V][\mathbf{1}/\sqrt{n}, V]^\top M[\mathbf{1}/\sqrt{n}, V][\mathbf{1}/\sqrt{n}, V]^\top \\ &\quad + V_\perp V_\perp^\top M[\mathbf{1}/\sqrt{n}, V][\mathbf{1}/\sqrt{n}, V]^\top + [\mathbf{1}/\sqrt{n}, V][\mathbf{1}/\sqrt{n}, V]^\top M V_\perp V_\perp^\top.\end{aligned}$$

Therefore,

$$\begin{aligned}\langle X-D, \hat{D}-A\rangle &= \langle \mathscr{P}_0(X-D), \mathscr{P}_0(\hat{D}-A)\rangle + \langle \mathscr{P}_1(X-D), \mathscr{P}_1(\hat{D}-A)\rangle \\ &= \langle \mathscr{P}_0(X-D), \mathscr{P}_0(\hat{D}-A)\rangle + \langle \mathscr{P}_1(X-D), \mathscr{P}_1\hat{D}\rangle \\ &\leqslant \|\mathscr{P}_0(X-D)\|\|\mathscr{P}_0(\hat{D}-A)\|_* + \|\mathscr{P}_1(X-D)\|\|\mathscr{P}_1\hat{D}\|_*\end{aligned}$$

where in the last inequality we used the fact that for any matrices $M_1, M_2 \in \mathbb{R}^{n\times n}$,

$$\langle M_1, M_2\rangle \leqslant \|M_1\|\|M_2\|_*,$$

and $\|\cdot\|$ and $\|\cdot\|_*$ represent the matrix spectral and nuclear norm respectively. It is clear that

$$\|\mathscr{P}_1(X-D)\| \leqslant \|X-D\|,$$

and

$$\|\mathscr{P}_0(X-D)\| \leqslant 2\|X-D\|.$$

Then,

$$\langle X-D, \hat{D}-A\rangle \leqslant \|X-D\|\left(2\|\mathscr{P}_0(\hat{D}-A)\|_* + \|\mathscr{P}_1\hat{D}\|_*\right)$$

On the other hand, recall that both D and $\hat{D}$ are hollow and $D_0 = (n-1)I - nJ$. Thus,

$$\begin{aligned}\langle D_0, \hat{D}-A\rangle &= n\langle A-\hat{D}, J\rangle \\ &= n\mathsf{trace}(J(A-\hat{D})J) \\ &= -n\mathsf{trace}(VV^\top(\hat{D}-A)VV^\top) - n\mathsf{trace}(\mathscr{P}_1\hat{D}) \\ &= -n\mathsf{trace}(VV^\top(\hat{D}-A)VV^\top) + n\|\mathscr{P}_1\hat{D}\|_* \\ &\geqslant -n\|VV^\top(\hat{D}-A)VV^\top\|_* + n\|\mathscr{P}_1\hat{D}\|_* \\ &\geqslant -n\|\mathscr{P}_0(\hat{D}-A)\|_* + n\|\mathscr{P}_1\hat{D}\|_*,\end{aligned}$$

where the last equality follows from the fact that $\mathscr{P}_1\hat{D}$ is negative semi-definite.

Taking $m\eta_n \geqslant \|X - D\|$ yields that

$$\langle X - D - \lambda_n D_0, \hat{D} - A \rangle \leqslant 3m\eta_n \|\mathscr{P}_0(\hat{D} - A)\|_*.$$

Note that, by Cauchy-Schwartz inequality, for any $M \in \mathbb{R}^{n \times n}$

$$\|M\|_* \leqslant \sqrt{\mathsf{rank}(M)}\|M\|_{\mathrm{F}}.$$

Therefore,

$$\begin{aligned} \|\mathscr{P}_0(\hat{D} - A)\|_* &\leqslant \sqrt{\mathsf{rank}(JAJ) + 1}\|\mathscr{P}_0(\hat{D} - A)\|_{\mathrm{F}} \\ &\leqslant \sqrt{\mathsf{rank}(JAJ) + 1}\|\hat{D} - A\|_{\mathrm{F}} \\ &= \sqrt{\dim(A) + 1}\|\hat{D} - A\|_{\mathrm{F}}, \end{aligned}$$

where the last equality follows from the fact that for any Euclidean distance matrix A, $\dim(A) = \mathsf{rank}(JAJ)$. See, e.g., Schönberg (1935) and Young and Householder (1938). As a result,

$$\langle A - \hat{D}, D - \hat{D} \rangle \leqslant 3m\eta_n \sqrt{\dim(A) + 1}\|\hat{D} - A\|_{\mathrm{F}}.$$

Simple algebraic manipulations show that

$$\langle A - \hat{D}, D - \hat{D} \rangle = \frac{1}{2}\left(\|\hat{D} - D\|_{\mathrm{F}}^2 + \|\hat{D} - A\|_{\mathrm{F}}^2 - \|A - D\|_{\mathrm{F}}^2\right).$$

Thus,

$$\|\hat{D} - D\|_{\mathrm{F}}^2 + \|\hat{D} - A\|_{\mathrm{F}}^2 \leqslant \|A - D\|_{\mathrm{F}}^2 + 6m\eta_n \sqrt{\dim(A) + 1}\|\hat{D} - A\|_{\mathrm{F}},$$

which implies that

$$\begin{aligned}\|\hat{D}-D\|_{\mathrm{F}}^2 &\leqslant \|A-D\|_{\mathrm{F}}^2+6m\eta_n\sqrt{\dim(A)+1}\|\hat{D}-A\|_{\mathrm{F}}-\|\hat{D}-A\|_{\mathrm{F}}^2\\ &= \|A-D\|_{\mathrm{F}}^2+9n^2\eta_n^2(\dim(A)+1)-\left(\|\hat{D}-A\|_{\mathrm{F}}-3m\eta_n\sqrt{\dim(A)+1}\right)^2\\ &\leqslant \|A-D\|_{\mathrm{F}}^2+9n^2\eta_n^2(\dim(A)+1).\end{aligned}$$

This completes the proof. □

Proof of Corollary 1.7. Observe first that

$$\hat{D}_r = \operatorname*{argmin}_{M\in\mathcal{D}_r}\|J(M-\hat{D})J\|_{\mathrm{F}}^2.$$

Therefore,

$$\begin{aligned}\|J(\hat{D}_r-D)J\|_{\mathrm{F}}^2 &\leqslant 2\|J(\hat{D}_r-\hat{D})J\|_{\mathrm{F}}^2+2\|J(\hat{D}-D)J\|_{\mathrm{F}}^2\\ &\leqslant 2\|J(D_r-\hat{D})J\|_{\mathrm{F}}^2+2\|\hat{D}-D\|_{\mathrm{F}}^2\\ &\leqslant 4\|J(D_r-D)J\|_{\mathrm{F}}^2+4\|J(\hat{D}-D)J\|_{\mathrm{F}}^2+2\|\hat{D}-D\|_{\mathrm{F}}^2\\ &\leqslant 4\min_{M\in\mathcal{D}_n(r)}\|J(D-M)J\|_{\mathrm{F}}^2+6\|\hat{D}-D\|_{\mathrm{F}}^2\end{aligned}$$

On the other hand, taking $M=D_r$ in Theorem 1.5 yields

$$\begin{aligned}\frac{1}{n^2}\|\widehat{D}-D\|_{\mathrm{F}}^2 &\leqslant \frac{1}{n^2}\|D_r-D\|_{\mathrm{F}}^2+9\eta_n^2(r+1)\\ &= \frac{1}{n^2}\min_{M\in\mathcal{D}_n(r)}\|J(D-M)J\|_{\mathrm{F}}^2+9\eta_n^2(r+1),\end{aligned}$$

where, as before, $\eta_n=\lambda_n/2n$. Therefore,

$$\frac{1}{n^2}\|J(\hat{D}_r-D)J\|_{\mathrm{F}}^2\leqslant\frac{10}{n^2}\min_{M\in\mathcal{D}_n(r)}\|J(D-M)J\|_{\mathrm{F}}^2+54\eta_n^2(r+1),$$

which completes the proof. □

Proofs in Chapter 3

Proof of Lemma 3.1. Let's consider a family of Ising models

$$\mathcal{P}_\theta = \left\{ \exp\left(\sum_{\substack{r,t \in \mathcal{N}_0 \cup \mathcal{N}_s \backslash 0 \\ r \neq t}} \theta_{rt} f_r f_t - A(\theta) \right) \middle| \theta \in \mathbb{R}^d, d = \frac{(d_s + d_0 - 1)(d_s + d_0 - 2)}{2}, A(\theta) < +\infty \right\}$$

The goal is to find a distribution $Q_{\tilde{\theta}} \in \mathcal{P}_\theta$ that has the minimal Kullback-Leibler divergence $D_{KL}(P_{\theta^*}(f_{\mathcal{N}_0 \cup \mathcal{N}_s \backslash 0}) \| Q)$.

$$\begin{aligned}
\min_{Q \in \mathcal{P}_\theta} D_{KL}(P_{\theta^*}(f_{\mathcal{N}_0 \cup \mathcal{N}_s \backslash 0}) \| Q) &\Longleftrightarrow \min_{Q \in \mathcal{P}_\theta} \sum_{y \in \{-1,+1\}^{d_s + d_0 - 1}} -p^*(y) \log q(y) \\
&\Longleftrightarrow \min_{Q \in \mathcal{P}_\theta} \sum_{y \in \{-1,+1\}^{d_s + d_0 - 1}} -p^*(y) \left(\sum_{\substack{r,t \in \mathcal{N}_0 \cup \mathcal{N}_s \backslash 0 \\ r \neq t}} \theta_{rt} f_r f_t - A(\theta) \right) \\
&\Longleftrightarrow \min_{Q \in \mathcal{P}_\theta} \left(A(\theta) - \sum_{\substack{r,t \in \mathcal{N}_0 \cup \mathcal{N}_s \backslash 0 \\ r \neq t}} \theta_{rt} \Big(p^*(f_r f_t = 1) - p^*(f_r f_t = -1) \Big) \right)
\end{aligned}$$

where we use $p^*(\cdot)$ as a shorthand for $P_{\theta^*}(f_{\mathcal{N}_0 \cup \mathcal{N}_s \backslash 0})$.

Since $\mathcal{P}_\theta$ is a subset of the regular exponential family, $A(\theta)$ is a convex function of θ. Therefore the optimal solution $Q_{\tilde{\theta}}(\cdot)$ can be given by the stationary condition:

$$\frac{\partial A(\theta)}{\theta_{rt}} = \mathbb{E}_{p^*}[f_r f_t] \tag{A.1}$$

where $\mathbb{E}_{p^*}[f_r f_t] = p^*(f_r f_t = 1) - p^*(f_r f_t = -1)$.

$\forall r \in \mathcal{N}_s \backslash 0, \forall t \in \mathcal{N}_0$,

$$\mathbb{E}_{p^*}[f_r f_t] = \frac{e^{2\theta^*_{rt}} - 1}{e^{2\theta^*_{rt}} + 1}$$

$\forall(r,t) \subseteq \mathcal{N}_0$,

$$\frac{p^*(f_r f_t = 1)}{p^*(f_r f_t = -1)} = \frac{e^{\theta^*_{0r}+\theta^*_{0t}+\theta^*_0} + e^{-\theta^*_{0r}-\theta^*_{0t}-\theta^*_0} + e^{-\theta^*_{0r}-\theta^*_{0t}+\theta^*_0} + e^{\theta^*_{0r}+\theta^*_{0t}-\theta^*_0}}{e^{\theta^*_{0r}-\theta^*_{0t}+\theta^*_0} + e^{-\theta^*_{0r}+\theta^*_{0t}-\theta^*_0} + e^{-\theta^*_{0r}+\theta^*_{0t}+\theta^*_0} + e^{\theta^*_{0r}-\theta^*_{0t}-\theta^*_0}} \triangleq b_{st}$$

$$\mathbb{E}_{p^*}[f_r f_t] = \frac{b_{st}-1}{b_{st}+1} \triangleq \mu_{st}$$

On the other hand, the moment matching conditions in (3.1) reveals that

$$\left.\frac{\partial A(\theta)}{\partial \theta_{rt}}\right|_{\tilde{\theta}_{rt}} = \mathbb{E}_{\tilde{\theta}}[f_r f_t] = \frac{e^{2\tilde{\theta}_{rt}}-1}{e^{2\tilde{\theta}_{rt}}+1} \tag{A.2}$$

Therefore, combining (A.1) and (A.2) together,

$$\tilde{\theta}_{rt} = \theta^*_{rt},\ \forall(r,t) \nsubseteq \mathcal{N}_0$$

$$\tilde{\theta}_{rt} = \frac{1}{2}\log(b_{st}),\ \forall(r,t) \subseteq \mathcal{N}_0$$

The uniqueness of $Q_{\tilde{\theta}}(\cdot)$ is naturally followed. □

Proof of Theorem 3.2. The proof basically goes through the same flow given in the proof of Lemma 3.1 and hence we dismiss the repetitive details by only providing a proving sketch. First, we start to build an Ising model family indexed by edge potential parameters $\theta \in \mathbb{R}^d$, where $d = \frac{p(p-1)}{2}$. Next, by using the stationary condition (A.1) combined with the moment matching equations (A.2), the desired result can be obtained. □

Proof of Proposition 3.3. We first show $\tilde{\theta}_{st} \neq 0$. Consider a bivariate function $f(x,y) = e^{x+y} + e^{-x-y} + e^{x-y} + e^{y-x}$, for a fixed $y \in \mathbb{R}$, $f(\cdot, y)$ is a symmetric and strictly convex function since $\frac{\partial^2 f}{\partial x^2} = (e^x + e^{-x})(e^y + e^{-y}) > 0$. To show $\tilde{\theta}_{st} \neq 0$, it is essential to show $f(\theta^*_{0s}+\theta^*_{0t}, \theta^*_0) \neq f(\theta^*_{0s}-\theta^*_{0t}, \theta^*_0)$. As $\theta^*_{0s}, \theta^*_{0t} \neq 0$, $\theta^*_{0s}+\theta^*_{0t} \neq \pm(\theta^*_{0s}-\theta^*_{0t})$, which ensures the argument inside the logarithm is not 1. Next, since $\tilde{\theta}_{st} > 0 \iff f(\theta^*_{0s}+\theta^*_{0t}, \theta^*_0) > f(\theta^*_{0s}-\theta^*_{0t}, \theta^*_0)$, $\theta^*_{0s}\theta^*_{0t} > 0 \iff |\theta^*_{0s}+\theta^*_{0t}| > |\theta^*_{0s}-\theta^*_{0t}|$,

and the strict convexity of $f(\cdot, \theta_0^*)$ ensures $f(\theta_{0s}^* + \theta_{0t}^*, \theta_0^*) > f(\theta_{0s}^* - \theta_{0t}^*, \theta_0^*) \iff |\theta_{0s}^* + \theta_{0t}^*| > |\theta_{0s}^* - \theta_{0t}^*|$, the proof is complete. □

Proof of Theorem 3.5. We first show $\forall s \in \mathcal{N}_0, s \in \mathcal{A}$. Obviously, $s \in A_s$. $\forall t \in \mathcal{N}_0 \backslash s$, Proposition 3.3 ensures $\tilde{\theta}_{st} \neq 0$ such that $s \in \tilde{\mathcal{N}}_t$. Therefore, $|\tilde{\mathcal{N}}_s| \geqslant d_0 - 1$. Suppose there exist $r \neq s$ such that $r \in A_s$, then $r \notin \mathcal{N}_0$ and $d_r \geqslant d_0 - 1$ since it must appear in each neighbourhood $\tilde{\mathcal{N}}_t, \forall t \in \mathcal{N}_0 \backslash s$, which contradicts with $(G1)$. Therefore, $s \in \mathcal{A}$. Next, we need to show $\forall s \in \mathcal{N}_0, |\tilde{\mathcal{N}}_s| \geqslant i_s - 1$. This is obvious due to the fact $|\tilde{\mathcal{N}}_s| \geqslant d_0 - 1$ and they correspond to the first d_0 largest $|\tilde{\mathcal{N}}_s|$. On the other side, if there exists a non-expert node $s \in \mathcal{A}$, as $|\tilde{\mathcal{N}}_s| < d_0 - 1$, then its corresponding index $i_s \geqslant d_0 + 1$, indicating $|\tilde{\mathcal{N}}_s| < i_s - 1$, therefore $\{s \in \mathcal{A} : |\tilde{\mathcal{N}}_s| \geqslant i_s - 1\} \subseteq \mathcal{N}_0$. □

REFERENCES

[1] Anderson, Marti J. 2001. A new method for non-parametric multivariate analysis of variance. *Austral ecology* 26(1):32–46.

[2] Anderson, TW. 1984. Multivariate statistical analysis. *VVi11ey and Sons, New York, NY*.

[3] Bishop, YM. 1977. M, fienberg, se & holland, pw 1975. discrete multivariate analysis theory and practice.

[4] Breiman, Leo. 2001. Random forests. *Machine learning* 45(1):5–32.

[5] Chen, L., and A. Buja. 2009. Local multidimensional scaling for nonlinear dimension reduction, graph drawing, and proximity analysis. *Journal of the American Statistical Association* 104(485):209–219.

[6] Chen, Lisha, and Andreas Buja. 2013. Stress functions for nonlinear dimension reduction, proximity analysis, and graph drawing. *Journal of Machine Learning Research* 14(4):1145–1173.

[7] Cross, George R, and Anil K Jain. 1983. Markov random field texture models. *IEEE Transactions on Pattern Analysis and Machine Intelligence* (1):25–39.

[8] Dattorro, J. 2013. Convex optimization and euclidean distance geometry. *Palo Alto: Meboo*.

[9] Dawid, Alexander Philip, and Allan M Skene. 1979. Maximum likelihood estimation of observer error-rates using the em algorithm. *Applied statistics* 20–28.

[10] Dempster, Arthur P, Nan M Laird, and Donald B Rubin. 1977. Maximum likelihood from incomplete data via the em algorithm. *Journal of the royal statistical society. Series B (methodological)* 1–38.

[11] Donmez, Pinar, Guy Lebanon, and Krishnakumar Balasubramanian. 2010. Unsupervised supervised learning i: Estimating classification and regression errors without labels. *Journal of Machine Learning Research* 11(Apr):1323–1351.

[12] Durbin, Richard, Sean R Eddy, Anders Krogh, and Graeme Mitchison. 1998. *Biological sequence analysis: probabilistic models of proteins and nucleic acids*. Cambridge university press.

[13] Dykstra, Richard L. 1983. An algorithm for restricted least squares regression. *Journal of the American Statistical Association* 78(384):837–842.

[14] Edgington, Eugene, and Patrick Onghena. 2007. *Randomization tests*. CRC Press.

[15] Freund, Yoav, and Robert E Schapire. 1995. A desicion-theoretic generalization of on-line learning and an application to boosting. In *European conference on computational learning theory*, 23–37. Springer.

[16] Freund, Yoav, Robert E Schapire, et al. 1996. Experiments with a new boosting algorithm. In *icml*, vol. 96, 148–156.

[17] Glunt, W, Tom L Hayden, S Hong, and J Wells. 1990. An alternating projection algorithm for computing the nearest euclidean distance matrix. *SIAM Journal on Matrix Analysis and Applications* 11(4):589–600.

[18] Hassner, Martin, and Jack Sklansky. 1980. The use of markov random fields as models of texture. *Computer Graphics and Image Processing* 12(4):357–370.

[19] Hastie, Trevor, Robert Tibshirani, and Jerome Friedman. 2009. Unsupervised learning. In *The elements of statistical learning*, 485–585. Springer.

[20] Hayden, Tom L, and Jim Wells. 1988. Approximation by matrices positive semidefinite on a subspace. *Linear Algebra and its Applications* 109:115–130.

[21] Ising, Ernst. 1925. Beitrag zur theorie des ferromagnetismus. *Zeitschrift für Physik A Hadrons and Nuclei* 31(1):253–258.

[22] Jaffe, Ariel, Ethan Fetaya, Boaz Nadler, Tingting Jiang, and Yuval Kluger. 2015. Unsupervised ensemble learning with dependent classifiers. *arXiv preprint arXiv:1510.05830*.

[23] Johnson, Richard Arnold, Dean W Wichern, et al. 2002. *Applied multivariate statistical analysis*. Prentice hall Upper Saddle River, NJ.

[24] Lu, Fan, Sündüz Keleş, Stephen J Wright, and Grace Wahba. 2005. Framework for kernel regularization with application to protein clustering. *Proceedings of the National Academy of Sciences of the United States of America* 102(35):12332–12337.

[25] Lu, Zhaosong, Renato DC Monteiro, and Ming Yuan. 2012. Convex optimization methods for dimension reduction and coefficient estimation in multivariate linear regression. *Mathematical Programming* 131(1):163–194.

[26] Manly, B. 1997. Randomization, bootstrap and monte carlo methods in biology. *Second Edition, Chapman & Hall, London*.

[27] Manning, Christopher D, Hinrich Schütze, et al. 1999. *Foundations of statistical natural language processing*, vol. 999. MIT Press.

[28] Martínez-Muñoz, Gonzalo, Daniel Hernández-Lobato, and Alberto Suárez. 2009. An analysis of ensemble pruning techniques based on ordered aggregation. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 31(2): 245–259.

[29] Negahban, Sahand, and Martin J Wainwright. 2011. Estimation of (near) low-rank matrices with noise and high-dimensional scaling. *The Annals of Statistics* 1069–1097.

[30] Olson, Chester L. 1974. Comparative robustness of six tests in multivariate analysis of variance. *Journal of the American Statistical Association* 69(348):894–908.

[31] Parisi, Fabio, Francesco Strino, Boaz Nadler, and Yuval Kluger. 2014. Ranking and combining multiple predictors without labeled data. *Proceedings of the National Academy of Sciences* 111(4):1253–1258.

[32] Pickering, Suzanne, Stephane Hué, Eun-Young Kim, Susheel Reddy, Steven M Wolinsky, and Stuart JD Neil. 2014. Preservation of tetherin and cd4 counteractivities in circulating vpu alleles despite extensive sequence variation within hiv-1 infected individuals. *PLoS Pathog* 10(1):e1003895.

[33] Pouzet, Maurice. 1979. Note sur le probleme de ulam. *Journal of Combinatorial Theory, Series B* 27(3):231–236.

[34] Ravikumar, Pradeep, Martin J Wainwright, John D Lafferty, et al. 2010. High-dimensional ising model selection using â„“1-regularized logistic regression. *The Annals of Statistics* 38(3):1287–1319.

[35] Ripley, Brian D. 2005. *Spatial statistics*, vol. 575. John Wiley & Sons.

[36] Rohde, Angelika, Alexandre B Tsybakov, et al. 2011. Estimation of high-dimensional low-rank matrices. *The Annals of Statistics* 39(2):887–930.

[37] Roy, Aidan. 2010. Minimal euclidean representations of graphs. *Discrete Mathematics* 310(4):727–733.

[38] Schoenberg, Isaac J. 1935. Remarks to maurice frechet's article"sur la definition axiomatique d'une classe d'espace distances vectoriellement applicable sur l'espace de hilbert. *Annals of Mathematics* 724–732.

[39] Schölkopf, Bernhard, Alexander Smola, and Klaus-Robert Müller. 1998. Nonlinear component analysis as a kernel eigenvalue problem. *Neural computation* 10(5):1299–1319.

[40] Sinai, Ya G, and Aleksandr Borisovich Soshnikov. 1998. A refinement of wigner's semicircle law in a neighborhood of the spectrum edge for random symmetric matrices. *Functional Analysis and Its Applications* 32(2):114–131.

[41] Smola, Alex J, and Bernhard Schölkopf. 1998. *Learning with kernels*. Citeseer.

[42] Székely, Gábor J, Maria L Rizzo, Nail K Bakirov, et al. 2007. Measuring and testing dependence by correlation of distances. *The Annals of Statistics* 35(6): 2769–2794.

[43] Tenenbaum, Joshua B, Vin De Silva, and John C Langford. 2000. A global geometric framework for nonlinear dimensionality reduction. *science* 290(5500): 2319–2323.

[44] Toh, Kim-Chuan, Michael J Todd, and Reha H Tütüncü. 1999. Sdpt3âĽ"a matlab software package for semidefinite programming, version 1.3. *Optimization methods and software* 11(1-4):545–581.

[45] Tütüncü, Reha H, Kim-Chuan Toh, and Michael J Todd. 2003. Solving semidefinite-quadratic-linear programs using sdpt3. *Mathematical programming* 95(2):189–217.

[46] Venna, Jarkko, and Samuel Kaski. 2006. Local multidimensional scaling. *Neural Networks* 19(6):889–899.

[47] Wainwright, Martin J, Michael I Jordan, et al. 2008. Graphical models, exponential families, and variational inference. *Foundations and Trends® in Machine Learning* 1(1–2):1–305.

[48] Woods, John. 1978. Markov image modeling. *IEEE Transactions on Automatic Control* 23(5):846–850.

[49] Wuthrich, Kurt. 1986. *Nmr of proteins and nucleic acids*. Wiley.

[50] Young, Gale, and Alston S Householder. 1938. Discussion of a set of points in terms of their mutual distances. *Psychometrika* 3(1):19–22.

[51] Yuan, Ming, Ali Ekici, Zhaosong Lu, and Renato Monteiro. 2007. Dimension reduction and coefficient estimation in multivariate linear regression. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 69(3):329–346.

[52] Zhang, Luwan, Grace Wahba, and Yuan Ming. 2016. Distance shrinkage and euclidean embedding via regularized kernel estimation. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*.