Information-centric Algorithms for Feature Extraction in High-Dimensional Sequential Data

by

Jiejun Jin

B.E. Southeast University (2019)
S.M. Massachusetts Institute of Technology (2021)

Submitted to the Department of Electrical Engineering and Computer Science
in Partial Fulfillment of the Requirements for the Degree of

DOCTOR OF PHILOSOPHY

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

February 2025

Authored by: Jiejun Jin
Department of Electrical Engineering and Computer Science
January 17, 2025

Certified by: Lizhong Zheng
Professor of Electrical Engineering and Computer Science
Thesis Supervisor

Accepted by: Leslie A. Kolodziejski
Professor of Electrical Engineering and Computer Science
Chair, Department Committee on Graduate Students

Information-centric Algorithms for Feature Extraction in High-Dimensional Sequential Data

by

Jiejun Jin

Submitted to the Department of Electrical Engineering and Computer Science on January 17, 2025 in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

ABSTRACT

Hidden Markov Models (HMMs) are a cornerstone of sequential data analysis, offering a robust framework for modeling observable events influenced by hidden internal states. With applications spanning speech recognition, video analysis, bioinformatics, and financial time series, HMMs enable the prediction and classification of raw data by leveraging their dual-layer stochastic structure: hidden Markov states and observable outputs. However, as real-world data grows increasingly high-dimensional, extracting meaningful features from observations becomes critical to reduce computational complexity while retaining relevant information.

This thesis addresses key challenges in feature extraction for high-dimensional HMMs. Current methods, such as neural networks (NNs), are widely used for nonlinear feature learning but lack mechanisms to prioritize useful features or incorporate known structural constraints. To bridge this gap, this work proposes novel algorithms to decouple representation learning from task-specific objectives and extract features aligned with predefined constraints.

The theoretical foundation, including local information geometry and Hirschfeld-Gebelein-Rényi (HGR) maximal correlation, is introduced in Chapter 2. Chapter 3 details three innovative feature extraction algorithms and their corresponding neural network architectures, highlighting their strengths and limitations. Convergence analyses and tail bounds for these methods are presented in Chapter 4. Numerical simulations validating the efficacy of the proposed approaches are provided in Chapter 5, while Chapter 6 concludes with a summary of contributions and potential future research directions.

This thesis advances the field by offering structured, constraint-aware feature extraction techniques tailored for high-dimensional sequential data, setting the stage for more effective and interpretable inference in HMMs.

Thesis supervisor: Lizhong Zheng
Title: Professor of Electrical Engineering and Computer Science

Acknowledgments

I still can't quite believe that I've completed this thesis and successfully defended it. The journey to this point has been filled with countless events, challenges, and milestones, as well as numerous incredible people who have supported me along the way. While these pages may be short, the list of people I wish to thank is long. I want to use this opportunity to express my heartfelt gratitude and make some acknowledgments.

First, I want to extend my deepest thanks to my thesis supervisor, Professor Lizhong Zheng, and my committee members, Professor Gregory W. Wornell and Professor Polina Golland.

Professor Lizhong has been an exceptionally kind, patient, and encouraging mentor throughout my PhD journey. His guidance has profoundly shaped the way I approach complex problems, and his unwavering support—both academically and personally—has been a cornerstone of my experience. As an international student, I was particularly grateful for his empathy and thoughtful care, which went far beyond research and academics. I feel incredibly fortunate to have had such an inspiring advisor, whose influence will continue to guide me.

I've had the privilege of knowing Professor Greg since my very first class at MIT, Algorithms for Inference, and later as a teaching assistant for his course. His clarity, insight, and dedication to teaching and mentorship have left a lasting impression on me. I thoroughly enjoyed our discussions on research problems and greatly benefited from his profound expertise and thoughtful guidance.

I am also deeply grateful to Professor Polina for her invaluable guidance and unwavering support, especially during the critical final stages of my PhD. Her prompt, thoughtful responses to my questions and her warm, encouraging messages made this journey feel much more manageable. Her kindness and mentorship have been a source of comfort and inspiration.

I would also like to thank my former and current colleagues in the Claude E. Shannon Communication and Network Research Group, Anuran Makur, David Qiu, Erixhen Sula, Mohamed Alhajri, Xiangxiang Xu, Melichan Erol, Xander Morgan, and Eric Wang. Your insightful discussions, constructive advice, and even our gym sessions have been a source of great joy and intellectual stimulation. I will always cherish the moments we shared, from research collaborations to celebrating special occasions together. I wish you all an incredible future!

I want to thank all my friends for your companionship and support. During my time at MIT, I've been fortunate made lifelong friends who have brought immense joy and balance to my life. The sweet memories we've created together will remain with me forever. Your presence helped me reduce the stress and homesickness, and for that, I am deeply grateful.

Finally, and most importantly, I want to thank my family, especially my dear parents. Despite the distance and challenges—such as COVID and visa issues—that have kept us apart for years, your love has been a constant source of strength and motivation for me. You've stood by me through every success and every challenge, offering unconditional support, patience, and encouragement. I am who I am today because of you, and I hope I've made you proud.

Contents

List of Figures

Chapter 1

Introduction

1.1 Background and Motivation

Within the sphere of sequential data analysis, Hidden Markov Models (HMMs) [1], [2] stand out as a powerful class of statistical models. It can be used to describe the evolution of observable events that depend on internal factors as shown in Figure 1.1. An HMM consists of two stochastic processes, namely, an invisible process of hidden states (usually denoted as X) and a visible process of observations (usually denoted as Y). The hidden states form a Markov chain, and the probability distribution of the observations depends on the underlying state.

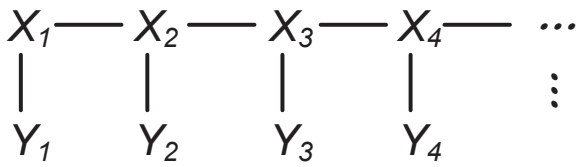


Figure 1.1: Hidden Markov Model (HMM)

Modeling observations in these two layers is very useful, since many real world problems deal with classifying raw observations into a number of categories, or predicting about new states. For example, let us consider the speech recognition problem [3]–[7], for which HMMs have been extensively used for several decades. In speech recognition, we are interested in predicting the uttered word from a recorded speech signal. For this purpose, the speech recognizer tries to find the sequence of phonemes (states) that gave rise to the actual uttered sound (observations). Since there can be a large variation in the actual pronunciation, the original phonemes (and ultimately, the uttered word) cannot be directly observed, and need to be predicted.

This approach is also useful in video analysis [8]–[13], bioinformatics for DNA sequencing [14]–[21], and financial time series research [22]–[30]. As a result, HMMs have become increasingly popular in the fields of modern statistics and machine learning.

Most of the time, extracting features [31]–[34] from observations is the first step when solving an inference problem, because it can reduce the dimension of the data by getting rid of the irrelevant or redundant information while keeping the useful one. This step is more important for high-dimensional HMMs [35]–[37], where the dimension of observations can be extremely large, because the complexity of using the original data directly becomes prohibited due to the increased computational requirements.

Neural networks (NNs) [38]–[40], as widely used, is an efficient and general way to learn good non-linear feature functions from data. However, there are some problems with using NNs directly:

- First, there is no specific metrics to tell which features are more useful, and which are not. The metrics (loss function) are designed based on the specific task of interest. For example, for a classification task [41], the proper loss function may be cross-entropy; for a regression task [42], the metrics of choice may be mean square error. If the goal is to extract general features from the data that could be useful for any further inference tasks, there's no corresponding metrics in literature.

- Second, there is no way to incorporate known structure or constraints of feature functions in current NNs. For example, it is hard to choose features that only represent the symmetric part of information inside the data. This could be useful when there's prior knowledge about the dataset.

This thesis manages to solve these problems. Specifically, we want to separate the representation learning with specific task requirement and be able to obtain features corresponding to certain constraints in the high-dimensional sequential data.

1.2 Notation

General Notation:

- $\mathbb{R}$ represents set of real numbers.
- $\times$ represents Cartesian product.
- exp represents exponential with base e.
- log represents logarithm with base e.
- $f(\epsilon) = o(g(\epsilon))$ means that $\lim_{\epsilon \to 0} \frac{f(\epsilon)}{g(\epsilon)} = 0$
- $f(\epsilon) = O(g(\epsilon))$ means that $\lim_{\epsilon \to 0} \frac{f(\epsilon)}{g(\epsilon)} \leq c$ for some positive real number c.

Probability and Random Variables:

- P, Q, and R represent distribution unless stated otherwise.
- Capital letters (e.g., X, Y, X_i, and Y_i) represent random variables unless stated otherwise.
- Low case letters (e.g., x, y, x_i and y_i) represent particular outcomes unless stated otherwise. Specifically, notation of x_i and $x^{(i)}$ are used interchangeably to represent the i-th observation for variable X.
- Calligraphic letters (e.g., $\mathcal{X}$ and $\mathcal{Y}$) represent the set of possible outcomes (alphabet) for the random variables.
- $|\mathcal{X}|$ represents the size of the set $\mathcal{X}$.
- $\mathcal{P}_{\mathcal{X}}$ represents the probability simplex over alphabet $\mathcal{X}$.
- relint($\mathcal{P}_{\mathcal{X}}$) represents the relative interior of $\mathcal{P}_{\mathcal{X}}$.
- $\mathbb{E}$ represents expectation.
- P_X and P_Y when written without any particular outcome are column vectors with the marginal probabilities in each entry.
- $\sqrt{P_X}$ and $\sqrt{P_Y}$ are the elementwise square root of P_X and P_Y, respectively.
- P_{YX} and $P_{Y|X}$ when written without any particular outcome are the matrix representations of the joint probability distribution and the conditional probability distribution, respectively, where Y varies in the 1st dimension and X varies in the 2nd dimension.

Vectors and Matrices

- Capital Greek letters (e.g., Φ, Ψ, Φ_i, and Ψ_i) represent matrices unless stated otherwise.
- Low case Greek letters (e.g., ϕ, ψ, ϕ_i, and ψ_i) represent vectors unless stated otherwise.
- e_i is a vector with all zero values except the i-th position with an one value.
- $[\cdot]_d$ is the diagonal matrix representation of a column vector.
- $\mathbb{R}^{m\times n}$ is the set of real $m \times n$ matrices.
- tr($\cdot$) represents the trace of a matrix.
- $\sigma_i(A)$ represents the i-th largest singular value of A.

- $\langle \cdot, \cdot \rangle$ represents the inner product between two vectors.
- $\| \cdot \|_2$ is the Euclidean norm of a vector.
- $\| \cdot \|_s$ is the spectrum norm of a matrix.
- $\| \cdot \|_F$ is the Frobenius norm of a matrix.
- $\| \cdot \|_{op}$ is the operator norm of a matrix.

Statistics and Machine Learning

- $\mathcal{L}$ represents the loss function in machine learning optimizations.
- $\check{\cdot}$ represents empirical version of some probabilistic object $\cdot$ (e.g.,random variable, distribution) unless stated otherwise.
- $\hat{\cdot}$ represents estimation of some target variable $\cdot$ unless stated otherwise.
- $I(X, Y)$ represents the Shannon mutual information between random variable X and Y.
- $D(P\|Q)$ represents the Kullback-Leibler (KL) divergence between distribution P and Q.

1.3 Related Work

HMMs have been extensively studied and applied across various fields due to their effectiveness in modeling sequential data. The foundational work of Rabiner [1], [3] laid the groundwork for understanding the structure and functionality of HMMs, particularly in speech recognition. This domain exemplifies the utility of HMMs in decoding hidden states from observed data.

Applications of HMMs are not confined to speech recognition. In video analysis, Boreczky et al. [8] employed HMMs to model temporal dependencies in video sequences, enabling tasks such as action recognition and event detection. In bioinformatics, Haussler et al. [14] used HMMs to analyze DNA sequences, facilitating the identification of gene structures and sequence alignment. Similarly, in financial time series, HMMs have been applied to model latent market states and predict price movements [22]. These examples highlight the versatility of HMMs in addressing diverse real-world problems.

Feature extraction plays a critical role in enhancing the performance of HMMs, particularly when dealing with high-dimensional data. Traditional approaches like principal component analysis (PCA) [43] are easy to implement, but these methods often lack flexibility and scalability. The central idea behind PCA is to reduce the dimensionality of a dataset consisting of a large number of interrelated variables, while retaining as much as

possible the variation presented in the original dataset [44]. However, the naïve application of PCA to high-dimensional objects requires their reshaping into vectors with high dimensionality, which obviously results in high processing costs in terms of computational and memory demands. Beyond implementation issues, reshaping the high-dimensional data also breaks the natural structure inside the original data, such as the spatial information in images and time information in videos. Therefore, it potentially loses more compact and useful representations that can be obtained in the original form [45]. Many variations of PCA have been proposed to overcome these shortcomings, such as representing the data as matrices (e.g., [46]–[48]) or higher-dimensional tensors (e.g., [49]–[51]) instead of column vectors. Nevertheless, due to the linearity essence of PCA, these methods can keep only the most principal linear relation inside the data and discard the non-linear information. This reduction can largely decrease the data complexity in some problems, but for many other scenarios that require more precise data processing, such as radar tracking and satellite observations, non-linear feature extraction methods need to be applied for better performance.

With the rise of machine learning, NNs have become a dominant tool for feature extraction, offering the ability to learn complex nonlinear relationships directly from data. One example is the autoencoder[52], which learns efficient coding to replicate its input data. This model can be applied to many problems , including facial recognition [53], anomaly detection, and learning the meaning of words [54], [55]. Despite these advancements, existing NN-based approaches have limitations when applied to sequential data. Specifically, these methods often prioritize task-specific objectives, such as classification accuracy, over the extraction of general-purpose features. Additionally, they rarely account for prior knowledge or structural constraints, such as symmetry or sparsity, which are vital in certain domains.

One of the successful non-linear feature extraction algorithms is the alternating conditional expectations (ACE) algorithm [56]–[58]. The ACE algorithm uses the Hirschfeld-Gebelein-R'enyi (HGR) maximal correlation [59]–[61] as a metric to evaluate the dependence inside high-dimensional data, and tries to maintain this correlation in the process of dimension reduction as much as possible. One major difference between the ACE algorithm and the PCA methods is that the ACE operates in the space of distributions rather than data. Consequently, "a strong advantage of the ACE procedure is the ability to incorporate variables of quite different type in terms of the set of values they can assume" [62]. Compared with other non-linear methods, the loss function is task-agnostic and can be modified as needed. However, the ACE algorithm considers only the correlation between two consecutive objects in the given dataset and ignores the hidden Markov property appearing in most sequential data. This important property can be used to enhance the performance of the selected features and further reduce the complexity of the algorithm.

This thesis aims to advance the state of the art by proposing algorithms that explicitly address these limitations. Inspired by the ACE algorithm, we leverage principles from information geometry and HGR maximal correlation, the proposed methods enable the extraction of structured and task-agnostic features from high-dimensional sequential data. These con-

tributions are expected to enhance the interpretability and flexibility of feature extraction processes, providing a solid foundation for downstream applications across various domains.

1.4 Outline

Chapter 2 introduces the local information geometry and HGR maximal correlation. This is the theoretical basis for the proposed feature extraction algorithms. Chapter 3 derives three different algorithms and the corresponding NN structures with their advantages and limitations. Chapter 4 analyzes the convergence and several tail bounds for the proposed algorithms. Numerical simulation results of the algorithms can be found in Chapter 5. Chapter 6 concludes this thesis and discusses future directions.

Chapter 2

Preliminaries

2.1 Local Information Geometry

Local information geometry, also known as Euclidean Information Geometry, has been extensively utilized in a series of works [63]–[67] to address problems in network information theory [68]–[70] and to develop novel machine learning algorithms [71]–[76]. For a detailed exposition on local information geometry, we refer the reader to [66]. Here, we introduce some important definitions and theorems to prepare for the proposed algorithms.

We use $\mathcal{F}_{\mathcal{X}} \triangleq \{\mathcal{X} \to \mathbb{R}\}$ to denote the collections of features (functions) on a given $\mathcal{X}$, and $\mathcal{P}_{\mathcal{X}}$ to represent the probability simplex over alphabet $\mathcal{X}$. We establish a correspondence between the distribution space $\mathcal{P}_{\mathcal{X}}$ and the feature space $\mathcal{F}_{\mathcal{X}}$ by the density ratio function.

Definition 2.1.1 (Density Ratio Function)**.** Given some fixed reference distribution $R \in \text{relint}(\mathcal{P}_{\mathcal{X}})$, for any other distribution $P \in \mathcal{P}_{\mathcal{X}}$, we define the (centered) density ratio function $\tilde{l}_{P;R} \in \mathcal{F}_{\mathcal{X}}$ as

$$\tilde{l}_{P;R}(x) = \frac{P(x) - R(x)}{R(x)}, \quad \text{for all } x \in \mathcal{X}. \tag{2.1}$$

Remark 2.1.1. $\tilde{l}_{P;R}$ has zero mean, i.e, $\mathbb{E}_R[\tilde{l}_{P;R}(X)] = 0$. Restricting the features (functions) to be zero-mean is without loss of generality, as there is an invertible mapping between any set of features and their zero-mean counterparts. As a result, we will generally impose this constraint.

For finite data alphabets [1], it can be convenient to introduce the vector and matrix representations of features. Assuming $|\mathcal{X}| < \infty$, we can represent the density ratio function $\tilde{l}_{P;R} \in \mathcal{F}_{\mathcal{X}}$ as an $|\mathcal{X}| \times 1$ vector ϕ.

[1] For the sake of simplicity, we will focus our development on finite alphabets and the associated discrete random variables. However, the corresponding results can be readily extended to general alphabets under certain regularity conditions, More specifically, our development on feature spaces can be extended to general Hilbert spaces [77], [78], where the alphabet $\mathcal{X}$ and the metric distribution $\mathcal{P}_{\mathcal{X}}$ are extended to a measurable set and the measure ([78], Example 6), respectively. For more details, please refer to [75].

Definition 2.1.2 (Information Vector)**.** Given some fixed reference distribution $R \in \text{relint}(\mathcal{P}_\mathcal{X})$, for any other distribution $P \in \mathcal{P}_\mathcal{X}$, we define the information vector ϕ corresponding to the density ratio function $\tilde{l}_{P;R}$ as:

$$\phi(x) = \tilde{l}_{P;R}(x)\sqrt{R(x)} = \frac{P(x) - R(x)}{\sqrt{R(x)}}, \quad \text{for all } x \in \mathcal{X}. \tag{2.2}$$

Remark 2.1.2. $\langle \sqrt{R}, \phi \rangle = 0$.

The benefits to introduce this information vector representation is the following lemma, where it interprets the squared-norm of an information vector in terms of KL divergence with respect to R in a local area ($\|\phi\|_2 \leq \epsilon$) on the (relative interior of the) probability simplex.

Lemma 2.1.1 ([79], Lemma 4.5)**.** If information vector $\|\phi\|_2 \leq \epsilon$, then we have the divergence $D(P\|R) = \frac{1}{2}\|\phi\|_2^2 + o(\epsilon^2)$.

Moreover, we have the following lemma, which follows from the second order Taylor expansion of the Kullback-Leibler (KL) divergence.

Lemma 2.1.2 ([57], Lemma 1)**.** If P_1 and P_2 are two distributions in the neighborhood of the reference distribution P_0 with information vector ϕ_1 and ϕ_2 respectively, where $\|\phi_1\|_2 \leq \epsilon$ and $\|\phi_2\|_2 \leq \epsilon$ hold, then:

$$D(P_1\|P_2) = \frac{1}{2}\|\phi_1 - \phi_2\|_2^2 + o(\epsilon^2)$$

The advantage for this vector space and local concept is that it transforms the asymmetric KL divergence between distributions into a symmetric measure, Euclidean distance, between the corresponding vectors, which we are familiar with. This will bring much convenience and many interesting results.

In the case where the random variables X and Y have the joint distribution $P_{X,Y}$ and reference distribution $R_{X,Y} = P_X P_Y$, the density ratio function can be expressed as,

$$\tilde{l}_{P_{X,Y};R_{X,Y}}(x,y) \triangleq \frac{P_{X,Y}(x,y) - P_X(x)P_Y(y)}{P_X(x)P_Y(y)}, \quad \text{for all } x \in \mathcal{X}, y \in \mathcal{Y}. \tag{2.3}$$

Similarly, for X and Y with $|\mathcal{X}| < \infty, |\mathcal{Y}| < \infty$, we can represent the function $\tilde{l}_{P_{X,Y};P_X P_Y} \in \mathcal{F}_{\mathcal{X}\times\mathcal{Y}}$ as an $|\mathcal{X}| \times |\mathcal{Y}|$ matrix $\tilde{B}_{X;Y}$:

$$\tilde{B}_{X;Y}(x,y) \triangleq \frac{P_{X,Y}(x,y) - P_X(x)P_Y(y)}{\sqrt{P_X(x)}\sqrt{P_Y(y)}}, \quad \text{for all } x \in \mathcal{X}, y \in \mathcal{Y}, \tag{2.4}$$

which is referred to as the canonical dependence matrix (CDM)[2] of X and Y.

[2]We will sometimes refer to this matrix as information matrix in the future due to its importance in capturing the information between X and Y.

Remark 2.1.3. Suppose the SVD for matrix $\tilde{B}_{X;Y}$ takes the form $\tilde{B}_{X;Y} = \sum_{i=1}^{K} \sigma_i \phi_i \psi_i^T$ with $K \triangleq \min\{|\mathcal{X}|, |\mathcal{Y}|\}$, where σ_i denotes the i-th singular value, ϕ_i and ψ_i are the corresponding left and right singular vectors, and by convention we order the singular values according to $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_K$. It can be shown that $\sigma_1 \leq 1, \sigma_K = 0, \phi_K = \sqrt{P_X}$ and $\psi_K = \sqrt{P_Y}$. That is, the CDM has one natural singular value that is equal to zero. In the future sections, when we talk about a full rank, positive definite or rank deficient CDM without further notice, it means that we ignore this zero singular value.

CDM is important because it measures the dependence between X and Y through the following lemma that is parallel to Lemma 2.1.1:

Lemma 2.1.3 (Local Approximation of Mutual Information)**.** If $\|\tilde{B}_{X;Y}\|_F \leq \epsilon$, we have

$$I(X,Y) = \frac{1}{2}\|\tilde{B}_{X;Y}\|_F^2 + o(\epsilon^2).$$

Proof.

$$\begin{aligned}
I(X;Y) &= \sum_{x,y} P_{X,Y}(x,y) \log \frac{P_{X,Y}(x,y)}{P_X(x)P_Y(y)} \\
&= \sum_{x,y} P_X(x)P_Y(y) \log \left(\frac{P_{X,Y}(x,y) - P_X(x)P_Y(y)}{P_X(x)P_Y(y)} + 1 \right) + \\
&\quad \sum_{x,y} \big(P_{X,Y}(x,y) - P_X(x)P_Y(y)\big) \log \left(\frac{P_{X,Y}(x,y) - P_X(x)P_Y(y)}{P_X(x)P_Y(y)} + 1 \right) \\
&= \sum_{x,y} P_X(x)P_Y(y) \left[\frac{P_{X,Y}(x,y) - P_X(x)P_Y(y)}{P_X(x)P_Y(y)} - \frac{1}{2} \left(\frac{P_{X,Y}(x,y) - P_X(x)P_Y(y)}{P_X(x)P_Y(y)} \right)^2 + \right. \\
&\quad \left. o\left(\left(\frac{P_{X,Y}(x,y) - P_X(x)P_Y(y)}{P_X(x)P_Y(y)} \right)^2 \right) \right] + \sum_{x,y} \big(P_{X,Y}(x,y) - P_X(x)P_Y(y)\big) \left[\right. \\
&\quad \left. \frac{P_{X,Y}(x,y) - P_X(x)P_Y(y)}{P_X(x)P_Y(y)} + o\left(\frac{P_{X,Y}(x,y) - P_X(x)P_Y(y)}{P_X(x)P_Y(y)} \right) \right] \qquad (2.5) \\
&= -\frac{1}{2}\|\tilde{B}_{X;Y}\|_F^2 + o(\epsilon^2) + \|\tilde{B}_{X;Y}\|_F^2 + o(\epsilon^2) \qquad (2.6) \\
&= \frac{1}{2}\|\tilde{B}_{X;Y}\|_F^2 + o(\epsilon^2) \qquad (2.7)
\end{aligned}$$

where to obtain (2.5), we use the first and second order Taylor expansion of $\log(1+x)$ when x is small; to obtain (2.6), we use the definition of CDM (2.4); to obtain (2.7), we use the fact that $o(\epsilon^2) + o(\epsilon^2) = o(\epsilon^2)$. □

In practice, the variables of interest typically have unknown and complicated probability distributions, with only data samples available for learning. We can similarly define

the feature geometry on data samples by replacing the distribution with the corresponding empirical ones, and all the lemmas still hold.

2.2 Modal Decomposition

Given $(X, Y) \sim P_{X,Y}$, we consider the space $\mathcal{F}_{\mathcal{X}\times\mathcal{Y}}$ with the reference distribution $R_{X,Y} = P_X P_Y$. We can characterize the statistical dependence between X and Y by the matrix $\tilde{B}_{X;Y}$ defined in (2.4). Specifically, it is easy to verify that $\|\tilde{B}_{X;Y}\|_F = 0$ if and only if X and Y are independent.

Suppose $\operatorname{rank}(\tilde{B}_{X;Y}) = K - 1$ and let the model decomposition be

$$\tilde{B}_{X;Y} = \sum_{i=1}^{K-1} \sigma_i \phi_i \psi_i^T = \sum_{i=1}^{K-1} \sigma_i (\sqrt{P_X} f_i^*)(\sqrt{P_Y} g_i^*)^T, \tag{2.8}$$

where $f_i^* \in \mathcal{F}_{\mathcal{X}}, g_i^* \in \mathcal{F}_{\mathcal{Y}}$ are the feature functions corresponding to the left and right singular vectors ϕ_i and ψ_i, receptively.

In particular, the features f_i^*'s, g_i^*'s are the maximally correlated features in $\mathcal{F}_{\mathcal{X}}, \mathcal{F}_{\mathcal{Y}}$, known as Hirschfeld-Gebelein-Rényi (HGR) maximal correlation functions [59]–[61]. To see this, let us denote the covariance for the given $f \in \mathcal{F}_{\mathcal{X}}, g \in \mathcal{F}_{\mathcal{Y}}$ as

$$\operatorname{cov}(f, g) \triangleq \mathbb{E}_{P_{X,Y}}[f(X)g(Y)] - \mathbb{E}_{P_X P_Y}[f(X)g(Y)]. \tag{2.9}$$

We have the following corollary.

Corollary 2.2.1 (HGR Maximal Correlation Functions)**.** For each $i = 1, \ldots, K-1$, we have $\sigma_i = \operatorname{cov}(f_i^*, g_i^*)$, and $(f_i^*, g_i^*) = \arg\max_{(f_i,g_i)\in\mathcal{D}_i} \operatorname{cov}(f_i, g_i)$, where we have recursively defined $\mathcal{D}_i$ as $\mathcal{D}_i = \big\{(f_i, g_i) \in \mathcal{F}_{\mathcal{X}} \times \mathcal{F}_{\mathcal{Y}} : \mathbb{E}_{P_X}[f_i^2(X)] = \mathbb{E}_{P_Y}[g_i^2(Y)] = 1$, and $\mathbb{E}_{P_X}[f_i(X)f_j(X)] = \mathbb{E}_{P_Y}[g_i(Y)g_j(Y)] = 0$ for all $j \in \{1, \ldots, i-1\}\big\}$.

This corollary, along with Lemma 2.1.3 indicates that we can decompose the dependence between X and Y according to the SVD of $\tilde{B}_{X;Y}$ matrix as $I(X;Y) \approx \frac{1}{2}\|\tilde{B}_{X;Y}\|_F^2 = \frac{1}{2}\sum_i \sigma_i^2$ and $\sigma_i = \operatorname{cov}(f_i^*, g_i^*)$. $\tilde{l}_i^* = \sigma_i f_i^* g_i^*$ captures the i-th dominant dependency structure between X and Y, i.e., $\tilde{l}_{P_{X,Y};R_{X,Y}}(x, y) = \sum_{i=1}^{K-1} \tilde{l}_i^*(x, y) = \sum_{i=1}^{K-1} \sigma_i f_i^*(x) g_i^*(y)$. Therefore, if we want to extract the i-th most informative features from the data, we should refer to f_i^* and g_i^*.

In other words, finding the "optimal" features is equivalent to computing the first few singular vectors of $\tilde{B}_{X;Y}$. In the case where $|\mathcal{X}|$ and $|\mathcal{Y}|$ are very large, computing the SVD of $\tilde{B}_{X;Y}$ is often a formidable task. Fortunately, this is a well-known problem in the literature, and a standard solution is the Alternating Conditional Expectations (ACE) algorithm [56], which is derived from the power method in numerical linear algebra [80]. More specifically, with the knowledge about $P_{X,Y}$, we can use the algorithm [57] in Algorithm 1 to compute the top feature pairs. When we have training samples of (x_i, y_i) instead of the population distribution, Algorithm 2 is used in practical.

Algorithm 1 ACE Algorithm

Require: Knowledge of $P_{X,Y}$
  Initialize: randomly pick $g(y), y \in \mathcal{Y}$
  Center: $g(y) \longleftarrow g(y) - \mathbb{E}[g(Y)]$
  repeat
    $f(x) \longleftarrow \mathbb{E}[g(Y)|X = x], \forall x \in \mathcal{X}$
    $g(y) \longleftarrow \mathbb{E}[f(X)|Y = y], \forall y \in \mathcal{Y}$
    Regularize: $g(y) \longleftarrow g(y)/\mathbb{E}[g^2(y)], \forall y \in \mathcal{Y}$
  until $\mathbb{E}[f(X)g(Y)]$ stops to increase

Algorithm 2 ACE Algorithm with Finite Samples

Require: Training samples $\{(x_i, y_i) : i = 1, ..., n\}$
  Initialize: randomly pick $g(y), y \in \mathcal{Y}$
  repeat
    $f(x) \longleftarrow \hat{\mathbb{E}}[g(Y)|X = x], \forall x \in \mathcal{X}$
    $g(y) \longleftarrow \hat{\mathbb{E}}[f(X)|Y = y], \forall y \in \mathcal{Y}$
    Regularize:
      $g(y) \longleftarrow g(y) - \hat{\mathbb{E}}[g(Y)], \forall y \in \mathcal{Y}$
      $g(y) \longleftarrow g(y)/\hat{\mathbb{E}}[g^2(y)], \forall y \in \mathcal{Y}$
  until $\hat{\mathbb{E}}[f(X)g(Y)]$ stops to increase

We could also obtain the top features through low-rank approximation point of view. The problem thus can be formulated as follows:

$$\begin{aligned} \min_{f,g} \quad & \|\tilde{B}_{X;Y} - \phi\psi^T\|_F^2 \\ s.t. \quad & \phi(x) = f(x)\sqrt{P(x)}, \quad \forall x \in \mathcal{X} \\ & \psi(y) = g(y)\sqrt{P(y)}, \quad \forall y \in \mathcal{Y} \end{aligned} \tag{2.10}$$

Focusing on the objective function, we have

$$\|\tilde{B}_{X;Y} - \phi\psi^T\| = -2\underbrace{\phi^T\tilde{B}_{X;Y}\psi}_{\mathbb{E}[f(X)g(Y)]} + \underbrace{\|\psi\|_2^2}_{\text{var}(f(X))} \cdot \underbrace{\|\phi\|_2^2}_{\text{var}(g(Y))} + \text{const},$$

and therefore, ignoring the constant terms and scale the objective function, we have

$$\mathscr{H}(f, g) \triangleq -\mathbb{E}[f(X)g(Y)] + \frac{1}{2}\text{var}(f(X))\text{var}(g(Y)), \tag{2.11}$$

which we call negative H-score ($\mathscr{H}$). Since this objective function only requires the data $(X, Y) \sim P_{X,Y}$, we could use the NN structure in Figure 2.1 to obtain the feature pairs f and g, where the loss function, $\mathcal{L} = \mathscr{H}(f, g)$, is given in (2.11). In this example, the data points for X and Y are sent into the NN from each end, and then pass through the NN

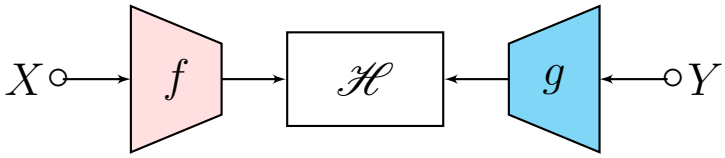


Figure 2.1: NN structure to extract the top feature pairs for X and Y.

structure that represents the function of f and g, respectively. The outputs for the f and g networks are both one-dimensional so that the top feature pair is selected. The algorithm for this method is given in Algorithm 3.

Algorithm 3 Negative H-score on a mini-batch to extract the top feature pair

Require: Training samples $\{(x_i, y_i) : i = 1, ..., m\}$ in a mini-batch of size m

Require: Two branches of parameterized NNs with one output units: f and g

1. Center:
   $f(x_i) \longleftarrow f(x_i) - \frac{1}{m}\sum_{j=1}^{m} f(x_j), i = 1, \ldots, m$
   $g(y_i) \longleftarrow g(y_i) - \frac{1}{m}\sum_{j=1}^{m} g(y_j), i = 1, \ldots, m$
2. Compute the empirical variance:
   $\text{var}(f) \longleftarrow \frac{1}{m-1}\sum_{i=1}^{m} f^2(x_i), i = 1, \ldots, m$
   $\text{var}(g) \longleftarrow \frac{1}{m-1}\sum_{i=1}^{m} g^2(y_i), i = 1, \ldots, m$
3. Compute the empirical objective $\mathscr{H}$:
   $-\frac{1}{m}\sum_{i=1}^{m} f(x_i)g(y_i) + \frac{1}{2}\text{var}(f)\text{var}(g)$

Compared with the previous algorithms, this model is more flexible in selecting the form of feature functions f and g using parametric NN. More specifically, it can utilize the powerful NN structure to represent different non-linear functions. For example, we can use a convolutional NN (CNN) [81]–[86] to deal with image or video processing problem, and Long Short-Term Memory (LSTM) [87]–[92] model for natural language processing problem. As long as the reachable functional space of the neural structures covers the optimal feature transformation, the algorithm and the HGR maximal correlation will always lead us to the equivalent solution.

This NN structure can be generalized to extract top k feature pairs as shwon in Figure 2.2. This time, the output for the f and g networks are both k-dimensional, and the objective function can be written as

$$\mathscr{H}(f, g) = -\mathbb{E}[f(X)^T g(Y)] + \frac{1}{2}\text{tr}(\text{cov}(f(X))\text{cov}(g(Y))).$$

The detailed algorithm is given in Algorithm 4.

One advantage to note here is that we do not need to decorrelate the functions of f_i and f_j (resp. for g) for $i \neq j$ as required by Corollary 2.2.1 since the negative H-score escapes whitening constraints but still arrives at an equivalent optimum.

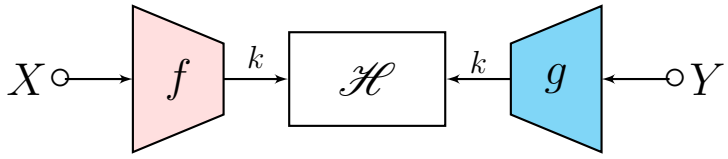


Figure 2.2: NN structure to extract the top k feature pairs for X and Y.

Algorithm 4 Negative H-score on a mini-batch to extract the top k feature pairs

Require: Training samples $\{(x_i, y_i) : i = 1, ..., m\}$ in a mini-batch of size m

Require: Two branches of parameterized NNs with k output units: f and g

1. Center:
   $f(x_i) \longleftarrow f(x_i) - \frac{1}{m}\sum_{j=1}^{m} f(x_j), i = 1, \ldots, m$
   $g(y_i) \longleftarrow g(y_i) - \frac{1}{m}\sum_{j=1}^{m} g(y_j), i = 1, \ldots, m$
2. Compute the empirical covariance:
   $\text{cov}(f) \longleftarrow \frac{1}{m-1}\sum_{i=1}^{m} f(x_i)f^T(x_i), i = 1, \ldots, m$
   $\text{cov}(g) \longleftarrow \frac{1}{m-1}\sum_{i=1}^{m} g(y_i)g^T(y_i), i = 1, \ldots, m$
3. Compute the empirical objective $\mathscr{H}$:
   $-\frac{1}{m}\sum_{i=1}^{m} f^T(x_i)g(y_i) + \frac{1}{2}\text{tr}(\text{cov}(f)\text{cov}(g))$

Chapter 3

Neural Networks Structures for Hidden Markov Model

In a typical feature extraction scenario, high-dimensional data $Y = y$ are observed while it is hard to directly make use of them in various tasks due to the curse of dimensionality. Suppose $X = x$ is a low-dimensional representation for each observation. We are interested in extracting the information of X from Y to make further inference problems possible. The relationship between the observation Y and its underlying label X can be regarded as a hidden Markov chain as shown in Fig. 3.1.

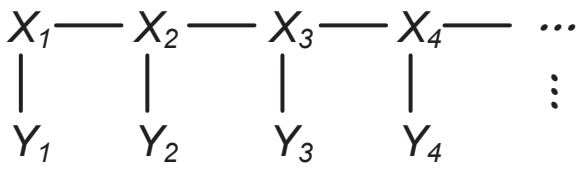


Figure 3.1: Hidden Markov Model (HMM)

Since the observations of Y usually come from the same source or are generated in a similar way, e.g., Y represents each frame in a video game or a single pixel in a particular image, it is reasonable to assume that the representation map between X and Y is fixed, i.e., $P_{Y_1|X_1} = P_{Y_2|X_2} = \cdots$. From Chapter 2, the best feature functions for Y can be obtained through the CDM $\tilde{B}_{X;Y}$, or more specifically, the row space of $\tilde{B}_{X;Y}$. Therefore, we will focus on deriving this information vector space in the following. To make it simple, we formulate the problem as a homogeneous hidden Markov model (HMM) with a fixed representation map $P_{Y|X} = P_{Y_1|X_1} = P_{Y_2|X_2}$ so that the probelm can be reduced to a two-tap HMM as shown in Figure 3.2. The probability distribution for this HMM is

$$P_{Y_1,Y_2}(y_1, y_2) = \sum_{x_1,x_2} P_{X_1,X_2}(x_1, x_2) P_{Y|X}(y_1|x_1) P_{Y|X}(y_2|x_2) \tag{3.1}$$

or equivalently in a matrix form,

$$P_{Y_1,Y_2} = P_{Y|X} \cdot P_{X_1,X_2} \cdot P_{Y|X}^T \tag{3.2}$$

Corollary 3.0.1. The CDM relationship corresponding to (3.2) with the fully independent distribution as the reference can be written as

$$\tilde{B}_{Y_1;Y_2} = \tilde{B}_{X;Y}^T \cdot \tilde{B}_{X_1;X_2} \cdot \tilde{B}_{X;Y}$$

Proof. By definition, in order to prove Corollary 3.0.1, it is equivalent to show that

$$\frac{P_{Y_1,Y_2}(y_1,y_2) - P_Y(y_1)P_Y(y_2)}{\sqrt{P_Y(y_1)P_Y(y_2)}} =$$
$$\sum_{x_1,x_2} \frac{P_{X,Y}(x_1,y_1) - P_X(x_1)P_Y(y_1)}{\sqrt{P_X(x_1)P_Y(y_1)}} \frac{P_{X_1,X_2}(x_1,x_2) - P_X(x_1)P_X(x_2)}{\sqrt{P_X(x_1)P_X(x_2)}} \frac{P_{X,Y}(x_2,y_2) - P_X(x_2)P_Y(y_2)}{\sqrt{P_X(x_2)P_Y(y_2)}},$$

which is equivalent to

$$P_{Y_1,Y_2}(y_1,y_2) - P_Y(y_1)P_Y(y_2) =$$
$$\sum_{x_1,x_2} \frac{P_{X,Y}(x_1,y_1) - P_X(x_1)P_Y(y_1)}{P_X(x_1)} \big(P_{X_1,X_2}(x_1,x_2) - P_X(x_1)P_X(x_2)\big) \frac{P_{X,Y}(x_2,y_2) - P_X(x_2)P_Y(y_2)}{P_X(x_2)} \tag{3.3}$$

Notice that the RHS of (3.3) is equivalent to

$$\sum_{x_1,x_2} \big(P_{Y|X}(y_1|x_1) - P_Y(y_1)\big)\big(P_{X_1,X_2}(x_1,x_2) - P_X(x_1)P_X(x_2)\big)\big(P_{Y|X}(y_2|x_2) - P_Y(y_2)\big),$$

and therefore, we need to prove

$$P_{Y_1,Y_2}(y_1,y_2) - P_Y(y_1)P_Y(y_2) =$$
$$\sum_{x_1,x_2} \big(P_{Y|X}(y_1|x_1) - P_Y(y_1)\big)\big(P_{X_1,X_2}(x_1,x_2) - P_X(x_1)P_X(x_2)\big)\big(P_{Y|X}(y_2|x_2) - P_Y(y_2)\big)$$

This equation holds by nature because of (3.1) and the following equations

$$\begin{aligned}
&\sum_{x_1,x_2} P_{Y|X}(y_1|x_1)P_{X_1,X_2}(x_1,x_2)P_Y(y_2) = \sum_{x_1,x_2} P_{Y|X}(y_1|x_1)P_X(x_1)P_X(x_2)P_{Y|X}(y_2|x_2) \\
=&\sum_{x_1,x_2} P_{Y|X}(y_1|x_1)P_X(x_1)P_X(x_2)P_Y(y_2) = \sum_{x_1,x_2} P_Y(y_1)P_{X_1,X_2}(x_1,x_2)P_{Y|X}(y_2|x_2) \\
=&\sum_{x_1,x_2} P_Y(y_1)P_{X_1,X_2}(x_1,x_2)P_Y(y_2) = \sum_{x_1,x_2} P_Y(y_1)P_X(x_1)P_X(x_2)P_{Y|X}(y_2|x_2) \\
=&\sum_{x_1,x_2} P_Y(y_1)P_X(x_1)P_X(x_2)P_Y(y_2) = P_Y(y_1)P_Y(y_2)
\end{aligned}$$

□

Notice the difference in this model compared to that in Chapter 2 is that only observations of Y are provided, which makes the representation map $\tilde{B}_{X;Y}$ harder to obtain.

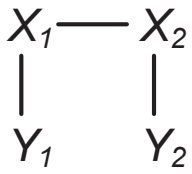


Figure 3.2: Two-tap HMM

Following the same idea of low-rank approximation, let $\check{B}_{Y_1;Y_2}$ to be the empirical version of $\tilde{B}_{Y_1;Y_2}$ matrix calculated from the samples. We solve

$$\begin{aligned}
\min_{\tilde{B}_{X;Y},\tilde{B}_{X_1;X_2}} \quad & \|\check{B}_{Y_1;Y_2} - \tilde{B}_{X;Y}^T \cdot \tilde{B}_{X_1;X_2} \cdot \tilde{B}_{X;Y}\|_F^2 \\
s.t. \quad & \tilde{B}_{X;Y}, \tilde{B}_{X_1;X_2} \in CDM
\end{aligned} \tag{3.4}$$

Notice here the vector space of $\tilde{B}_{Y_1;Y_2}$ may be different from that of $\tilde{B}_{X;Y}$ due to the existence of $\tilde{B}_{X_1;X_2}$. If $\tilde{B}_{X_1;X_2}$ is rank deficient, some modes will be killed. However, this is the best we can do given the observations of Ys. In the following, we will show how our approach approximates the row space of $\tilde{B}_{X;Y}$ in different cases and examine the associated trade-offs.

Before we dive into solving the problem (3.4), we propose another way of formulating this feature extraction problem. Suppose we want to find $P_{Y|X}$ and P_{X_1,X_2} to fit the model

(3.2) with observations Y. This is a maximum likelihood (ML) problem:

$$\begin{aligned}
\hat{P}_{Y|X}, \hat{P}_{X_1,X_2} &= \underset{P_{Y|X}, P_{X_1,X_2}}{\arg\max} \ \log\Big(\prod_{i=1}^{N} P_{Y_1,Y_2}(y_1^{(i)}, y_2^{(i)})\Big) \\
&= \underset{P_{Y|X}, P_{X_1,X_2}}{\arg\max} \ \frac{1}{N}\sum_{i=1}^{N} \log P_{Y_1,Y_2}(y_1^{(i)}, y_2^{(i)}) \\
&= \underset{P_{Y|X}, P_{X_1,X_2}}{\arg\max} \ \mathbb{E}_{\check{P}_{Y_1,Y_2}}\big[\log P_{Y_1,Y_2}\big] \\
&= \underset{P_{Y|X}, P_{X_1,X_2}}{\arg\min} \ D\big(\check{P}_{Y_1,Y_2} \| P_{Y_1Y_2}\big)
\end{aligned}$$

where $(y_1^{(i)}, y_2^{(i)})$ is the i-th pair of observation of (Y_1, Y_2), $\check{P}_{Y_1,Y_2}$ is the empirical P_{Y_1,Y_2} distribution calculated from observed samples, and $\hat{P}_{Y|X}$ and $\hat{P}_{X_1,X_2}$ are the estimations of $P_{Y|X}$ and P_{X_1,X_2}, respectively. Plug (3.2) into this optimization problem:

$$\min_{P_{YX}, P_{X_1X_2}} D\big(\check{P}_{Y_1Y_2} \| P_{Y|X} \cdot P_{X_1,X_2} \cdot P_{Y|X}^T\big) \tag{3.5}$$

Using a local notation similar to Lemma 2.1.2, KL divergence between distributions can be transformed into distances between information matrix space measured by Frobenius norm. Therefore, the optimization problem (3.4) is equivalent to the optimization problem (3.5), which should be no surprise due to the fact that the information in probability distribution space is the same as the information in information matrix space.

3.1 Difficulties for the HMM Problem

As discussed so far, the key is to figure out the row space of $B_{X;Y}$ through Problem (3.4) so that we could construct the features accordingly. If we turn this optimization problem into a more abstract form, we have

$$\min_{P \in \mathbb{R}^{N\times k}, M \in \mathbb{R}^{k\times k}} \ \|B - PMP^T\|_F^2, \tag{3.6}$$

where $k \leq N$ for some matrix B. This is not in the form of any well-known matrix decompositions, for example, eigenvalue decomposition (EVD), singular value decomposition (SVD), LU decomposition, or QR decomposition. In the Appendix A, we provide the mathematical solution to this problem in some special cases. However, in general, there is no closed-form solution. Therefore, in next sections, we will revise the optimization target a little to make it easier to analyze.

3.2 First Neural Networks Structure

If we only care about the row space of $\tilde{B}_{X;Y}$, we can relax the Problem 3.4 into the following form:

$$\min_{\Phi,\Psi\in\mathbb{R}^{N\times k}} \quad \|\check{B}_{Y_1;Y_2} - \Phi\Psi^T\|_F^2, \tag{3.7}$$

where $N = |\mathcal{Y}|$ and k is the number of selected features.

In order to solve this problem, we introduce the following theorem:

Theorem 3.2.1 (Eckart-Young-Mirsky Theorem [93])**.** Suppose $A = U\Sigma V^T$, then $A_r = U_{1:r}\Sigma_{1:r}V_{1:r}^T = \sum_{i=1}^r \sigma_i u_i v_i^T$ is the optimal solution to the following low-rank approximation problem:

$$\begin{aligned} \min_{A_r} \quad & \|A - A_r\|_F^2 \\ s.t. \quad & \text{rank}(A_r) \leq r \end{aligned} \tag{3.8}$$

Therefore, suppose $\check{B}_{Y_1;Y_2} = U\Sigma V^T$ is the SVD for $\check{B}_{Y_1;Y_2}$ matrix, then the optimal Φ^* and Ψ^* should follow:

$$\Phi^*\Psi^{*T} = \sum_{i=1}^k \sigma_i u_i v_i^T = U_{1:k}\Sigma_{1:k}V_{1:k}^T$$

The optimal Φ^* and Ψ^* are not unique. Given any constant decomposition of $\Sigma_{1:k} = H_1H_2^T$, where H_1 and H_2 are full rank, there is an associated solution $\Phi^* = U_{1:k}H_1, \Psi_* = V_{1:k}H_2$. However, the appearance of H_1 will not change the vector space Φ^* lies. Φ^* should be in the same column space as $\check{B}_{Y_1;Y_2}$, so close to the column space of $\tilde{B}_{Y_1;Y_2}$ as we hoped.

Notice that this new problem is the generalized version of Problem (2.10). Therefore, following the same derivation, we have

$$\|\check{B}_{Y_1;Y_2} - \Phi\Psi^T\| = -2\underbrace{\text{tr}(\Phi^T\check{B}_{Y_1;Y_2}\Psi)}_{\mathbb{E}[f(Y_1)^Tg(Y_2)]} + \text{tr}(\underbrace{\Phi^T\Phi}_{\text{cov}(f(Y_1))}\underbrace{\Psi^T\Psi}_{\text{cov}(g(Y_2))}) + \text{const}$$

This leads to the following objective function

$$\mathscr{H}(f,g) = -\mathbb{E}[f(Y_1)^Tg(Y_2)] + \frac{1}{2}\text{tr}(\text{cov}(f(Y_1))\text{cov}(g(Y_2))), \tag{3.9}$$

and the neural network structure given in Figure 3.3. The corresponding algorithm is given in Algorithm 5.

Remember we mention that Φ^* and Ψ^* are not unique, so are f^* and g^*. Equivalent expressions for f^* and g^* are:

$$\begin{aligned} f_i^*(y_1) &= [U_{1:k}H_1]_{y_1,i}/\sqrt{P_Y(y_1)}, i = 1,\ldots,k, y_1 \in \mathcal{Y} \\ g_i^*(y_2) &= [V_{1:k}H_2]_{y_2,i}/\sqrt{P_Y(y_2)}, i = 1,\ldots,k, y_2 \in \mathcal{Y} \end{aligned}$$

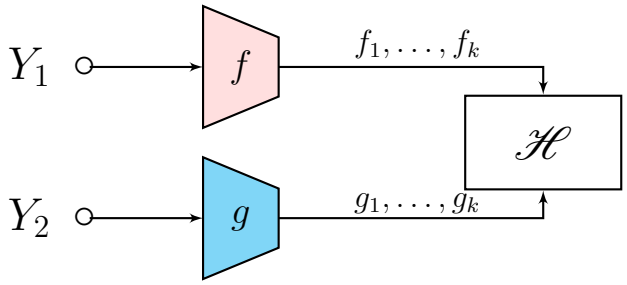


Figure 3.3: NN structure to extract the top k feature pairs for HMM.

Algorithm 5 Negative H-score on a mini-batch to extract the top k feature pairs for HMM

Require: Training samples $\{(y_1^{(i)}, y_2^{(i)}) : i = 1, ..., m\}$ in a mini-batch of size m
Require: Two branches of parameterized NNs with k output units: f and g

1. Center:
   $f(y_1^{(i)}) \longleftarrow f(y_1^{(i)}) - \frac{1}{m}\sum_{j=1}^m f(y_1^{(j)}), i = 1, \ldots, m$
   $g(y_2^{(i)}) \longleftarrow g(y_2^{(i)}) - \frac{1}{m}\sum_{j=1}^m g(y_2^{(j)}), i = 1, \ldots, m$
2. Compute the empirical covariance:
   $\mathrm{cov}(f) \longleftarrow \frac{1}{m-1}\sum_{i=1}^m f(y_1^{(i)})f^T(y_1^{(i)}), i = 1, \ldots, m$
   $\mathrm{cov}(g) \longleftarrow \frac{1}{m-1}\sum_{i=1}^m g(y_2^{(i)})g^T(y_2^{(i)}), i = 1, \ldots, m$
3. Compute the empirical objective $\mathscr{H}$:
   $-\frac{1}{m}\sum_{i=1}^m f^T(y_1^{(i)})g(y_2^{(i)}) + \frac{1}{2}\mathrm{tr}(\mathrm{cov}(f)\mathrm{cov}(g))$

Since H_1 and H_2 are invertible, one can conclude that the optimal features obtained through Algorithm 5 are linearly transformable from the one to the other. Namely, $\mathrm{span}\{f_1^*, \ldots, f_k^*\}$ (resp. for g) spans the same feature space for different H_1, H_2, and therefore describes the same amount of information. One way to understand this equivalence is to image that the features are feed into an extra linear dense layer of dimension $k \times k$ and output the linearly transformable result.

Compared with the structure shown in Figure 2.2, we only need to replace X with Y_1 and Y with Y_2 to obtain Figure 3.3. This NN solves the problem in (3.7), and as a result, it will recover the row space of $\tilde{B}_{X;Y}$ when the number of selected features k is large enough and $\tilde{B}_{X_1;X_2}$ does not kill any mode, i.e., $\mathrm{rank}(\tilde{B}_{X_1;X_2}) = |\mathcal{X}| - 1$.

3.3 Second Neural Networks Structure

Because of the special structure of the homogeneous HMM, we are trying to exploit its property to make the algorithm more efficient. One thing to note is the fixed representation map, i.e., $P_{Y_1|X_1} = P_{Y_2|X_2}$. This inspires us instead of using two different networks f and g in the NN structure, we can use the same network repeatedly to save some computational complexity. Therefore, a second NN structure is proposed in Figure 3.4, and the corresponding algorithm is given in Algorithm 6. In this case, the objective function reduces to

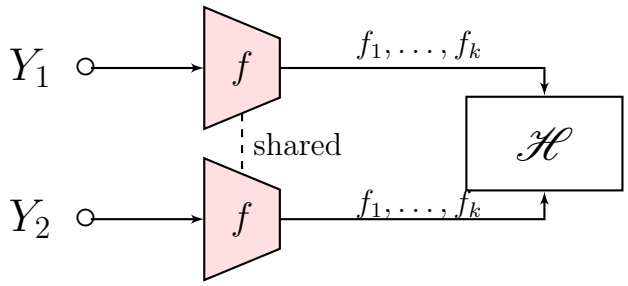


Figure 3.4: Symmetric NN structure to extract the top k feature pairs for HMM.

Algorithm 6 Symmetric negative H-score on a mini-batch to extract the top k feature pairs for HMM

Require: Training samples $\{(y_1^{(i)}, y_2^{(i)}) : i = 1, ..., m\}$ in a mini-batch of size m

Require: Two branches of parameterized NNs with k output units: both branches are f

1. Center:
   $f(y_1^{(i)}) \longleftarrow f(y_1^{(i)}) - \frac{1}{m}\sum_{j=1}^{m} f(y_1^{(j)}), i = 1, \ldots, m$
   $f(y_2^{(i)}) \longleftarrow f(y_2^{(i)}) - \frac{1}{m}\sum_{j=1}^{m} f(y_2^{(j)}), i = 1, \ldots, m$
2. Compute the empirical covariance:
   $\mathrm{cov}(f(Y_1)) \longleftarrow \frac{1}{m-1}\sum_{i=1}^{m} f(y_1^{(i)}) f^T(y_1^{(i)}), i = 1, \ldots, m$
   $\mathrm{cov}(f(Y_2)) \longleftarrow \frac{1}{m-1}\sum_{i=1}^{m} f(y_2^{(i)}) f^T(y_2^{(i)}), i = 1, \ldots, m$
3. Compute the empirical objective $\mathscr{H}$:
   $-\frac{1}{m}\sum_{i=1}^{m} f^T(y_1^{(i)}) f(y_2^{(i)}) + \frac{1}{2}\mathrm{tr}(\mathrm{cov}(f(Y_1))\mathrm{cov}(f(Y_2)))$

$$\mathscr{H}(f) = -\mathbb{E}[f(Y_1)^T f(Y_2)] + \frac{1}{2}\mathrm{tr}(\mathrm{cov}(f(Y_1))\mathrm{cov}(f(Y_2))), \tag{3.10}$$

and correspondingly, we can map the function space back to the information matrix space:

$$\|\check{B}_{Y_1;Y_2} - \Phi\Phi^T\| = -2\mathbb{E}[f(Y_1)^T f(Y_2)] + \mathrm{tr}(\mathrm{cov}(f(Y_1))\mathrm{cov}(f(Y_2))) + \mathrm{const.}$$

Therefore, this NN tries to solve the following optimization problem:

$$\min_{\Phi\in\mathbb{R}^{N\times k}} \quad \|\check{B}_{Y_1;Y_2} - \Phi\Phi^T\|_F^2, \tag{3.11}$$

where N and k are defined as before.

If we turn this optimization problem into a more abstract form, we have

$$\min_{P\in\mathbb{R}^{N\times k}} \quad \|B - PP^T\|_F^2, \tag{3.12}$$

where $k \leq N$ for some matrix B. In the following, we will prove that solving the Problem (3.12) is equivalent to solving the following problem in terms of the vector space for the optimal matrix P:

$$\min_{P\in\mathbb{R}^{N\times k}} \quad \|B^{\mathrm{pd}} - PP^T\|_F^2, \tag{3.13}$$

where B^{pd} represents the positive definite part of the matrix B. It is obtained by doing EVD for the symmetric matrix $\frac{B+B^T}{2}$ and setting all the negative eigenvalues to 0. Correpondingly, we use B^{nd} to represent the negative definite part of the matrix B, which is obtained by doing EVD for the symmetric matrix $\frac{B+B^T}{2}$ and setting all the positive eigenvalues to 0. Specifically, $B^{\mathrm{pd}} + B^{\mathrm{nd}} = \frac{B+B^T}{2}$.

Proof.

$$\begin{aligned}
& \arg\min_{P\in\mathbb{R}^{N\times k}} \|B - PP^T\|_F^2 \\
= \ & \arg\min_{P\in\mathbb{R}^{N\times k}} \|\frac{B+B^T}{2} + \frac{B-B^T}{2} - PP^T\|_F^2 \\
= \ & \arg\min_{P\in\mathbb{R}^{N\times k}} \|\frac{B+B^T}{2} - PP^T\|_F^2 + \|\frac{B-B^T}{2}\|_F^2 && (3.14) \\
= \ & \arg\min_{P\in\mathbb{R}^{N\times k}} \|\frac{B+B^T}{2} - PP^T\|_F^2 && (3.15) \\
= \ & \arg\min_{P\in\mathbb{R}^{N\times k}} \|B^{\mathrm{pd}} + B^{\mathrm{nd}} - PP^T\|_F^2 \\
= \ & \arg\min_{P\in\mathbb{R}^{N\times k}} \|B^{\mathrm{pd}} - PP^T\|_F^2 + \|B^{\mathrm{nd}}\|_F^2 && (3.16) \\
= \ & \arg\min_{P\in\mathbb{R}^{N\times k}} \|B^{\mathrm{pd}} - PP^T\|_F^2 && (3.17)
\end{aligned}$$

where (3.14) is obtained because the symmetric and anti-symmetric matrices are orthogonal in terms of Frobenius norm, i.e., $\|A+B\|_F^2 = \|A\|_F^2 + \|B\|_F^2$ when $A^T = A$ and $B^T = -B$; (3.15) is obtained because $\|\frac{B-B^T}{2}\|_F^2$ is irrelevant with respect to P; (3.16) is obtained because the positive definite and negative definite matrices are orthogonal in terms of Frobenius norm, i.e., $\|A+B\|_F^2 = \|A\|_F^2 + \|B\|_F^2$ when A is a positive definite matrix and B is a negative definite matrix; (3.17) is obtained because $\|B^{\text{nd}}\|_F^2$ is irrelevant with respect to P. □

Therefore, the second structure tries to recover the column space of the positive definite part of $\check{B}_{Y_1;Y_2}$. Compared with the NN structure in Figure 3.3, it saves from learning only half of the parameters in the network; it will recover the row space of $\tilde{B}_{X;Y}$ when the number of selected features k is large enough and $\tilde{B}_{X_1;X_2}$ is a positive definite (PD) matrix, i.e, $\text{rank}(\tilde{B}^{\text{pd}}_{X_1;X_2}) = |\mathcal{X}| - 1$.

3.4 Third Neural Networks Structure

It is nice to utilize the symmetric property hidden in the model. However, the second NN structure seems to be too simple to capture all valuable information - only the positive definite part of the matrix is effective. Encouraged by the second structure, we further dig into the symmetric property and introduce a third structure based on that.

Here, we consider a symmetric problem by replacing everything except $\tilde{B}_{X;Y}$ in Problem (3.4) with its symmetric version:

$$\begin{aligned} \min_{\tilde{B}_{X;Y}, \tilde{B}^{\text{sym}}_{X_1;X_2}} \quad & \|\check{B}^{\text{sym}}_{Y_1;Y_2} - \tilde{B}^T_{X;Y} \cdot \tilde{B}^{\text{sym}}_{X_1;X_2} \cdot \tilde{B}_{X;Y}\|_F^2 \\ s.t. \quad & \tilde{B}_{X;Y}, \tilde{B}^{\text{sym}}_{X_1;X_2} \in CDM \end{aligned} \tag{3.18}$$

where $\tilde{B}^{\text{sym}}_{X_1;X_2} = \frac{\tilde{B}_{X_1;X_2} + \tilde{B}^T_{X_1;X_2}}{2}$ and $\check{B}^{\text{sym}}_{Y_1;Y_2} = \frac{\check{B}_{Y_1;Y_2} + \check{B}^T_{Y_1;Y_2}}{2}$.

This is also a valid formulation because we have

$$\frac{P_{Y_1Y_2} + P^T_{Y_1Y_2}}{2} = P_{Y|X} \cdot \frac{P_{X_1X_2} + P^T_{X_1X_2}}{2} \cdot P^T_{Y|X}$$

from the model, and thus the corresponding CDM relationship holds:

$$\tilde{B}^{\text{sym}}_{Y_1;Y_2} = \tilde{B}^T_{X;Y} \cdot \tilde{B}^{\text{sym}}_{X_1;X_2} \cdot \tilde{B}_{X;Y}$$

If we turn the optimization problem (3.18) into a more abstract form, we have

$$\min_{P \in \mathbb{R}^{N\times k}, M \in \mathbb{S}^{k\times k}} \quad \|\frac{B+B^T}{2} - PMP^T\|_F^2, \tag{3.19}$$

where $k \leq N$, and $\mathbb{S}^{k\times k}$ represents the set of all $k \times k$ symmetric matrices. In the following, we will prove that solving the Problem (3.19) is equivalent to solving the following problem

in terms of the vector space for the optimal matrix P:

$$\min_{P\in\mathbb{R}^{N\times k}, D\in diag(k)} \quad \|B - PDP^T\|_F^2, \tag{3.20}$$

where $diag(k)$ represents the set of all $k \times k$ diagonal matrices. Actually, we will show that D could be further constrained to be a diagonal matrix whose diagonal values are either 1 or -1.

As pointed out, $\frac{B+B^T}{2}$ is a symmetric matrix. Considering the problem of finding a low-rank approximation to $\frac{B+B^T}{2}$, i.e.,

$$\min_{P\in\mathbb{R}^{N\times k}, Q\in\mathbb{R}^{N\times k}} \quad \|\frac{B+B^T}{2} - PQ^T\|_F^2. \tag{3.21}$$

Since the set of $\{PQ^T\}$ in problem (3.21) - all $N \times N$ matrices whose rank is less than or equal to k, is larger than that of $\{PMP^T\}$ in problem (3.19) - all $N \times N$ symmetric matrices whose rank is less than or equal to k, so the optimal value in (3.21) should be smaller than that in (3.19). According to Theorem 3.2.1, the optimal PQ^T to (3.21) should be the rank-k truncated SVD of $\frac{B+B^T}{2}$, which has the form of PMP^T (actually PMP^T can be regarded as EVD since we allow the elements of M to be negative). Therefore, the solution to problem (3.19) is also the rank-k truncated SVD of $\frac{B+B^T}{2}$. Moreover, the solution to this problem is equivalent to that to (3.20) according to the following proof.

Proof.

$$\begin{aligned}
&\arg\min_{P\in\mathbb{R}^{N\times k}, D\in diag(k)} \|B - PDP^T\|_F^2 \\
=&\arg\min_{P\in\mathbb{R}^{N\times k}, D\in diag(k)} \|\frac{B+B^T}{2} + \frac{B-B^T}{2} - PDP^T\|_F^2 \\
=&\arg\min_{P\in\mathbb{R}^{N\times k}, D\in diag(k)} \|\frac{B+B^T}{2} - PDP^T\|_F^2 + \|\frac{B-B^T}{2}\|_F^2 \\
=&\arg\min_{P\in\mathbb{R}^{N\times k}, D\in diag(k)} \|\frac{B+B^T}{2} - PDP^T\|_F^2
\end{aligned}$$

Again, the set of $\{PQ^T\}$ in Problem (3.21) - all $N \times N$ matrices whose rank is less than or equal to k, is larger than that of $\{PDP^T\}$ in Problem (3.20) - all $N \times N$ symmetric matrices whose rank is less than or equal to k, so the optimal value in (3.21) should be smaller than that in (3.20). The optimal PQ^T to (3.21) should be the rank-k truncated SVD of $\frac{B+B^T}{2}$, which has the form of PDP^T. Therefore, the solution to Problem (3.20) is also the rank-k truncated SVD of $\frac{B+B^T}{2}$, and thus equivalent to problem (3.19). We further note here that since the optimal P and D is not unique, we can absorb the diagonals of D into P so that the diagonal values of D to be either 1 or -1. □

Therefore, we can reformulate the problem in the following form:

$$\min_{\Phi\in\mathbb{R}^{N\times k}, D\in diag(k)} \quad \|\check{B}_{Y_1;Y_2} - \Phi D\Phi^T\|_F^2. \tag{3.22}$$

The optimal Φ^* should be in the same column space as $\check{B}^{\text{sym}}_{Y_1;Y_2}$, so close to the column space of $\tilde{B}^{sym}_{Y_1;Y_2}$,which is the same as $\tilde{B}_{Y_1;Y_2}$ as we hoped.

Following similar steps, we have

$$\|\check{B}_{Y_1;Y_2} - \Phi D\Phi^T\| = -2\underbrace{\text{tr}(D\Phi^T\check{B}_{Y_1;Y_2}\Phi)}_{\mathbb{E}[f(Y_1)^T g(Y_2)]} + \text{tr}(\underbrace{\Phi^T\Phi}_{\text{cov}(f(Y_1))}\underbrace{(\Phi D)^T\Phi D)}_{\text{cov}(g(Y_2))} + \text{const.}$$

This leads to the following objective function

$$\mathscr{H}(f,g) = -\mathbb{E}[f(Y_1)^T g(Y_2)] + \frac{1}{2}\text{tr}(\text{cov}(f(Y_1))\text{cov}(g(Y_2))). \tag{3.23}$$

Although the objective function looks similar to the one we have for the first NN structure, there is a constraint for f and g to satisfy in this case. That is,

$$f_i^*(y_1) = [\Phi]_{y_1,i}/\sqrt{P_Y(y_1)}, i = 1,\ldots,k, y_1 \in \mathcal{Y}$$
$$g_i^*(y_2) = [\Phi D]_{y_2,i}/\sqrt{P_Y(y_2)}, i = 1,\ldots,k, y_2 \in \mathcal{Y}$$

which means that features of g can be obtained through applying D on the features of f. More specifically, after obtaining the features of f, a new module with sign switch (multiply with 1 or -1) can be applied to each dimension of f to obtain g. The corresponding neural network structure is shown in Figure 3.5, where D is used to select the correct sign (positive or negative) for each feature function. The detailed algorithm is given in Algorithm 7.

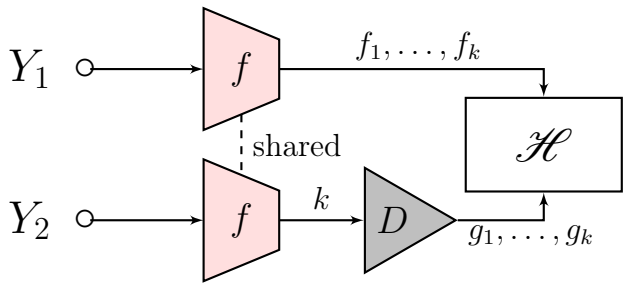


Figure 3.5: Partial symmetric NN structure to extract the top k feature pairs for HMM.

Compared with the second NN structure in Figure 3.4, this one learns k extra parameters in the network through D module; it will recover the row space of $\tilde{B}_{X;Y}$ when the number of selected features k is large enough and $\tilde{B}^{\text{sym}}_{X_1;X_2}$ has full rank (compared with the requirement that $\tilde{B}^{\text{sym}}_{X_1;X_2}$ only has positive eigenvalues in second NN structure).

Algorithm 7 Partial symmetric negative H-score on a mini-batch to extract the top k feature pairs for HMM

Require: Training samples $\{(y_1^{(i)}, y_2^{(i)}) : i = 1, ..., m\}$ in a mini-batch of size m

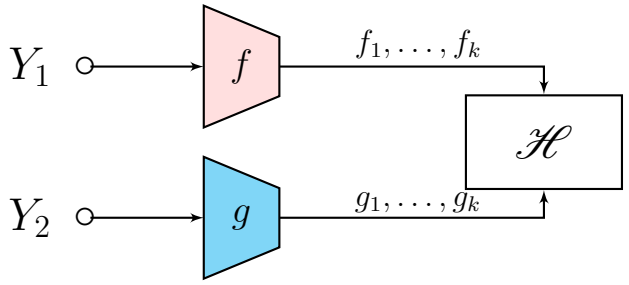
Require: Two branches of parameterized NNs with k output units: f and g, where g is f with a sign correction module

1. Center:
   $f(y_1^{(i)}) \longleftarrow f(y_1^{(i)}) - \frac{1}{m}\sum_{j=1}^{m} f(y_1^{(j)}), i = 1, \ldots, m$
   $g(y_2^{(i)}) \longleftarrow g(y_2^{(i)}) - \frac{1}{m}\sum_{j=1}^{m} g(y_2^{(j)}), i = 1, \ldots, m$
2. Compute the empirical covariance:
   $\mathrm{cov}(f) \longleftarrow \frac{1}{m-1}\sum_{i=1}^{m} f(y_1^{(i)}) f^T(y_1^{(i)}), i = 1, \ldots, m$
   $\mathrm{cov}(g) \longleftarrow \frac{1}{m-1}\sum_{i=1}^{m} g(y_2^{(i)}) g^T(y_2^{(i)}), i = 1, \ldots, m$
3. Compute the empirical objective $\mathscr{H}$:
   $-\frac{1}{m}\sum_{i=1}^{m} f^T(y_1^{(i)}) g(y_2^{(i)}) + \frac{1}{2}\mathrm{tr}(\mathrm{cov}(f)\mathrm{cov}(g))$

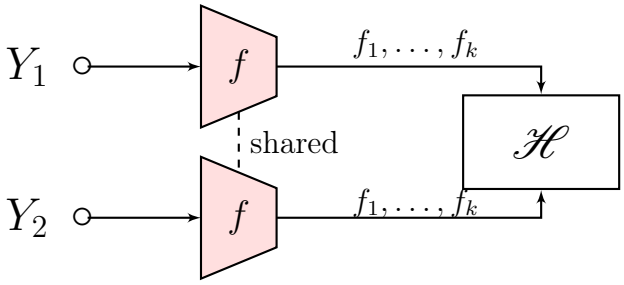
Chapter 4

Convergence Analysis

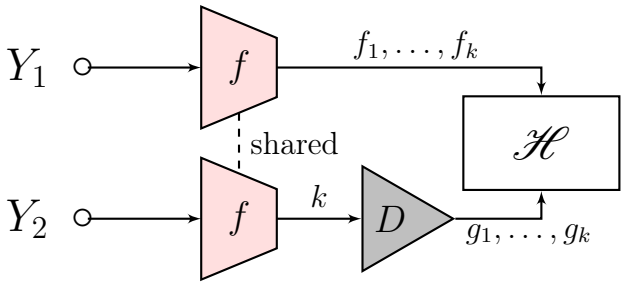
In Chapter 3, we introduce three different structures to extract features from HMM as concluded in Figure 4.1. Here is a summary:



(a) NN structure to extract the top k feature pairs.



(b) Symmetric NN structure to extract the top k feature pairs.



(c) Partial symmetric NN structure to extract the top k feature pairs.

Figure 4.1: Comparison of the three NN structures

The first model (a) tries to solve

$$\Phi^*_{(\mathsf{a})} = \arg \min_{\Phi \in \mathbb{R}^{N \times k}} \left(\min_{\Psi \in \mathbb{R}^{N \times k}} \|\check{B}_{Y_1;Y_2} - \Phi \Psi^T\|_F^2 \right),$$

and is able to extract **all** feature functions for k (number of selected features) large enough and $\tilde{B}_{X_1;X_2}$ has full rank.

The second model (b) tries to solve

$$\Phi^*_{(\mathsf{pd})} = \arg\min_{\Phi\in\mathbb{R}^{N\times k}} \|\check{B}_{Y_1;Y_2} - \Phi\Phi^T\|_F^2,$$

and is able to extract feature functions corresponding to the **positive definite** part of $\check{B}_{Y_1;Y_2}$ for k large enough and $\tilde{B}_{X_1;X_2}$ has full rank.

The third model (c) tries to solve

$$\Phi^*_{(\mathsf{s})} = \arg\min_{\Phi\in\mathbb{R}^{N\times k}} \Big(\min_{D\in diag(k)} \|\check{B}_{Y_1;Y_2} - \Phi D\Phi^T\|_F^2 \Big),$$

and is able to extract feature functions corresponding to the **symmetric** part of $\check{B}_{Y_1;Y_2}$ for k large enough and $\tilde{B}_{X_1;X_2}$ has full rank.

In this chapter, we will analyze the convergence properties for each model and demonstrate the trade-off between the number of learned parameters, and the accuracy of the features.

4.1 Closeness between CDMs

We first claim that the empirical $\check{B}_{Y_1;Y_2}$ calculated through empirical distribution is close to the exact $\tilde{B}_{Y_1;Y_2}$ calculated through population distribution. To be more precise, we have the following theorem:

Theorem 4.1.1. Assume there are n set of samples $\{(y_1^{(i)}, y_2^{(i)}) : i = 1, \dots n\}$, then

$$\mathbb{E}\big[\|\tilde{B}_{Y_1;Y_2} - \check{B}_{Y_1;Y_2}\|_F^2\big] = O\Big(\frac{1}{n}\Big), \tag{4.1}$$

$$\mathbb{E}\big[\|\tilde{B}^{\mathrm{sym}}_{Y_1;Y_2} - \check{B}^{\mathrm{sym}}_{Y_1;Y_2}\|_F^2\big] = O\Big(\frac{1}{n}\Big), \tag{4.2}$$

$$\mathbb{E}\big[\|\tilde{B}^{\mathrm{pd}}_{Y_1;Y_2} - \check{B}^{\mathrm{pd}}_{Y_1;Y_2}\|_F^2\big] = O\Big(\frac{1}{n}\Big), \tag{4.3}$$

where $\tilde{B}^{\mathrm{sym}}_{Y_1;Y_2} = \frac{\tilde{B}_{Y_1;Y_2} + \tilde{B}^T_{Y_1;Y_2}}{2}$, $\tilde{B}^{\mathrm{pd}}_{Y_1;Y_2}$ is the positive definite part of $\tilde{B}^{\mathrm{sym}}_{Y_1;Y_2}$, i.e., replace all the negative eigenvalues of $\tilde{B}^{\mathrm{sym}}_{Y_1;Y_2}$ with zeros, and $\check{B}^{\mathrm{pd}}_{Y_1;Y_2}$ is the empirical version of it.

Proof. • Proof of (4.1)

$$\check{P}_{Y_1,Y_2}(a,b) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}\{y_1^{(i)} = a, y_2^{(i)} = b\}, \forall a, b \in \mathcal{Y} \tag{4.4}$$

where $\mathbb{1}\{\cdot\}$ represents an indicator function, i.e., $\mathbb{1}\{y_1^{(i)} = a, y_2^{(i)} = b\}$ is 1 when $y_1^{(i)} = a$ and $y_2^{(i)} = b$ are satisfied, and is 0 otherwise. Let $X_i = \mathbb{1}\{y_1^{(i)} = a, y_2^{(i)} = b\}, i = 1, \dots n$. Then $X_i \overset{i.i.d}{\sim} \mathrm{Ber}(p)$ where $p = P_{Y_1,Y_2}(a,b)$.

Assume $\exists P_0 > 0$, such that $P_Y(y) \geq P_0, \forall y \in \mathcal{Y}$. This is a reasonable assumption since for a finite discrete random variable, outcomes with zero probability can be removed out of alphabet. We have

$$\begin{aligned}
& \mathbb{E}\big[\|\tilde{B}_{Y_1;Y_2} - \check{B}_{Y_1;Y_2}\|_F^2\big] \\
= \; & \mathbb{E}\left[\sum_{a\in\mathcal{Y}}\sum_{b\in\mathcal{Y}}\Big(\frac{\check{P}_{Y_1,Y_2}(a,b) - P_{Y_1,Y_2}(a,b)}{\sqrt{P_Y(a)}\sqrt{P_Y(b)}}\Big)^2\right] && (4.5)\\
= \; & \sum_{a\in\mathcal{Y}}\sum_{b\in\mathcal{Y}}\mathbb{E}\left[\frac{\big(\frac{1}{n}\sum_{i=1}^n \mathbb{1}\{y_1^{(i)} = a, y_2^{(i)} = b\} - P_{Y_1,Y_2}(a,b)\big)^2}{P_Y(a)P_Y(b)}\right] \\
\leq \; & \frac{\sum_{a\in\mathcal{Y}}\sum_{b\in\mathcal{Y}}\operatorname{var}\big(\frac{1}{n}\sum_{i=1}^n \mathbb{1}\{y_1^{(i)} = a, y_2^{(i)} = b\}\big)}{P_0^2} && (4.6)\\
= \; & \frac{\sum_{a\in\mathcal{Y}}\sum_{b\in\mathcal{Y}}\operatorname{var}(\mathbb{1}\{y_1^{(i)} = a, y_2^{(i)} = b\})}{nP_0^2} \\
= \; & \frac{\sum_{a\in\mathcal{Y}}\sum_{b\in\mathcal{Y}}P_{Y_1,Y_2}(a,b)(1 - P_{Y_1,Y_2}(a,b))}{nP_0^2} && (4.7)\\
\leq \; & \frac{1}{nP_0^2}
\end{aligned}$$

where to obtain (4.5), we use the assumption that marginal can be estimated accurately; to obtain (4.6), we use the assumption that $P_Y(y) \geq P_0, \forall y \in \mathcal{Y}$, and $\mathbb{E}\big[\frac{1}{n}\sum_{i=1}^n \mathbb{1}\{y_1^{(i)} = a, y_2^{(i)} = b\}\big] = p = P_{Y_1,Y_2}(a,b)$; to obtain (4.7), we use the fact that the variance of a Bernoulli variable with parameter p is $p(1-p)$.

On the other hand, assume $\exists P_1 < 1$, such that $P_{Y_1Y_2}(y_1, y_2) \leq P_1, \forall y_1, y_2 \in \mathcal{Y}$, and $P_Y(y) \leq 1, \forall y \in \mathcal{Y}$, we have $\mathbb{E}\big[\|\tilde{B}_{Y_1;Y_2} - \check{B}_{Y_1;Y_2}\|_F^2\big] \geq \frac{1-P_1}{n}$.

Therefore, $\mathbb{E}\big[\|\tilde{B}_{Y_1;Y_2} - \check{B}_{Y_1;Y_2}\|_F^2\big] = O\big(\frac{1}{n}\big)$.

- Proof of (4.2) and (4.3)

  The proof of (4.2) and (4.3) are similar to each other. So here, we will show the proof of (4.2) as an example. The proof of (4.3) can be obtained accordingly.

  Recall the empirical distribution $\check{P}^{\text{sym}}_{Y_1Y_2}$ has the following definition:

$$\check{P}^{\text{sym}}_{Y_1,Y_2}(a,b) = \frac{1}{2n}\sum_{i=1}^n \big(\mathbb{1}\{y_1^{(i)} = a, y_2^{(i)} = b\} + \mathbb{1}\{y_1^{(i)} = b, y_2^{(i)} = a\}\big), \forall a, b \in \mathcal{Y} \qquad (4.8)$$

  Let $X_i = \mathbb{1}\{y_1^{(i)} = a, y_2^{(i)} = b\}, i = 1, \ldots n$. Then $X_i \overset{i.i.d}{\sim} \text{Ber}(p)$ where $p = P_{Y_1,Y_2}(a,b)$. It also interesting to analyze the relationship between X_i and $Z_i = \mathbb{1}\{y_1^{(i)} = b, y_2^{(i)} =$

$a\}, i = 1, \dots n \overset{i.i.d}{\sim} \text{Ber}(P_{Y_1,Y_2}(b,a))$. We have for $a \neq b$:

$$\begin{aligned} \text{cov}(X_i, Z_i) &= \mathbb{E}[X_i Z_i] - \mathbb{E}[X_i]\mathbb{E}[Z_i] \\ &= 0 - P_{Y_1,Y_2}(a,b) \times P_{Y_1,Y_2}(b,a) \\ &= -P_{Y_1,Y_2}(a,b) P_{Y_1,Y_2}(b,a) \end{aligned} \tag{4.9}$$

where to obtain (4.9), we use the fact that $X_i Z_i$ is always 0 unless $a = b$. When $a = b$, the covariance should be $\text{Cov}(X_i, Z_i) = P_{Y_1,Y_2}(a,a) - P^2_{Y_1,Y_2}(a,a)$.

Assume $\exists P_0 > 0$, such that $P_Y(y) \geq P_0, \forall y \in \mathcal{Y}$. We have

$$\begin{aligned} & \mathbb{E}\big[\tilde{B}^{\text{sym}}_{Y_1;Y_2} - \check{B}^{\text{sym}}_{Y_1;Y_2}\|^2_F\big] \\ = \; & \mathbb{E}\left[\sum_{a\in\mathcal{Y}}\sum_{b\in\mathcal{Y}}\Big(\frac{\check{P}^{\text{sym}}_{Y_1,Y_2}(a,b) - P^{\text{sym}}_{Y_1,Y_2}(a,b)}{\sqrt{P_Y(a)}\sqrt{P_Y(b)}}\Big)^2\right] && (4.10) \\ = \; & \sum_{a\in\mathcal{Y}}\sum_{b\in\mathcal{Y}}\mathbb{E}\left[\frac{\Big(\frac{1}{2n}\sum_{i=1}^n\big(\mathbb{1}\{y_1^{(i)} = a, y_2^{(i)} = b\} + \mathbb{1}\{y_1^{(i)} = b, y_2^{(i)} = a\}\big) - P^{\text{sym}}_{Y_1,Y_2}(a,b)\Big)^2}{P_Y(a)P_Y(b)}\right] \\ \leq \; & \frac{\sum_{a\in\mathcal{Y}}\sum_{b\in\mathcal{Y}}\text{var}\Big(\frac{1}{2n}\sum_{i=1}^n\big(\mathbb{1}\{y_1^{(i)} = a, y_2^{(i)} = b\} + \mathbb{1}\{y_1^{(i)} = b, y_2^{(i)} = a\}\big)\Big)}{P_0^2} && (4.11) \\ = \; & \frac{\sum_{a\in\mathcal{Y}}\sum_{b\in\mathcal{Y}}\text{var}\Big(\mathbb{1}\{y_1^{(i)} = a, y_2^{(i)} = b\} + \mathbb{1}\{y_1^{(i)} = b, y_2^{(i)} = a\}\Big)}{4nP_0^2} && (4.12) \\ = \; & \frac{\sum_{a\in\mathcal{Y}}\sum_{b\in\mathcal{Y}}\Big(\text{var}\big(\mathbb{1}\{y_1^{(i)} = a, y_2^{(i)} = b\}\big) + \text{var}\big(\mathbb{1}\{y_1^{(i)} = b, y_2^{(i)} = a\}\big) + }{4nP_0^2} \\ & \frac{2\cdot\text{cov}\big(\mathbb{1}\{y_1^{(i)} = a, y_2^{(i)} = b\}, \mathbb{1}\{y_1^{(i)} = b, y_2^{(i)} = a\}\big)\Big)}{} && (4.13) \\ = \; & \frac{\sum_{a\in\mathcal{Y}}\sum_{b\in\mathcal{Y}}\Big(P_{Y_1,Y_2}(a,b)(1 - P_{Y_1,Y_2}(a,b)) + P_{Y_1,Y_2}(b,a)(1 - P_{Y_1,Y_2}(b,a))\Big) - }{4nP_0^2} \\ & \frac{\sum_{a\neq b} 2\cdot P_{Y_1,Y_2}(a,b)P_{Y_1,Y_2}(b,a) + \sum_{a=b} 2\cdot\big(P_{Y_1,Y_2}(a,a) - P^2_{Y_1,Y_2}(a,a)\big)}{} && (4.14) \\ \leq \; & \frac{1}{nP_0^2} \end{aligned}$$

where to obtain (4.10), we use the assumption that marginal can be estimated accurately; to obtain (4.11), we use the assumption that $P_Y(y) \geq P_0, \forall y \in \mathcal{Y}$, and $\mathbb{E}\big[\frac{1}{2n}\sum_{i=1}^n\big(1\{y_1^{(i)} = a, y_2^{(i)} = b\} + 1\{y_1^{(i)} = b, y_2^{(i)} = a\}\big)\big] = \frac{P_{Y_1Y_2}(a,b)+P_{Y_1Y_2}(b,a)}{2} = P^{\text{sym}}_{Y_1Y_2}(a,b)$; to obtain (4.12), we use the fact that $X_i + Z_i = 1\{y_1^{(i)} = a, y_2^{(i)} = b\} + 1\{y_1^{(i)} = b, y_2^{(i)} = a\}$ are i.i.d for all $i = 1, \dots, n$; to obtain (4.13), we use

the formula for the variance of the sum, i.e., for any random variable A and B, $\mathrm{var}(A+B) = \mathrm{var}(A) + \mathrm{var}(B) + 2 \cdot \mathrm{cov}(A, B)$; to obtain (4.14), we use the fact that the variance of a Bernoulli variable with parameter p is $p(1-p)$, and the covariance between X_i and Z_i.

Therefore, $\mathbb{E}\big[\|\tilde{B}^{\mathrm{sym}}_{Y_1;Y_2} - \check{B}^{\mathrm{sym}}_{Y_1;Y_2}\|_F^2\big] = O\big(\frac{1}{n}\big)$.

□

We then analyze the distance between the singular vector space of $\tilde{B}_{Y_1;Y_2}$ and $\check{B}_{Y_1;Y_2}$. We claim that the singular vector space is continuous in the sense that if $\|\tilde{B}_{Y_1;Y_2} - \check{B}_{Y_1;Y_2}\|_F \to 0$, then the singular vector space of $\tilde{B}_{Y_1;Y_2}$ and $\check{B}_{Y_1;Y_2}$ is close to each other.

Singular vector space is a subspace spanned by a set of orthonormal basis. Here, we introduce the $\sin\Theta$ distance [94] to measure the difference between two singular vector spaces, U and $\check{U}$, which are both of size $n \times k$. Suppose the singular values of $U^T\check{U}$ are $\sigma_1 \geq \cdots \geq \sigma_k \geq 0$, then we call

$$\Theta(U, \check{U}) = \mathrm{diag}(\cos^{-1}(\sigma_1), \cdots, \cos^{-1}(\sigma_k))$$

as principle angles. A quantitative measure of distance between U and $\check{U}$ is then $\|\sin\Theta\|_F$.

The following theorem about SVD perturbation is important for our analysis.

Theorem 4.1.2 (Wedin [95])**.** Let A and $\tilde{A} = A + E$ be two $m \times n$ $(m \geq n)$ matrices with SVDs:

$$A = U\Sigma V^T \equiv (U_1, U_2, U_3) \begin{pmatrix} \Sigma_1 & 0 \\ 0 & \Sigma_2 \\ 0 & 0 \end{pmatrix} \begin{pmatrix} V_1^T \\ V_2^T \end{pmatrix} \tag{4.15}$$

$$\tilde{A} = \tilde{U}\tilde{\Sigma}\tilde{V}^T \equiv (\tilde{U}_1, \tilde{U}_2, \tilde{U}_3) \begin{pmatrix} \tilde{\Sigma}_1 & 0 \\ 0 & \tilde{\Sigma}_2 \\ 0 & 0 \end{pmatrix} \begin{pmatrix} \tilde{V}_1^T \\ \tilde{V}_2^T \end{pmatrix} \tag{4.16}$$

where U and $\tilde{U}$ are $m \times m$ orthogonal matrices, V and $\tilde{V}$ are $n \times n$ orthogonal matrices, $U_1, \tilde{U}_1 \in \mathbb{R}^{m\times k}(1 \leq k < n)$ are first k columns of U and $\tilde{U}$, respectively, $V_1, \tilde{V}_1 \in \mathbb{R}^{n\times k}$ are first k columns of V and $\tilde{V}$, respectively, and

$$\Sigma_1 = \mathrm{diag}(\sigma_1, \ldots, \sigma_k), \quad \Sigma_2 = \mathrm{diag}(\sigma_{k+1}, \ldots, \sigma_n)$$
$$\tilde{\Sigma}_1 = \mathrm{diag}(\tilde{\sigma}_1, \ldots, \tilde{\sigma}_k), \quad \tilde{\Sigma}_2 = \mathrm{diag}(\tilde{\sigma}_{k+1}, \ldots, \tilde{\sigma}_n)$$

If $\exists \alpha, \delta > 0$, s.t

$$\min_{1\leq i\leq k} \sigma_i \geq \alpha + \delta, \quad \max_{k+1\leq j\leq n} \tilde{\sigma}_j \leq \alpha,$$

then

$$\max\big\{||\sin\Theta(U_1, \tilde{U}_1)||_{\mathrm{F}}, ||\sin\Theta(V_1, \tilde{V}_1)||_{\mathrm{F}}\big\} \leq \frac{||E||_{\mathrm{F}}}{\delta} \tag{4.17}$$

Proof. See [96]. □

Notice although the second condition is placed on $\tilde{\sigma}$, if E is small compared to δ, we can place it on σ since the Weyl's theorem [97], [98] guarantees that $\tilde{\sigma}$'s are very close to σ's.

Therefore, Theorem 4.1.2 tells us that as long as the k-th singular value of the exact B is bounded below by 0, the singular vector space is continuous in the sense that if $\|\tilde{B}_{Y_1;Y_2} - \check{B}_{Y_1;Y_2}\|_F \to 0$, then $\|\sin\Theta(U, \check{U})\|_F \to 0$.

To summarize, theorems 4.1.1 and 4.1.2 are important because it indicates the convergence of the proposed algorithms to their targeted features. Take the first NN structure for an example, because (4.1) holds, and when $\|\tilde{B}_{Y_1;Y_2} - \check{B}_{Y_1;Y_2}\|_F^2 \to 0$, the distance between their singular vector spaces converges to 0, this model's output as the low-rank approximation of $\check{B}_{Y_1;Y_2}$ will recover the subspace that converges to the true singular vector space of $\tilde{B}_{Y_1;Y_2}$, which is the same as the row space of $\tilde{B}_{X;Y}$ (the key component that we hope to find) as long as $\tilde{B}_{X_1;X_2}$ has full rank. For the second and third structure, the targets are no longer $\tilde{B}_{Y_1;Y_2}$, but $\tilde{B}^{\text{sym}}_{Y_1;Y_2}$ and $\tilde{B}^{\text{pd}}_{Y_1;Y_2}$, respectively. The difference in targets reveals the trade-off between the number of learned parameters and the accuracy of the features.

4.2 Sample Complexity

4.2.1 Analysis in $Y_1 - Y_2$ Space

In this subsection, we determine the number of samples required to obtain accurate estimates of $\mathbf{f}_{(k)} = [f_1, f_2, \ldots, f_k]^T$ of the exact features $\mathbf{f}^*_{(k)} = [f_1^*, f_2^*, \ldots, f_k^*]^T$. Despite the existence of invariant subspace stability results, the individual singular vectors of a matrix often vary greatly under perturbations. So, instead of directly analyzing the convergence of a single feature function to another, we analyze the whole space defined by the feature functions. To be more specific, our development focuses on measuring the accuracy of these estimates with the following metrics:

$$\Delta_{(\mathsf{a})} = \|\tilde{B}_{Y_1;Y_2}\Phi_{(\mathsf{a})}\|_F^2 - \|\tilde{B}_{Y_1;Y_2}\hat{\Phi}_{(\mathsf{a})}\|_F^2, \tag{4.18}$$

$$\Delta_{(\mathsf{s})} = \|\tilde{B}^{\text{sym}}_{Y_1;Y_2}\Phi_{(\mathsf{s})}\|_F^2 - \|\tilde{B}^{\text{sym}}_{Y_1;Y_2}\hat{\Phi}_{(\mathsf{s})}\|_F^2, \tag{4.19}$$

$$\Delta_{(\mathsf{pd})} = \|\tilde{B}^{\text{pd}}_{Y_1;Y_2}\Phi_{(\mathsf{pd})}\|_F^2 - \|\tilde{B}^{\text{pd}}_{Y_1;Y_2}\hat{\Phi}_{(\mathsf{pd})}\|_F^2, \tag{4.20}$$

where $\Phi_{(\mathsf{a})}, \Phi_{(\mathsf{s})}, \Phi_{(\mathsf{pd})} \in \mathbb{R}^{|\mathcal{Y}|\times k}$ are the exact orthonormal subspaces of $\tilde{B}_{Y_1;Y_2}, \tilde{B}^{\text{sym}}_{Y_1;Y_2}$, and $\tilde{B}^{\text{pd}}_{Y_1;Y_2}$, receptively. $\hat{\Phi}_{(\cdot)} \in \mathbb{R}^{|\mathcal{Y}|\times k}$ is the corresponding orthonormal subspace obtained by our algorithms. We explain the reason to use such a metric. In a local sense,

$$I(Y_1, Y_2) = \frac{1}{2}\|\tilde{B}_{Y_1;Y_2}\|_F^2 + o(\epsilon^2) = \frac{1}{2}\sum \sigma_i^2 + o(\epsilon^2).$$

Therefore, in (4.18), $\|\tilde{B}_{Y_1;Y_2}\Phi_{(\mathsf{a})}\|_F^2 = \sum \sigma_i^2$ captures the most information from Y inside the data. This metric, denoted as $\Delta_{(\mathsf{a})}$, measures how effectively the algorithm captures information from Y. For (4.19) and (4.20), as explained before, some information will be

lost at the cost of a fewer number of parameters to learn. Therefore, different targets are used, representing the symmetric and positive definite subspaces these algorithms are trying to represent, respectively. In general, a smaller value of $\Delta_{(\cdot)}$ indicates better algorithm performance. In the end, the goal would be getting some results for the tail bound $\mathbb{P}(\Delta_{(\cdot)} \geq \delta)$ and mean square error (MSE) bound $\mathbb{E}[\Delta_{(\cdot)}^2]$.

First note that $\Delta_{(\cdot)} \geq 0$, where the proof through a related lemma can be found in the Appendix B. This loss function captures the extent to which the estimates preserve as much of a "rank-k approximation" of the mutual information between Y_1 and Y_2 as possible under local approximations.

The next theorem portrays an exponential concentration of measure inequality for the loss $\Delta_{(\cdot)}$. For simplicity, we denote $\Delta_{(\cdot)}$ as Δ, $\hat{\Phi}_{(\cdot)}$ as $\hat{\Phi}$, and $K_{(\cdot)}$ as K, where $K_{(\mathsf{a})} = \operatorname{rank}(\tilde{B}_{Y_1;Y_2})$, $K_{(\mathsf{s})} = \operatorname{rank}(\tilde{B}^{\text{sym}}_{Y_1;Y_2})$, and $K_{(\mathsf{pd})} = \operatorname{rank}(\tilde{B}^{\text{pd}}_{Y_1;Y_2})$. n represents the number of training samples. For $k \in \{1, \ldots, K\}$:

Theorem 4.2.1. For $\hat{\Phi}$ obtained by our algorithm and $0 \leq \delta \leq 4k$, we have

$$\mathbb{P}(\Delta \geq \delta) \leq 2|\mathcal{Y}| \exp\left(- \frac{P_0 \delta^2 n}{64 k^2} \right)$$

where $P_{Y_1}(y) = P_{Y_2}(y) \geq P_0, \forall y \in \mathcal{Y}$ for some $P_0 > 0$.

Proof. • Proof for (4.18)

Lemma 4.2.2. Given $A_1, A_2 \in \mathbb{R}^{k_1 \times k_2}$, and $k \in \{1, \ldots, \min(k_1, k_2)\}$, we have

$$0 \leq \|A_1 \mathbf{\Psi}^{A_1}_{(k)}\|_F^2 - \|A_1 \mathbf{\Psi}^{A_2}_{(k)}\|_F^2 \leq 4k \|A_1\|_s \|A_1 - A_2\|_s,$$

where $\psi^A_{(i)}$ denotes the right singular vector of A corresponding to $\sigma_i(A)$, and

$$\mathbf{\Psi}^A_{(k)} = [\psi^A_{(1)} \cdots \psi^A_{(k)}],$$

which has orthonormal columns.

Proof for Lemma 4.2.2: See [79].

$$\begin{aligned} \Delta &= \|\tilde{B}_{Y_1;Y_2} \Phi_{(\mathsf{a})}\|_F^2 - \|\tilde{B}_{Y_1;Y_2} \hat{\Phi}_{(\mathsf{a})}\|_F^2 \\ &\leq 4k \|\tilde{B}_{Y_1;Y_2}\|_s \|\tilde{B}_{Y_1;Y_2} - \check{B}_{Y_1;Y_2}\|_s \qquad (4.21) \\ &\leq 4k \|\tilde{B}_{Y_1;Y_2} - \check{B}_{Y_1;Y_2}\|_s, \qquad (4.22) \end{aligned}$$

where to obtain (4.21), we use Lemma 4.2.2; to obtain (4.22), we notice that the largest singular value for $\tilde{B}_{Y_1;Y_2}$ is less than 1, i.e., $\|\tilde{B}_{Y_1;Y_2}\|_s \leq 1$.

Now let $\check{\mathbf{B}}_i$ denote an $|\mathcal{Y}| \times |\mathcal{Y}|$ matrix with (y_1, y_2)-th entry $\check{B}_i(y_1, y_2) = \frac{\mathbb{1}\{Y_1^{(i)} = y_1, Y_2^{(i)} = y_2\}}{\sqrt{P_Y(y_1) P_Y(y_2)}} - \sqrt{P_Y(y_1) P_Y(y_2)}$, and let $\tilde{Z}_i = \check{\mathbf{B}}_i - \tilde{B}_{Y_1;Y_2}$. Then, it is obvious that $\mathbb{E}[\tilde{Z}_i] = 0$, and we

can show that for $P_{Y_1}(y) = P_{Y_2}(y) \geq P_0, \forall y \in \mathcal{Y}$ for some $P_0 > 0$,

$$\begin{aligned}
\|\tilde{Z}_i\|_s &= \|\check{\mathbf{B}}_i - \tilde{B}_{Y_1;Y_2}\|_s \\
&= \left\|\check{\mathbf{B}}_i + \sqrt{P_Y(Y_1^{(i)})}\sqrt{P_Y(Y_2^{(i)})} - \tilde{B}_{Y_1;Y_2} - \sqrt{P_Y(Y_1^{(i)})}\sqrt{P_Y(Y_2^{(i)})}\right\|_s \\
&\leq \left\|\tilde{B}_{Y_1;Y_2} + \sqrt{P_Y(Y_1^{(i)})}\sqrt{P_Y(Y_2^{(i)})}\right\|_s + \left\|\check{\mathbf{B}}_i + \sqrt{P_Y(Y_1^{(i)})}\sqrt{P_Y(Y_2^{(i)})}\right\|_s \quad (4.23) \\
&= 1 + \frac{1}{\sqrt{P_Y(Y_1^{(i)})}\sqrt{P_Y(Y_2^{(i)})}} \quad (4.24) \\
&\leq 1 + \frac{1}{P_0} \triangleq c_1, \quad (4.25)
\end{aligned}$$

where to obtain (4.23) we have used the spectral norm triangle inequality, to obtain (4.24) we have used that $\|\tilde{B}_{Y_1;Y_2} + \sqrt{P_Y(Y_1^{(i)})}\sqrt{P_Y(Y_2^{(i)})}\|_s = 1$ and $\check{\mathbf{B}}_i + \sqrt{P_Y(Y_1^{(i)})}\sqrt{P_Y(Y_2^{(i)})}$ has a single nonzero entry so $e_{Y_1^{(i)}}$ and $e_{Y_2^{(i)}}$ are its principal left and right singular vectors, respectively, and to obtain (4.25) we have used the definition of P_0.

Similarly, we can derive that

$$\left\|\frac{1}{n}\sum_{i=1}^{n}\mathrm{cov}(\tilde{Z}_i)\right\|_s \leq 1 + \frac{1}{P_0} \triangleq \bar{c}_1$$

We introduce the following matrix generation of Bernstein's inequality Theorem [99].

Lemma 4.2.3 (Bernstein's Inequality (Matrix Version))**.** : for some dimensions d_1 and d_2, let $\tilde{Z}_1, \ldots, \tilde{Z}_n \in \mathbb{R}^{d_1 \times d_2}$ be independent zero-mean random matrices such that for some constant $c > 0$,

$$\mathbb{P}(\|\tilde{Z}_i\|_s \leq c) = 1, \quad i = 1, \ldots, n.$$

Moreover, let $\bar{c} \in (0, c^2]$ be a constant such that

$$\max\{\left\|\frac{1}{n}\sum_{i=1}^{n}\mathrm{cov}(\tilde{Z}_i)\right\|_s, \left\|\frac{1}{n}\sum_{i=1}^{n}\mathrm{cov}(\tilde{Z}_i^T)\right\|_s\} \leq \bar{c}.$$

Then, for all $0 \leq \delta \leq \bar{c}/c$,

$$\mathbb{P}(\left\|\frac{1}{n}\sum_{i=1}^{n}\tilde{Z}_i\right\|_s \geq \delta) \leq (d_1 + d_2)\exp\left(-\frac{3\delta^2 n}{8\bar{c}}\right)$$

Finally, we have

$$\begin{aligned}
\mathbb{P}(\Delta \geq \delta) &\leq \mathbb{P}(\|\tilde{B}_{Y_1;Y_2} - \check{B}_{Y_1;Y_2}\|_s \geq \frac{\delta}{4k}) && (4.26)\\
&\leq \mathbb{P}(\left\|\frac{1}{n}\sum_{i=1}^{n}\tilde{Z}_i\right\|_s \geq \frac{\delta}{4k}) && (4.27)\\
&\leq 2|\mathcal{Y}|\exp\left(-\frac{3n}{8}(\frac{1}{1+1/P_0})(\frac{\delta}{4k})^2\right) && (4.28)\\
&\leq 2|\mathcal{Y}|\exp\left(-\frac{P_0\delta^2 n}{64k^2}\right), && (4.29)
\end{aligned}$$

where to obtain (4.26), we use (4.22); to obtain (4.27), we use the fact that $\check{B}_{Y_1;Y_2} - \tilde{B}_{Y_1;Y_2} = \frac{1}{n}\sum_{i=1}^{n}\tilde{Z}_i$; to obtain (4.28), we use Lemma 4.2.3; to obtain (4.29), we use that $P_0 \leq 1/2$ since $|\mathcal{Y}| \geq 2$.

- Proof for (4.19) and (4.20)

  Proof for (4.19) and (4.20) are similar to each other. So here, we will show the proof for (4.19) as an example. The proof of (4.20) can be obtained accordingly.

$$\begin{aligned}
\Delta &= \|\tilde{B}^{\text{sym}}_{Y_1;Y_2}\Phi_{(\mathsf{s})}\|_F^2 - \|\tilde{B}^{\text{sym}}_{Y_1;Y_2}\hat{\Phi}_{(\mathsf{s})}\|_F^2\\
&\leq 4k\|\tilde{B}^{\text{sym}}_{Y_1;Y_2}\|_s\|\tilde{B}^{\text{sym}}_{Y_1;Y_2} - \check{B}^{\text{sym}}_{Y_1;Y_2}\|_s && (4.30)\\
&\leq 4k\|\tilde{B}^{\text{sym}}_{Y_1;Y_2} - \check{B}^{\text{sym}}_{Y_1;Y_2}\|_s, && (4.31)
\end{aligned}$$

where to obtain (4.30), we use Lemma 4.2.2; to obtain (4.31), we notice that the largest singular value for $\tilde{B}^{\text{sym}}_{Y_1;Y_2}$ is less than 1, i.e., $\|\tilde{B}^{\text{sym}}_{Y_1;Y_2}\|_s \leq 1$.

Now let $\check{\mathbf{B}}_i^{\text{sym}}$ denotes an $|\mathcal{Y}| \times |\mathcal{Y}|$ matrix with (y_1, y_2)-th entry

$$\check{B}_i(y_1,y_2) = \frac{\mathbb{1}\{Y_1^{(i)} = y_1, Y_2^{(i)} = y_2\} + \mathbb{1}\{Y_1^{(i)} = y_2, Y_2^{(i)} = y_1\}}{2\sqrt{P_Y(y_1)P_Y(y_2)}} - \sqrt{P_Y(y_1)P_Y(y_2)},$$

and $\tilde{Z}_i = \check{\mathbf{B}}_i^{\text{sym}} - \tilde{B}^{\text{sym}}_{Y_1;Y_2}$. Then, it is obvious that $\mathbb{E}[\tilde{Z}_i] = 0$, and similar to before, we can show that for $P_{Y_1}(y) = P_{Y_2}(y) \geq P_0, \forall y \in \mathcal{Y}$ for some $P_0 > 0$,

$$\begin{aligned}
\|\tilde{Z}_i\|_s &\leq 1 + \frac{1}{P_0} \triangleq c_1,\\
\left\|\frac{1}{n}\sum_{i=1}^{n}\text{cov}(\tilde{Z}_i)\right\|_s &\leq 1 + \frac{1}{P_0} \triangleq \bar{c}_1.
\end{aligned}$$

Finally, we have

$$\begin{aligned}
\mathbb{P}(\Delta \geq \delta) &\leq \mathbb{P}(\|\tilde{B}^{\text{sym}}_{Y_1;Y_2} - \check{B}^{\text{sym}}_{Y_1;Y_2}\|_s \geq \frac{\delta}{4k}) && (4.32)\\
&\leq \mathbb{P}(\left\|\frac{1}{n}\sum_{i=1}^{n}\tilde{Z}_i\right\|_s \geq \frac{\delta}{4k}) && (4.33)\\
&\leq 2|\mathcal{Y}|\exp\left(-\frac{3n}{8}(\frac{1}{1+1/P_0})(\frac{\delta}{4k})^2\right) && (4.34)\\
&\leq 2|\mathcal{Y}|\exp\left(-\frac{P_0\delta^2 n}{64k^2}\right), && (4.35)
\end{aligned}$$

where to obtain (4.32), we use (4.31); to obtain (4.33), we use the fact that $\check{B}^{\text{sym}}_{Y_1;Y_2} - \tilde{B}^{\text{sym}}_{Y_1;Y_2} = \frac{1}{n}\sum_{i=1}^{n}\tilde{Z}_i$; to obtain (4.34), we use Lemma 4.2.3; to obtain (4.35), we use that $P_0 \leq 1/2$ since $|\mathcal{Y}| \geq 2$.

□

This theorem illustrates that estimating $\Phi_{(\cdot)}$ (or $\mathbf{f}^*_{(k)}$) via $\hat{\Phi}_{(\cdot)}$ (or $\mathbf{f}_{(k)}$) to within a fixed error and confidence level requires n to grow quadratically with k. A key consequence of Theorem 4.2.1 is the following corollary, which presents a bound on the MSE between $\|\tilde{B}^{-}_{Y_1;Y_2}\Phi_{(\cdot)}\|_F^2$ and $\|\tilde{B}^{-}_{Y_1;Y_2}\hat{\Phi}_{(\cdot)}\|_F^2$.

Corollary 4.2.4. For the same P_0 defined above, and sufficiently large n such that $\frac{P_0 n}{32} \geq \frac{1}{|\mathcal{Y}|}$ and $\frac{P_0 n}{4} \geq \ln(\frac{P_0 n}{32}|\mathcal{Y}|)$, we have

$$\mathbb{E}[\Delta^2] \leq \frac{64k^2\left[\ln\left(P_0 n|\mathcal{Y}|\right) - 2\right]}{P_0 n}$$

For simplicity, we will use the result for (4.18) as an example to prove the corollary.

Proof.

$$\begin{aligned}
\Delta &= \|\tilde{B}_{Y_1;Y_2}\Phi_{(\mathsf{a})}\|_F^2 - \|\tilde{B}_{Y_1;Y_2}\hat{\Phi}_{(\mathsf{a})}\|_F^2\\
&\leq \|\tilde{B}_{Y_1;Y_2}\Phi_{(\mathsf{a})}\|_F^2\\
&\leq \|\tilde{B}_{Y_1;Y_2}\|_s^2\|\Phi_{(\mathsf{a})}\|_F^2 && (4.36)\\
&\leq \sum_{i=1}^{k}\|\phi_i\|^2 && (4.37)\\
&= k, && (4.38)
\end{aligned}$$

where to obtain (4.36), we use the fact that for any compatible matrices A_1 and A_2, we have $\|A_1A_2\|_F \leq \|A_1\|_s\|A_2\|_F$; to obtain (4.37), we use the fact that the largest singular value for $\tilde{B}_{Y_1;Y_2}$ is less than 1, i.e., $\|\tilde{B}_{Y_1;Y_2}\|_s \leq 1$; to obtain (4.38), we use the fact that singular vectors have unit norm.

Let ϵ_δ to be the event that $\{\Delta \geq \delta\}$, where $0 \leq \delta \leq 4k$, then

$$\begin{aligned}\mathbb{E}[\Delta^2] &= \mathbb{E}[\Delta^2|\epsilon_\delta^c]\mathbb{P}(\epsilon_\delta^c) + \mathbb{E}[\Delta^2|\epsilon_\delta]\mathbb{P}(\epsilon_\delta) \\ &\leq \delta^2 + 2k^2|\mathcal{Y}| \exp\bigg(-\frac{P_0\delta^2 n}{64k^2}\bigg), \end{aligned} \tag{4.39}$$

where to obtain (4.39), we use that $P(\epsilon_\delta^c) \leq 1$, (4.38), and Theorem 4.2.1.

Next, we optimize (4.39) over δ to obtain the tightest bound.

$$\begin{aligned}\mathbb{E}[\Delta^2] &\leq \min_\delta \quad \bigg(\delta^2 + 2k^2|\mathcal{Y}| \exp\bigg(-\frac{P_0\delta^2 n}{64k^2}\bigg)\bigg) \\ &= \frac{64k^2}{p_0 n}\bigg[1 + \ln\bigg(k^2|\mathcal{Y}|\frac{P_0 n}{32k^2}\bigg)\bigg] \qquad (4.40) \\ &= \frac{64k^2}{p_0 n}\Big[\ln(|\mathcal{Y}|P_0 n) + 1 - \ln(32)\Big] \\ &\leq \frac{64k^2}{p_0 n}\Big[\ln(|\mathcal{Y}|P_0 n) - 2\Big], \qquad (4.41)\end{aligned}$$

where to obtain (4.40), we take the derivative of (4.39) with respective to δ^2 and force it to be zero to get the optimal $\delta^{*2} = \frac{64k^2}{P_0 n}\ln\frac{|\mathcal{Y}|P_0 n}{32}$ since the function is convex with respect to δ^2; to obtain (4.41), we use $\ln(32) - 1 \geq 2$. Note that since δ in (4.39) has the constraint that $0 \leq \delta \leq 4k$ by Theorem 4.2.1, this constraint should be imposed on the minimizer δ^* as well. Therefore,

$$0 \leq \frac{64k^2}{P_0 n}\ln\frac{|\mathcal{Y}|P_0 n}{32} \leq 16k^2,$$

and equivalently, we have

$$\frac{P_0 n}{32} \geq \frac{1}{|\mathcal{Y}|},$$

and

$$\frac{P_0 n}{4} \geq \ln\bigg(\frac{P_0 n}{32}\bigg)|\mathcal{Y}|$$

□

Alternatively, if we choose a different proof methods, we have the following theorem and its corresponding corollary:

Theorem 4.2.5. For $\hat{\Phi} \in \mathbb{R}^{|\mathcal{Y}|\times k}$ obtained by our algorithm and $0 \leq \delta \leq \frac{4}{P_0}\sqrt{\frac{k}{2}}$, we have

$$\mathbb{P}(\Delta \geq \delta) \leq \exp\bigg(\frac{1}{4} - \frac{P_0^2\delta^2 n}{128k}\bigg)$$

where $P_{Y_1}(y) = P_{Y_2}(y) \geq P_0, \forall y \in \mathcal{Y}$ for some $P_0 > 0$.

Proof. We adapt the proof Theorem 4.2.1, replacing the use of the spectral norm bound of Lemma 4.2.2 with the following Frobenius norm bound:

Lemma 4.2.6. Given $A_1, A_2 \in \mathbb{R}^{k_1 \times k_2}$, and $k \in \{1, \ldots, \min(k_1, k_2)\}$, we have

$$0 \leq \|A_1 \mathbf{\Psi}_{(k)}^{A_1}\|_F^2 - \|A_1 \mathbf{\Psi}_{(k)}^{A_2}\|_F^2 \leq 4\sqrt{k}\|A_1\|_s \|A_1 - A_2\|_F,$$

where $\mathbf{\Psi}_{(k)}^{A}$ is defined as in Lemma 4.2.2.

Proof for Lemma 4.2.6: See [79].

From now on, takes (4.18) as an example, and the proof for (4.19) and (4.20) follow accordingly.

$$\begin{aligned} \Delta &= \|\tilde{B}_{Y_1;Y_2}\Phi_{(\mathsf{a})}\|_F^2 - \|\tilde{B}_{Y_1;Y_2}\hat{\Phi}_{(\mathsf{a})}\|_F^2 \\ &\leq 4\sqrt{k}\|\tilde{B}_{Y_1;Y_2}\|_s \|\tilde{B}_{Y_1;Y_2} - \check{B}_{Y_1;Y_2}\|_F \qquad (4.42) \\ &\leq 4\sqrt{k}\|\tilde{B}_{Y_1;Y_2} - \check{B}_{Y_1;Y_2}\|_F, \qquad (4.43) \end{aligned}$$

where to obtain (4.42), we use Lemma 4.2.6; to obtain (4.43), we use that $\|\tilde{B}_{Y_1;Y_2}\|_s \leq 1$.

If we use the same definition as before, the Frobenius norm of $\tilde{Z}_i$ satisfies the following inequalities:

$$\|\tilde{Z}_i\|_F^2 \leq \frac{2}{p_0^2} \triangleq c_2^2,$$
$$\frac{1}{n}\sum_{i=1}^{n} \mathbb{E}[\|\tilde{Z}_i\|_F^2] \leq \frac{1}{p_0^2} \triangleq \bar{c}_2.$$

We introduce the following vector generation of Bernstein's inequality Theorem [100].

Lemma 4.2.7 (Bernstein's Inequality (Vector Version))**.** : for some dimensions d, let $\tilde{Z}_1, \ldots, \tilde{Z}_n \in \mathbb{R}^d$ be independent zero-mean random matrices such that for some constant $c > 0$,

$$\mathbb{P}(\|\tilde{Z}_i\|_F \leq c) = 1, \quad i = 1, \ldots, n.$$

Moreover, let $\bar{c} \in (0, c^2]$ be a constant such that

$$\frac{1}{n}\sum_{i=1}^{n} \mathbb{E}[\|\tilde{Z}_i\|_F^2] \leq \bar{c}.$$

Then, for all $0 \leq \delta \leq \bar{c}/c$,[1]

$$\mathbb{P}\left(\left\|\frac{1}{n}\sum_{i=1}^{n}\tilde{Z}_i\right\| \geq \delta\right) \leq \exp\left(\frac{1}{4} - \frac{\delta^2 n}{8\bar{c}}\right)$$

[1] As notes in [100], this bound does not depend on d.

Finally, we have

$$\begin{aligned}
\mathbb{P}(\Delta \geq \delta) &\leq \mathbb{P}(\|\tilde{B}_{Y_1;Y_2} - \check{B}_{Y_1;Y_2}\|_F \geq \frac{\delta}{4\sqrt{k}}) && (4.44)\\
&\leq \mathbb{P}(\left\|\frac{1}{n}\sum_{i=1}^{n}\tilde{Z}_i\right\|_F \geq \frac{\delta}{4\sqrt{k}}) && (4.45)\\
&\leq \exp\left(\frac{1}{4} - \frac{P_0^2\delta^2 n}{128k}\right) && (4.46)
\end{aligned}$$

where to obtain (4.44), we use (4.43); to obtain (4.45), we use the fact that $\check{B}_{Y_1;Y_2} - \tilde{B}_{Y_1;Y_2} = \frac{1}{n}\sum_{i=1}^{n}\tilde{Z}_i$; to obtain (4.46), we use Lemma 4.2.7. □

Corollary 4.2.8. For the same P_0 defined above, and sufficiently large n such that $n \geq \frac{128e^{-\frac{1}{4}}}{P_0^2 k}$ and $n \geq 4[1 + 4\ln\frac{kP_0^2 n}{128}]$, we have

$$\mathbb{E}[\Delta^2] \leq \frac{32k}{P_0^2 n}\left[5 + 4\ln\left(\frac{kP_0^2 n}{128}\right)\right]$$

Proof. We adapt the proof Corollary 4.2.4, replacing the use of the sample complexity bound of Theorem 4.2.1 with Theorem 4.2.5:

$$\begin{aligned}
\mathbb{E}[\Delta^2] &= \mathbb{E}[\Delta^2|\epsilon_\delta^c]\mathbb{P}(\epsilon_\delta^c) + \mathbb{E}[\Delta^2|\epsilon_\delta]\mathbb{P}(\epsilon_\delta)\\
&\leq \delta^2 + k^2\exp\left(\frac{1}{4} - \frac{P_0^2\delta^2 n}{128k}\right), && (4.47)
\end{aligned}$$

where to obtain (4.47), we use that $P(\epsilon_\delta^c) \leq 1$, (4.38), and Theorem 4.2.5.

Next, we optimize (4.47) over δ to obtain the tightest bound.

$$\begin{aligned}
\mathbb{E}[\Delta^2] &\leq \min_\delta \left(\delta^2 + k^2\exp\left(\frac{1}{4} - \frac{P_0^2\delta^2 n}{128k}\right)\right)\\
&= \frac{32k}{P_0^2 n}\left[5 + 4\ln\left(\frac{kP_0^2 n}{128}\right)\right] && (4.48)
\end{aligned}$$

where to obtain (4.48), we take the derivative of (4.47) with respective to δ^2 and force it to be zero to get the optimal $\delta^{*2} = \frac{32k}{P_0^2 n}\left[1 + 4\ln\left(\frac{kP_0^2 n}{128}\right)\right]$ since the function is convex with respect to δ^2. Note that since δ in (4.47) has the constraint that $0 \leq \delta \leq \frac{4}{P_0}\sqrt{\frac{k}{2}}$ by Theorem 4.2.5, this constraint should be imposed on the minimizer δ^* as well. Therefore,

$$0 \leq \frac{32k}{P_0^2 n}\left[1 + 4\ln\left(\frac{kP_0^2 n}{128}\right)\right] \leq \frac{8k}{P_0^2},$$

and equivalently, we have

$$n \geq \frac{128e^{-\frac{1}{4}}}{P_0^2 k},$$

and

$$n \geq 4\left[1 + 4\ln\left[1 + 4\ln\left(\frac{kP_0^2 n}{128}\right)\right]\right]$$

□

4.2.2 Analysis in $X - Y$ Space

Since the row space of $\tilde{B}_{X;Y}$ is the target, we can also provide some analysis about the sample complexity in $X - Y$ space in addition to the results given above. More specifically, here we will focus on the proposed first and third structure in Figure 4.1. The corresponding result for the second structure in Figure 4.1 is similar to that for the third structurem and can be obtained accordingly.

To evaluate our algorithm's performance, we use the following metric:

$$\Delta = \|\tilde{B}_{X;Y}\Phi\|_F^2 - \|\tilde{B}_{X;Y}\hat{\Phi}\|_F^2, \tag{4.49}$$

where Φ is the exact orthonormal subspace of $\tilde{B}_{XY}$, i.e., target, and $\hat{\Phi}$ is the orthonormal subspace obtained by one of our algorithms. This metric quantifies how effectively the algorithm captures information from Y about X. A smaller value of Δ indicates better algorithm performance. In the end, the goal would be getting an upper bound for $\mathbb{P}(\Delta \geq \delta)$.

We analyze the metric Δ from two perspectives:

- The differences caused by $\tilde{B}_{X_1;X_2}$, i.e., between $\tilde{B}_{X;Y}$ and $\tilde{B}_{Y_1;Y_2}$;
- The differences due to random sampling, i.e., between $\tilde{B}_{Y_1;Y_2}$ and $\check{B}_{Y_1;Y_2}$.

If we lose some modes due to the rank deficiency of $\tilde{B}_{X_1;X_2}$, $\tilde{B}^{\text{sym}}_{X_1;X_2}$, or $\tilde{B}^{\text{pd}}_{X_1;X_2}$, our algorithm cannot recover those modes, resulting in a minimum Δ value around $\sum_{j\in\mathcal{L}} a_j\sigma_j^2$, where $\mathcal{I}$ denotes the set of lost mode indices, a_j represents the degree of the loss in that index direction, and σ_j corresponds to the singular values of $\tilde{B}_{X;Y}$ for these indices. $|\mathcal{I}| = |\mathcal{X}| -$ rank $(\tilde{B}_{X_1;X_2})$ ($\tilde{B}^{\text{sym}}_{X_1;X_2}$, or $\tilde{B}^{\text{pd}}_{X_1;X_2}$).

Let us consider the problem of comparing the methods of using two different f and g networks (Figure 4.1 (a)), and using one f with sign adjustments networks (Figure 4.1 (c)). That is to compare the following two optimization solutions,

$$\begin{aligned}
&\underset{\Phi\in\mathbb{R}^{N\times k},\Psi\in\mathbb{R}^{N\times k}}{\arg\min} && \|\check{B}_{Y_1Y_2} - \Phi\Psi^T\|_F^2, \\
&\underset{\Phi\in\mathbb{R}^{N\times k},D\in diag(k)}{\arg\min} && \|\check{B}_{Y_1Y_2} - \Phi D\Phi^T\|_F^2,
\end{aligned}$$

where we use the orthonormal column space of Φ matrix as the features.

Full rank $\tilde{B}_{X_1;X_2}$ and $\tilde{B}^{\text{sym}}_{X_1;X_2}$

First, let us assume no differences between the subspace of $\tilde{B}_{X;Y}, \tilde{B}_{Y_1;Y_2}$ and $\tilde{B}^{\text{sym}}_{Y_1;Y_2}$.

Let the SVD of $\tilde{B}_{XY} = U\Sigma\Phi^T$ ($\Phi \in \mathbb{R}^{|\mathcal{Y}|\times|\mathcal{X}|-1}$), and $\hat{\Phi}$ ($\hat{\Phi} \in \mathbb{R}^{|\mathcal{Y}|\times k}, k \geq |\mathcal{X}| - 1$) is obtained by one of the two methods. We plug it into the expression for Δ in (4.49):

$$\begin{aligned}
\Delta &= \|\tilde{B}_{XY}\Phi\|_F^2 - \|\tilde{B}_{XY}\hat{\Phi}\|_F^2 \\
&= \|\Sigma\Phi^T\Phi\|_F^2 - \|\Sigma\Phi^T\hat{\Phi}\|_F^2 \\
&= \sum_{i=1}^{|\mathcal{X}|-1} \sigma_i^2 - \sum_{i=1}^{|\mathcal{X}|-1} \sigma_i^2 \|\Phi_i^T\hat{\Phi}\|_2^2 \\
&= \sum_{i=1}^{|\mathcal{X}|-1} \sigma_i^2 (1 - \|\Phi_i^T\hat{\Phi}\|_2^2),
\end{aligned}$$

where Φ_i is the i-th column of the matrix Φ. $\|\Phi_i^T\hat{\Phi}\|_2^2 = \cos^2\theta_i$, where θ_i is the angle between the vector Φ_i and the subspace $\hat{\Phi}$. Therefore,

$$\begin{aligned}
\Delta &= \sum_{i=1}^{|\mathcal{X}|-1} \sigma_i^2 (1 - \|\Phi_i^T\hat{\Phi}\|_2^2) \\
&\leq \sigma_{\max}^2(\tilde{B}_{X;Y}) \sum_{i=1}^{|\mathcal{X}|-1} (1 - \|\Phi_i^T\hat{\Phi}\|_2^2) \\
&= \sigma_{\max}^2(\tilde{B}_{X;Y})(|\mathcal{X}| - 1 - \|(\Phi Q)^T\hat{\Phi}\|_2^2),
\end{aligned}$$

where Q is an orthogonal matrix such that the columns of ΦQ, Φ'_i, are the corresponding vectors in the column space of Φ that are selected to make up the principal angles. In this sense, we have

$$\begin{aligned}
\Delta &\leq \sigma_{\max}^2(\tilde{B}_{X;Y})(|\mathcal{X}| - 1 - \|(\Phi Q)^T\hat{\Phi}\|_2^2) \\
&= \sigma_{\max}^2(\tilde{B}_{XY}) \sum_{i=1}^{|\mathcal{X}|-1} (1 - \|\Phi_i'^T\hat{\Phi}\|_2^2) \\
&= \sigma_{\max}^2(\tilde{B}_{XY}) \sum_{i=1}^{|\mathcal{X}|-1} (1 - \cos^2\theta_i) \\
&= \sigma_{\max}^2(\tilde{B}_{XY}) \sum_{i=1}^{|\mathcal{X}|-1} \sin^2\theta_i \\
&= \sigma_{\max}^2(\tilde{B}_{XY}) \|\sin\Theta(\Phi, \hat{\Phi})\|_F^2,
\end{aligned}$$

where $\Theta(\Phi, \hat{\Phi})$ denotes the $|\mathcal{X}| - 1 \times |\mathcal{X}| - 1$ diagonal matrix whose i-th diagonal entry is the i-th principal angle, and let $\sin\Theta(\Phi, \hat{\Phi})$ be defined entrywise.

Notice that in our problem, $\Phi \in \mathbb{R}^{|\mathcal{Y}|\times|\mathcal{X}|-1}$, but $\hat{\Phi} \in \mathbb{R}^{|\mathcal{Y}|\times k} (k \geq |\mathcal{X}| - 1)$. So the

dimensions of these two matrices are different. However, if we let $\hat{\Phi}'$ to be the matrix that is composed of the first $|\mathcal{X}|-1$ columns of $\hat{\Phi}$, for each $i \in |\mathcal{X}|-1$, $\theta_i(\Phi, \hat{\Phi}') \geq \theta_i(\Phi, \hat{\Phi})$ since the column space of $\hat{\Phi}'$ is inside of the column space of $\hat{\Phi}$. Therefore, we have (without considering the effect of k)

$$\|\sin\Theta(\Phi, \hat{\Phi})\|_F^2 \leq \|\sin\Theta(\Phi, \hat{\Phi}')\|_F^2,$$

and

$$\|\tilde{B}_{X;Y}\Phi\|_F^2 - \|\tilde{B}_{X;Y}\hat{\Phi}\|_F^2 \leq \sigma_{\max}^2(\tilde{B}_{X;Y})\|\sin\Theta(\Phi, \hat{\Phi}')\|_F^2,$$

where $\Phi \in \mathbb{R}^{|\mathcal{Y}|\times|\mathcal{X}|-1}$ is the exact orthonormal column space of $\tilde{B}_{XY}$ as our target and $\hat{\Phi}' \in \mathbb{R}^{|\mathcal{Y}|\times|\mathcal{X}|-1}$ is the orthonormal column space calculated by one of our methods. $\Theta(\Phi, \hat{\Phi}')$ denotes the $|\mathcal{X}|-1 \times |\mathcal{X}|-1$ diagonal matrix whose i-th diagonal entry is the i-th principal angle, and let $\sin\Theta(\Phi, \hat{\Phi}')$ be defined entrywise.

There are several results to bound $\|\sin\Theta(\Phi, \hat{\Phi}')\|_F^2$. Two results from [101] are given here.

Theorem 4.2.9. Let $A, \hat{A} \in \mathbb{R}^{p\times p}$ be symmetric, with eigenvalues $\lambda_1 \geq \ldots \geq \lambda_p$ and $\hat{\lambda}_1 \geq \ldots \geq \hat{\lambda}_p$ respectively. Fix $1 \leq r \leq s \leq p$ and assume that $\min(\lambda_{r-1} - \lambda_r, \lambda_s - \lambda_{s+1}) > 0$, where $\lambda_0 := \infty$, and $\lambda_{p+1} := -\infty$. Let $d := s - r + 1$, and let $V = (v_r, v_{r+1}, \ldots, v_s) \in \mathbb{R}^{p\times d}$ and $\hat{V} = (\hat{v}_r, \hat{v}_{r+1}, \ldots, \hat{v}_s) \in \mathbb{R}^{p\times d}$ have orthonormal columns satisfying $Av_j = \lambda_j v_j$ and $\hat{A}\hat{v}_j = \hat{\lambda}_j\hat{v}_j$ for $j = r, r+1, \ldots, s$. Then,

$$\|\sin\Theta(V, \check{V})\|_F \leq \frac{2\min(d^{\frac{1}{2}}\|\hat{A} - A\|_{op}, \|\hat{A} - A\|_F)}{\min(\lambda_{r-1} - \lambda_r, \lambda_s - \lambda_{s+1})}.$$

Theorem 4.2.10. Let $A, \hat{A} \in \mathbb{R}^{p\times q}$ have singular values $\sigma_1 \geq \ldots \geq \sigma_{\min(p,q)}$ and $\hat{\sigma}_1 \geq \ldots \geq \hat{\sigma}_{\min(p,q)}$ respectively. Fix $1 \leq r \leq s \leq \text{rank}(A)$ and assume that $\min(\sigma_{r-1}^2 - \sigma_r^2, \sigma_s^2 - \sigma_{s+1}^2) > 0$, where $\sigma_0^2 := \infty$, and $\sigma_{p+1}^2 := -\infty$. Let $d := s - r + 1$, and let $V = (v_r, v_{r+1}, \ldots, v_s) \in \mathbb{R}^{q\times d}$ and $\hat{V} = (\hat{v}_r, \hat{v}_{r+1}, \ldots, \hat{v}_s) \in \mathbb{R}^{q\times d}$ have orthonormal columns satisfying $Av_j = \sigma_j u_j$ and $\hat{A}\hat{v}_j = \hat{\sigma}_j\hat{u}_j$ for $j = r, r+1, \ldots, s$. Then,

$$\|\sin\Theta(V, \check{V})\|_F \leq \frac{2(2\sigma_1 + \|\hat{A} - A\|_{op})\min(d^{\frac{1}{2}}\|\hat{A} - A\|_{op}, \|\hat{A} - A\|_F)}{\min(\sigma_{r-1}^2 - \sigma_r^2, \sigma_s^2 - \sigma_{s+1}^2)}.$$

Plug in the result for $f - g$ method (Figure 4.1 (a)) and f sign method (Figure 4.1 (c)),

we have [2],

$$\sqrt{\|\tilde{B}_{X;Y}\Phi\|_F^2 - \|\tilde{B}_{X;Y}\hat{\Phi}_{(\mathrm{a})}\|_F^2} \leq \sigma_{\max}(\tilde{B}_{X;Y})\times$$
$$\frac{2(2\sigma_{\max}(\tilde{B}_{Y_1;Y_2}) + \|\tilde{B}_{Y_1;Y_2} - \check{B}_{Y_1;Y_2}\|_{op})\min(\sqrt{|\mathcal{X}|-1}\|\tilde{B}_{Y_1;Y_2} - \check{B}_{Y_1;Y_2}\|_{op}, \|\tilde{B}_{Y_1;Y_2} - \check{B}_{Y_1;Y_2}\|_F)}{\sigma_{\min}^2(\tilde{B}_{Y_1;Y_2})},$$
$$\sqrt{\|\tilde{B}_{X;Y}\Phi\|_F^2 - \|\tilde{B}_{X;Y}\hat{\Phi}_{(\mathrm{a})}\|_F^2} \leq \sigma_{\max}(\tilde{B}_{X;Y}) \times \frac{4(\min(\sqrt{|\mathcal{X}|-1}\|\tilde{B}_{Y_1;Y_2} - \check{B}_{Y_1;Y_2}\|_{op}, \|\tilde{B}_{Y_1;Y_2} - \check{B}_{Y_1;Y_2}\|_F)}{\sigma_{\min}^2(\tilde{B}_{Y_1;Y_2})},$$
$$\sqrt{\|\tilde{B}_{X;Y}\Phi\|_F^2 - \|\tilde{B}_{X;Y}\hat{\Phi}_{(\mathrm{s})}\|_F^2} \leq \sigma_{\max}(\tilde{B}_{X;Y})\times$$
$$\frac{2(2\sigma_{\max}(\tilde{B}^{\mathrm{sym}}_{Y_1;Y_2}) + \|\tilde{B}^{\mathrm{sym}}_{Y_1;Y_2} - \check{B}^{\mathrm{sym}}_{Y_1;Y_2}\|_{op})\min(\sqrt{|\mathcal{X}|-1}\|\tilde{B}^{\mathrm{sym}}_{Y_1;Y_2} - \check{B}^{\mathrm{sym}}_{Y_1;Y_2}\|_{op}, \|\tilde{B}^{\mathrm{sym}}_{Y_1;Y_2} - \check{B}^{\mathrm{sym}}_{Y_1;Y_2}\|_F)}{\sigma_{\min}^2(B^{\mathrm{sym}}_{Y_1;Y_2})},$$

where $\check{\Phi}_{(\mathrm{a})}$ is the orthonormal feature space obtained through $f-g$ method (Figure 4.1 (a)) and $\check{\Phi}_{(\mathrm{s})}$ is the orthonormal feature space obtained through f sign method (Figure 4.1 (c)). The first and last inequalities are obtained by using Theorem 4.2.9, and the second one is obtained by using Theorem 4.2.10.

As we proved in the previous section, for $0 \leq \delta \leq 1$,

$$\mathbb{P}(\|\tilde{B}_{Y_1;Y_2} - \check{B}_{Y_1;Y_2}\|_{op} \geq \delta) \leq 2|\mathcal{Y}|\exp\Big(-\frac{P_0\delta^2 n}{4}\Big),$$
$$\mathbb{P}(\|\tilde{B}^{\mathrm{sym}}_{Y_1;Y_2} - \check{B}^{\mathrm{sym}}_{Y_1;Y_2}\|_{op} \geq \delta) \leq 2|\mathcal{Y}|\exp\Big(-\frac{P_0\delta^2 n}{4}\Big),$$

for $0 \leq \delta \leq \frac{1}{P_0\sqrt{2}}$,

$$\mathbb{P}(\|\tilde{B}_{Y_1;Y_2} - \check{B}_{Y_1;Y_2}\|_F \geq \delta) \leq \exp\Big(\frac{1}{4} - \frac{P_0^2\delta^2 n}{8}\Big),$$
$$\mathbb{P}(\|\tilde{B}^{\mathrm{sym}}_{Y_1;Y_2} - \check{B}^{\mathrm{sym}}_{Y_1;Y_2}\|_F \geq \delta) \leq \exp\Big(\frac{1}{4} - \frac{P_0^2\delta^2 n}{8}\Big),$$

where P_0 us defined as before.

Since the bounds for $\mathbb{P}(\|\tilde{B}_{Y_1;Y_2} - \check{B}_{Y_1;Y_2}\|_{op} \geq \delta)$ and $\mathbb{P}(\|\tilde{B}^{\mathrm{sym}}_{Y_1;Y_2} - \check{B}^{\mathrm{sym}}_{Y_1;Y_2}\|_{op}$ are the same, so are $\mathbb{P}(\|\tilde{B}_{Y_1;Y_2} - \check{B}_{Y_1;Y_2}\|_F \geq \delta)$ and $\mathbb{P}(\|\tilde{B}^{\mathrm{sym}}_{Y_1;Y_2} - \check{B}^{\mathrm{sym}}_{Y_1;Y_2}\|_F \geq \delta)$, the difference in bounds for these 2 methods in the case when $\tilde{B}_{X_1;X_2}$ and $\tilde{B}^{\mathrm{sym}}_{X_1;X_2}$ are full rank comes from the difference between $\frac{\sigma_{\max}(\tilde{B}_{Y_1;Y_2})}{\sigma^2_{\min}(\tilde{B}_{Y_1;Y_2})}$ and $\frac{\sigma_{\max}(\tilde{B}^{\mathrm{sym}}_{Y_1;Y_2})}{\sigma^2_{\min}(\tilde{B}^{\mathrm{sym}}_{Y_1;Y_2})}$.

We calculate the exact bound for these two methods in the following. There are two

[2] Theorem 4.2.9 requires the order of eigenvalues while a symmetric matrix may have negative eigenvalues, so it is hard to apply directly on $\tilde{B}^{\mathrm{sym}}_{Y_1;Y_2}$.

inequalities used in the process to bound the relation:

$$\|A\|_F \geq \|A\|_{op}, \forall \text{ matrix } A,$$
$$\|A\|_{op} \geq \frac{1}{k}\|A\|_F, \text{where } k = \text{rank}(A).$$

With some calculations, we have

$$\begin{aligned}
&\mathbb{P}(\|\tilde{B}_{X;Y}\Phi\|_F^2 - \|\tilde{B}_{X;Y}\hat{\Phi}_{(\text{a})}\|_F^2 \geq \delta) \\
\leq &\min\Bigg(2|\mathcal{Y}|\exp\Bigg(-\frac{P_0 n}{4}\Bigg(\sqrt{\sigma_{\max}^2(\tilde{B}_{Y_1;Y_2}) + \frac{\sqrt{\delta}\sigma_{\min}^2(B_{Y_1;Y_2})}{2\sqrt{|\mathcal{X}|-1}\sigma_{\max}(\tilde{B}_{X;Y})}} - \sigma_{\max}(\tilde{B}_{Y_1;Y_2})\Bigg)^2\Bigg), \\
&\exp\Bigg(\frac{1}{4} - \frac{P_0^2 n}{8}\Bigg(\sqrt{\sigma_{\max}^2(\tilde{B}_{Y_1;Y_2}) + \frac{\sqrt{\delta}\sigma_{\min}^2(\tilde{B}_{Y_1;Y_2})}{2\sigma_{\max}(\tilde{B}_{X;Y})}} - \sigma_{\max}(\tilde{B}_{Y_1;Y_2})\Bigg)^2\Bigg), \\
&2|\mathcal{Y}|\exp\Bigg(-\frac{P_0 n\delta\sigma_{\min}^4(\tilde{B}_{Y_1;Y_2})}{64(|\mathcal{X}|-1)\sigma_{\max}^2(\tilde{B}_{X;Y})}\Bigg), \exp\Bigg(\frac{1}{4} - \frac{P_0^2 n\delta\sigma_{\min}^4(\tilde{B}_{Y_1;Y_2})}{128\sigma_{\max}^2(\tilde{B}_{X;Y})}\Bigg)\Bigg). \\
&\mathbb{P}(\|\tilde{B}_{X;Y}\Phi\|_F^2 - \|\tilde{B}_{X;Y}\hat{\Phi}_{(\text{s})}\|_F^2 \geq \delta) \\
\leq &\min\Bigg(2|\mathcal{Y}|\exp\Bigg(-\frac{P_0 n}{4}\Bigg(\sqrt{\sigma_{\max}^2(\tilde{B}^{\text{sym}}_{Y_1;Y_2}) + \frac{\sqrt{\delta}\sigma_{\min}^2(\tilde{B}^{\text{sym}}_{Y_1;Y_2})}{2\sqrt{|\mathcal{X}|-1}\sigma_{\max}(\tilde{B}_{X;Y})}} - \sigma_{\max}(\tilde{B}^{\text{sym}}_{Y_1;Y_2})\Bigg)^2\Bigg), \\
&\exp\Bigg(\frac{1}{4} - \frac{P_0^2 n}{8}\Bigg(\sqrt{\sigma_{\max}^2(\tilde{B}^{\text{sym}}_{Y_1;Y_2}) + \frac{\sqrt{\delta}\sigma_{\min}^2(\tilde{B}^{\text{sym}}_{Y_1;Y_2})}{2\sigma_{\max}(\tilde{B}_{X;Y})}} - \sigma_{\max}(\tilde{B}^{\text{sym}}_{Y_1;Y_2})\Bigg)^2\Bigg)\Bigg)
\end{aligned}$$

Full rank $\tilde{\boldsymbol{B}}_{\boldsymbol{X_1;X_2}}$ and rank deficient $\tilde{\boldsymbol{B}}^{\text{sym}}_{\boldsymbol{X_1;X_2}}$

In this case, there are no differences between the subspace of $\tilde{B}_{X;Y}$ and $\tilde{B}_{Y_1;Y_2}$, but $\tilde{B}^{\text{sym}}_{Y_1;Y_2}$ is in a lower subspace of $\tilde{B}_{X;Y}$. Therefore, the conclusion for $f-g$ method (Figure 4.1 (a)) is unchanged. To analyze the metric for f sign method (Figure 4.1 (c)), again we have

$$
\begin{aligned}
\Delta &= \sum_{i=1}^{|\mathcal{X}|-1} \sigma_i^2 (1 - \|\Phi_i^T \hat{\Phi}_{(\mathrm{s})}\|_2^2) \\
&\leq \sigma_{\max}^2(\tilde{B}_{X;Y}) \sum_{i=1}^{|\mathcal{X}|-1} (1 - \|\Phi_i^T \hat{\Phi}_{(\mathrm{s})}\|_2^2) \\
&\leq (|\mathcal{X}| - \mathrm{rank}(\tilde{B}^{\mathrm{sym}}_{Y_1;Y_2}))\sigma_{\max}^2(\tilde{B}_{X;Y}) + \sigma_{\max}^2(\tilde{B}_{X;Y}) \sum_{i=1}^{\mathrm{rank}(\tilde{B}^{\mathrm{sym}}_{Y_1;Y_2})} (1 - \|\Phi_i^T \hat{\Phi}_{(\mathrm{s})}\|_2^2) \\
&= (|\mathcal{X}| - \mathrm{rank}(\tilde{B}^{\mathrm{sym}}_{Y_1;Y_2}))\sigma_{\max}^2(\tilde{B}_{X;Y}) + \sigma_{\max}^2(\tilde{B}_{X;Y})(\mathrm{rank}(\tilde{B}^{\mathrm{sym}}_{Y_1;Y_2}) - \|(\Phi_{\mathrm{trc}} Q)^T \hat{\Phi}_{(\mathrm{s})}\|_2^2), \\
&= (|\mathcal{X}| - \mathrm{rank}(\tilde{B}^{\mathrm{sym}}_{Y_1;Y_2}))\sigma_{\max}^2(\tilde{B}_{X;Y}) + \sigma_{\max}^2(\tilde{B}_{X;Y}) \sum_{i=1}^{\mathrm{rank}(\tilde{B}^{\mathrm{sym}}_{Y_1;Y_2})} (1 - \cos^2\theta_i) \\
&= (|\mathcal{X}| - \mathrm{rank}(\tilde{B}^{\mathrm{sym}}_{Y_1;Y_2}))\sigma_{\max}^2(\tilde{B}_{X;Y}) + \sigma_{\max}^2(\tilde{B}_{X;Y}) \sum_{i=1}^{\mathrm{rank}(\tilde{B}^{\mathrm{sym}}_{Y_1;Y_2})} \sin^2\theta_i \\
&= (|\mathcal{X}| - \mathrm{rank}(\tilde{B}^{\mathrm{sym}}_{Y_1;Y_2}))\sigma_{\max}^2(\tilde{B}_{X;Y}) + \sigma_{\max}^2(\tilde{B}_{X;Y}) \| \sin\Theta(\Phi_{\mathrm{trc}}, \hat{\Phi}_{(\mathrm{s})})\|_F^2,
\end{aligned}
$$

where Φ_{trc} is the truncated version of Φ which only contains the first $\mathrm{rank}(\tilde{B}^{\mathrm{sym}}_{Y_1;Y_2})$ columns of Φ, Q is an orthogonal matrix such that the columns of $\Phi_{\mathrm{trc}}Q$, Φ'_i, are the corresponding vectors in the column space of Φ_{trc} that are selected to make up the principal angles, $\Theta(\Phi_{\mathrm{trc}}, \hat{\Phi}_{(\mathrm{s})})$ denotes the $\mathrm{rank}(\tilde{B}^{\mathrm{sym}}_{Y_1;Y_2}) \times \mathrm{rank}(\tilde{B}^{\mathrm{sym}}_{Y_1;Y_2})$ diagonal matrix whose i-th diagonal entry is the i-th principal angle, and let $\sin\Theta(\Phi_{\mathrm{trc}}, \hat{\Phi}_{(\mathrm{s})})$ be defined entrywise.

Notice that in our problem, $\Phi_{\mathrm{trc}} \in \mathbb{R}^{|\mathcal{Y}| \times \mathrm{rank}(\tilde{B}^{\mathrm{sym}}_{Y_1;Y_2})}$, but $\hat{\Phi}_{(\mathrm{s})} \in \mathbb{R}^{|\mathcal{Y}| \times k} (k \geq |\mathcal{X}| - 1)$. So the dimensions of these two matrices are different. However, if we let $\hat{\Phi}'_{(\mathrm{s})}$ to be the matrix that is composed of the first $\mathrm{rank}(\tilde{B}^{\mathrm{sym}}_{Y_1;Y_2})$ columns of $\hat{\Phi}_{(\mathrm{s})}$, for each $i \in \mathrm{rank}(\tilde{B}^{\mathrm{sym}}_{Y_1;Y_2})$, $\theta_i(\Phi_{\mathrm{trc}}, \hat{\Phi}'_{(\mathrm{s})}) \geq \theta_i(\Phi_{\mathrm{trc}}, \hat{\Phi}_{(\mathrm{s})})$ since the column space of $\hat{\Phi}'_{(\mathrm{s})}$ is inside of the column space of $\hat{\Phi}_{(\mathrm{s})}$. Therefore, we have (without considering the effect of k)

$$
\| \sin\Theta(\Phi_{\mathrm{trc}}, \hat{\Phi}_{(\mathrm{s})})\|_F^2 \leq \| \sin\Theta(\Phi_{\mathrm{trc}}, \hat{\Phi}'_{(\mathrm{s})})\|_F^2,
$$

and

$$
\|\tilde{B}_{X;Y}\Phi\|_F^2 - \|\tilde{B}_{X;Y}\hat{\Phi}_{(\mathrm{s})}\|_F^2 \leq (|\mathcal{X}| - \mathrm{rank}(\tilde{B}^{\mathrm{sym}}_{Y_1;Y_2}))\sigma_{\max}^2(\tilde{B}_{X;Y}) + \sigma_{\max}^2(\tilde{B}_{X;Y}) \| \sin\Theta(\Phi_{\mathrm{trc}}, \hat{\Phi}'_{(\mathrm{s})})\|_F^2,
$$

where $\Phi_{\mathrm{trc}} \in \mathbb{R}^{|\mathcal{Y}| \times \mathrm{rank}(\tilde{B}^{\mathrm{sym}}_{Y_1;Y_2})}$ is composed of the first $\mathrm{rank}(\tilde{B}^{\mathrm{sym}}_{Y_1;Y_2})$ columns of the exact orthonormal column space of $\tilde{B}_{X;Y}$ as our target and $\hat{\Phi}'_{(\mathrm{s})} \in \mathbb{R}^{|\mathcal{Y}| \times \mathrm{rank}(\tilde{B}^{\mathrm{sym}}_{Y_1;Y_2})}$ is the orthonormal column space calculated by one of our methods. $\Theta(\Phi_{\mathrm{trc}}, \hat{\Phi}'_{(\mathrm{s})})$ denotes the $\mathrm{rank}(\tilde{B}^{\mathrm{sym}}_{Y_1;Y_2}) \times \mathrm{rank}(\tilde{B}^{\mathrm{sym}}_{Y_1;Y_2})$ diagonal matrix whose i-th diagonal entry is the i-th principal angle, and let

$\sin\Theta(\Phi_{\text{trc}}, \hat{\Phi}'_{(\text{s})})$ be defined entrywise.

Again, using Theorem 4.2.10 to bound the metric, we have

$$\|\tilde{B}_{X;Y}\Phi\|_F^2 - \|\tilde{B}_{X;Y}\hat{\Phi}_{(\text{s})}\|_F^2 \leq (|\mathcal{X}| - \text{rank}(\tilde{B}^{\text{sym}}_{Y_1;Y_2}))\sigma^2_{\max}(\tilde{B}_{X;Y}) + \sigma^2_{\max}(\tilde{B}_{X;Y})\times$$
$$\Big(\frac{2(2\sigma_{\max}(\tilde{B}^{\text{sym}}_{Y_1;Y_2}) + \|\tilde{B}^{\text{sym}}_{Y_1;Y_2} - \check{B}^{\text{sym}}_{Y_1;Y_2}\|_{op})\min(\sqrt{\text{rank}(\tilde{B}^{\text{sym}}_{Y_1;Y_2}) - 1}\|\tilde{B}^{\text{sym}}_{Y_1;Y_2} - \check{B}^{\text{sym}}_{Y_1;Y_2}\|_{op}, \|\tilde{B}^{\text{sym}}_{Y_1;Y_2} - \check{B}^{\text{sym}}_{Y_1;Y_2}\|_F)}{\sigma^2_{\min}(\tilde{B}^{\text{sym}}_{Y_1;Y_2})}\Big)^2.$$

For comparison, the result for $f-g$ method (Figure 4.1 (a)) is

$$\sqrt{\|\tilde{B}_{X;Y}\Phi\|_F^2 - \|\tilde{B}_{X;Y}\hat{\Phi}_{(\text{a})}\|_F^2} \leq \sigma_{\max}(\tilde{B}_{X;Y})\times$$
$$\frac{2(2\sigma_{\max}(\tilde{B}_{Y_1;Y_2}) + \|\tilde{B}_{Y_1;Y_2} - \check{B}_{Y_1;Y_2}\|_{op})\min(\sqrt{|\mathcal{X}| - 1}\|\tilde{B}_{Y_1;Y_2} - \check{B}_{Y_1;Y_2}\|_{op}, \|\tilde{B}_{Y_1;Y_2} - \check{B}_{Y_1;Y_2}\|_F)}{\sigma^2_{\min}(\tilde{B}_{Y_1;Y_2})},$$
$$\sqrt{\|\tilde{B}_{X;Y}\Phi\|_F^2 - \|\tilde{B}_{X;Y}\hat{\Phi}_{(\text{a})}\|_F^2} \leq \sigma_{\max}(\tilde{B}_{X;Y}) \times \frac{4(\min(\sqrt{|\mathcal{X}| - 1}\|\tilde{B}_{Y_1;Y_2} - \check{B}_{Y_1;Y_2}\|_{op}, \|\tilde{B}_{Y_1;Y_2} - \check{B}_{Y_1;Y_2}\|_F)}{\sigma^2_{\min}(\tilde{B}_{Y_1;Y_2})},$$

With some calculations, we have

$$\mathbb{P}(\|\tilde{B}_{X;Y}\Phi\|_F^2 - \|\tilde{B}_{X;Y}\hat{\Phi}_{(\text{a})}\|_F^2 \geq \delta)$$
$$\leq \min\Bigg(2|\mathcal{Y}|\exp\Big(-\frac{P_0 n}{4}\Big(\sqrt{\sigma^2_{\max}(\tilde{B}_{Y_1;Y_2}) + \frac{\sqrt{\delta}\sigma^2_{\min}(\tilde{B}_{Y_1;Y_2})}{2\sqrt{|\mathcal{X}| - 1}\sigma_{\max}(\tilde{B}_{X;Y})}} - \sigma_{\max}(\tilde{B}_{Y_1;Y_2})\Big)^2\Big),$$
$$\exp\Big(\frac{1}{4} - \frac{P_0^2 n}{8}\Big(\sqrt{\sigma^2_{\max}(\tilde{B}_{Y_1;Y_2}) + \frac{\sqrt{\delta}\sigma^2_{\min}(\tilde{B}_{Y_1;Y_2})}{2\sigma_{\max}(\tilde{B}_{X;Y})}} - \sigma_{\max}(\tilde{B}_{Y_1;Y_2})\Big)^2\Big),$$
$$2|\mathcal{Y}|\exp\Big(-\frac{P_0 n\delta\sigma^4_{\min}(\tilde{B}_{Y_1;Y_2})}{64(|\mathcal{X}| - 1)\sigma^2_{\max}(\tilde{B}_{X;Y})}\Big), \exp\Big(\frac{1}{4} - \frac{P_0^2 n\delta\sigma^4_{\min}(\tilde{B}_{Y_1;Y_2})}{128\sigma^2_{\max}(\tilde{B}_{X;Y})}\Big)\Bigg).$$

$$\mathbb{P}(\|\tilde{B}_{X;Y}\Phi\|_F^2 - \|\tilde{B}_{X;Y}\hat{\Phi}_{(\text{s})}\|_F^2 \geq \delta)$$
$$\leq \min\Bigg(2|\mathcal{Y}|\exp\Big(-\frac{P_0 n}{4}\Big(\sqrt{\sigma^2_{\max}(\tilde{B}^{\text{sym}}_{Y_1;Y_2}) + \frac{\sqrt{\delta - (|\mathcal{X}| - \text{rank})\sigma^2_{\max}(\tilde{B}_{X;Y})}\sigma^2_{\min}(\tilde{B}^{\text{sym}}_{Y_1;Y_2})}{2\sqrt{\text{rank} - 1}\sigma_{\max}(\tilde{B}_{X;Y})}} - \sigma_{\max}(\tilde{B}^{\text{sym}}_{Y_1;Y_2})\Big)^2\Big)$$
$$\exp\Big(\frac{1}{4} - \frac{P_0^2 n}{8}\Big(\sqrt{\sigma^2_{\max}(\tilde{B}^{\text{sym}}_{Y_1;Y_2}) + \frac{\sqrt{\delta - (|\mathcal{X}| - \text{rank})\sigma^2_{\max}(\tilde{B}_{X;Y})}\sigma^2_{\min}(\tilde{B}^{\text{sym}}_{Y_1;Y_2})}{2\sigma_{\max}(\tilde{B}_{X;Y})}} - \sigma_{\max}(\tilde{B}^{\text{sym}}_{Y_1;Y_2})\Big)^2\Big)\Bigg)$$

where $\text{rank} = \text{rank}(\tilde{B}^{\text{sym}}_{Y_1;Y_2})$.

In conclusion, we have the following theorem for the sample complexity analysis in the $X-Y$ space:

Theorem 4.2.11. [3] For $\hat{\Phi}_{(\text{a})}$ obtained by different $f-g$ (Figure 4.1 (a)) feature functions,

[3]

we have

$$\begin{aligned}
&\mathbb{P}(\|\tilde{B}_{X;Y}\Phi\|_F^2 - \|\tilde{B}_{X;Y}\hat{\Phi}_{(\mathrm{a})}\|_F^2 \geq \delta) \\
\leq &\min\Bigg(2|\mathcal{Y}|\exp\Bigg(-\frac{P_0 n}{4}\Bigg(\sqrt{\sigma_{\max}^2(\tilde{B}_{Y_1;Y_2}) + \frac{\sqrt{\delta}\sigma_{\min}^2(B_{Y_1;Y_2})}{2\sqrt{|\mathcal{X}|-1}\sigma_{\max}(\tilde{B}_{X;Y})}} - \sigma_{\max}(\tilde{B}_{Y_1;Y_2})\Bigg)^2\Bigg), \\
&\exp\Bigg(\frac{1}{4} - \frac{P_0^2 n}{8}\Bigg(\sqrt{\sigma_{\max}^2(\tilde{B}_{Y_1;Y_2}) + \frac{\sqrt{\delta}\sigma_{\min}^2(\tilde{B}_{Y_1;Y_2})}{2\sigma_{\max}(\tilde{B}_{X;Y})}} - \sigma_{\max}(\tilde{B}_{Y_1;Y_2})\Bigg)^2\Bigg), \\
&2|\mathcal{Y}|\exp\Bigg(-\frac{P_0 n\delta\sigma_{\min}^4(\tilde{B}_{Y_1;Y_2})}{64(|\mathcal{X}|-1)\sigma_{\max}^2(\tilde{B}_{X;Y})}\Bigg), \exp\Bigg(\frac{1}{4} - \frac{P_0^2 n\delta\sigma_{\min}^4(\tilde{B}_{Y_1;Y_2})}{128\sigma_{\max}^2(\tilde{B}_{X;Y})}\Bigg)\Bigg).
\end{aligned}$$

For $\hat{\Phi}_{(\mathrm{pd})}$ obtained by one f feature function (Figure 4.1 (b)), we have

$$\begin{aligned}
&\mathbb{P}(\|\tilde{B}_{X;Y}\Phi\|_F^2 - \|\tilde{B}_{X;Y}\hat{\Phi}_{(\mathrm{pd})}\|_F^2 \geq \delta) \\
\leq &\min\Bigg(2|\mathcal{Y}|\exp\Bigg(-\frac{P_0 n}{4}\Bigg(\sqrt{\sigma_{\max}^2(\tilde{B}_{Y_1;Y_2}^{\mathrm{pd}}) + \frac{\sqrt{\delta - (|\mathcal{X}| - \mathrm{rank}_1)\sigma_{\max}^2(\tilde{B}_{X;Y})}\sigma_{\min}^2(\tilde{B}_{Y_1;Y_2}^{\mathrm{pd}})}{2\sqrt{\mathrm{rank}_1 - 1}\sigma_{\max}(\tilde{B}_{X;Y})}} - \sigma_{\max}(\tilde{B}_{Y_1;Y_2}^{\mathrm{pd}})\Bigg)^2\Bigg), \\
&\exp\Bigg(\frac{1}{4} - \frac{P_0^2 n}{8}\Bigg(\sqrt{\sigma_{\max}^2(\tilde{B}_{Y_1;Y_2}^{\mathrm{pd}}) + \frac{\sqrt{\delta - (|\mathcal{X}| - \mathrm{rank}_1)\sigma_{\max}^2(\tilde{B}_{X;Y})}\sigma_{\min}^2(\tilde{B}_{Y_1;Y_2}^{\mathrm{pd}})}{2\sigma_{\max}(\tilde{B}_{X;Y})}} - \sigma_{\max}(\tilde{B}_{Y_1;Y_2}^{\mathrm{pd}})\Bigg)^2\Bigg), \\
&2|\mathcal{Y}|\exp\Bigg(-\frac{P_0 n(\delta - (|\mathcal{X}| - \mathrm{rank}_1)\sigma_{\max}^2(\tilde{B}_{X;Y}))\sigma_{\min}^4(\tilde{B}_{Y_1;Y_2}^{\mathrm{pd}})}{64(\mathrm{rank}_1 - 1)\sigma_{\max}^2(\tilde{B}_{X;Y})}\Bigg), \\
&\exp\Bigg(\frac{1}{4} - \frac{P_0^2 n(\delta - (|\mathcal{X}| - \mathrm{rank}_1)\sigma_{\max}^2(\tilde{B}_{X;Y}))\sigma_{\min}^4(\tilde{B}_{Y_1;Y_2}^{\mathrm{pd}})}{128\sigma_{\max}^2(\tilde{B}_{X;Y})}\Bigg)\Bigg).
\end{aligned}$$

For $\hat{\Phi}_{(\mathrm{s})}$ obtained by one f feature function with sign adjustments (Figure 4.1 (c)), we

- $\|\tilde{B} - \hat{B}\|_{op/F}$ is same for $\tilde{B}$, $\tilde{B}^{\mathrm{sym}}$, and $\tilde{B}^{\mathrm{pd}}$, so we need to make sure three methods give different $\|\sin\Theta(\Phi, \hat{\Phi})\|_F$.
- Note that Theorem 4.2.9 requires the order of eigenvalues while a symmetric matrix may have negative eigenvalues, so it is hard to apply directly on $\tilde{B}_{Y_1;Y_2}^{\mathrm{sym}}$.
- The ideal δ should be smaller than $\|\tilde{B}_{X;Y}\Phi\|_F^2$. Otherwise, it is not reasonable.

have

$$\mathbb{P}(\|\tilde{B}_{X;Y}\Phi\|_F^2 - \|\tilde{B}_{X;Y}\hat{\Phi}_{(\mathrm{s})}\|_F^2 \geq \delta)$$
$$\leq \min\Bigg(2|\mathcal{Y}|\exp\Bigg(-\frac{P_0 n}{4}\Bigg(\sqrt{\sigma_{\max}^2(\tilde{B}_{Y_1;Y_2}^{\mathrm{sym}}) + \frac{\sqrt{\delta - (|\mathcal{X}| - \mathrm{rank}_2)\sigma_{\max}^2(\tilde{B}_{X;Y})}\sigma_{\min}^2(\tilde{B}_{Y_1;Y_2}^{\mathrm{sym}})}{2\sqrt{\mathrm{rank}_2 - 1}\sigma_{\max}(\tilde{B}_{X;Y})}} - \sigma_{\max}(\tilde{B}_{Y_1;Y_2}^{\mathrm{sym}})\Bigg)^2\Bigg),$$
$$\exp\Bigg(\frac{1}{4} - \frac{P_0^2 n}{8}\Bigg(\sqrt{\sigma_{\max}^2(\tilde{B}_{Y_1;Y_2}^{\mathrm{sym}}) + \frac{\sqrt{\delta - (|\mathcal{X}| - \mathrm{rank}_2)\sigma_{\max}^2(\tilde{B}_{X;Y})}\sigma_{\min}^2(\tilde{B}_{Y_1;Y_2}^{\mathrm{sym}})}{2\sigma_{\max}(\tilde{B}_{X;Y})}} - \sigma_{\max}(\tilde{B}_{Y_1;Y_2}^{\mathrm{sym}})\Bigg)^2\Bigg)\Bigg)$$

where $\mathrm{rank}_1 = \mathrm{rank}(\tilde{B}_{Y_1;Y_2}^{\mathrm{pd}}), \mathrm{rank}_2 = \mathrm{rank}(\tilde{B}_{Y_1;Y_2}^{\mathrm{sym}})$, $P_{Y_1}(y) = P_{Y_2}(y) \geq P_0, \forall y \in \mathcal{Y}$ for some P_0, and we require $\delta \geq (|\mathcal{X}| - \mathrm{rank}(\tilde{B}_{Y_1;Y_2}^{\mathrm{pd}}))\sigma_{\max}^2(\tilde{B}_{X;Y})$. Otherwise, if $\delta < (|\mathcal{X}| - \mathrm{rank}(\tilde{B}_{Y_1;Y_2}^{\mathrm{pd}}))\sigma_{\max}^2(\tilde{B}_{X;Y})$, we have $\mathbb{P}(\|\tilde{B}_{X;Y}\Phi\|_F^2 - \|\tilde{B}_{XY}\hat{\Phi}\|_F^2 \geq \delta) \leq 1$. Also note that for $\delta > \|\tilde{B}_{X;Y}\|_F^2$, $\mathbb{P}(\|\tilde{B}_{X;Y}\Phi\|_F^2 - \|\tilde{B}_{X;Y}\hat{\Phi}\|_F^2 \geq \delta) = 0$.

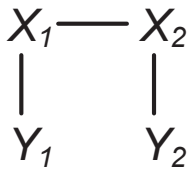
Chapter 5

Simulation Results

In this part, we implement our proposed three models on different tasks to verify the effectiveness of the proposed algorithm. More specifically, we show the trade-off between the number of learned parameters and the accuracy of the features.

5.1 $\tilde{B}_{X_1;X_2}$ is Positive Definite

In the first scenario, we have the following experiment setup:



- $|\mathcal{X}| = 4, |\mathcal{Y}| = 8$
- $(y_1^{(i)}, y_2^{(i)})_{i=1}^n$ comes from a homogeneous HMM where the state diagram for MC is shown in Figure 5.1. That is, the self transition probability is $\frac{5}{8}$ while the transition probability between states is $\frac{1}{8}$.

It is obvious that this structure is fully symmetric, and with a careful analysis, the matrix $\tilde{B}_{X_1;X_2}$ actually has no negative eigenvalues. The transition matrix of $X - Y$ link is set to be

$$P_{Y|X} = \begin{bmatrix} 0.3 & 0.1 & 0.07 & 0.03 & 0.125 & 0.125 & 0.125 & 0.125 \\ 0.07 & 0.03 & 0.3 & 0.1 & 0.125 & 0.125 & 0.125 & 0.125 \\ 0.03 & 0.07 & 0.1 & 0.3 & 0.125 & 0.125 & 0.125 & 0.125 \\ 0.1 & 0.3 & 0.03 & 0.07 & 0.125 & 0.125 & 0.125 & 0.125 \end{bmatrix},$$

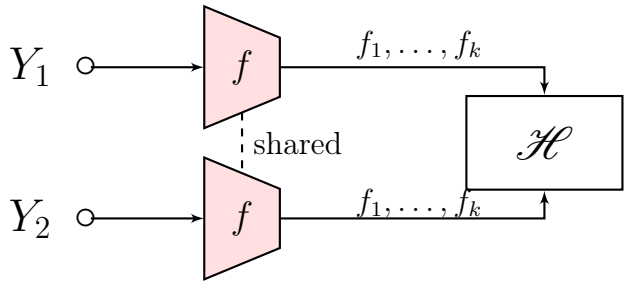
(a) (b)

Figure 5.1: (a) State diagram for scenario 1; (b) Transition probabilities for scenario 1.

so that the 3 feature functions have the following patterns

$$\boldsymbol{\phi}_1 = \begin{bmatrix} 1 \\ 1 \\ -1 \\ -1 \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix}, \boldsymbol{\phi}_2 = \begin{bmatrix} 1 \\ -1 \\ 1 \\ -1 \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix}, \boldsymbol{\phi}_3 = \begin{bmatrix} 1 \\ -1 \\ -1 \\ 1 \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix}.$$

Because of the nice positive definiteness property of this problem, the features can be recovered by any of our proposed structures. Specifically, if we use the following network (the most efficient one):



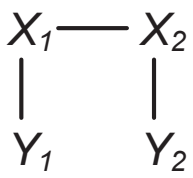
we have the NN output as follows:

$$\mathbf{F} = \begin{pmatrix} 1.4372609 & -7.289173e^{-1} & -9.523574e^{-1} \\ 1.4255644 & 7.341049e^{-1} & 9.4092274e^{-1} \\ -1.430149 & -7.591157e^{-1} & 9.408605e^{-1} \\ -1.4549847 & 7.5401753e^{-1} & -9.3699163e^{-1} \\ 2.170068e^{-2} & -1.64817e^{-3} & -2.621175e^{-2} \\ -6.47862e^{-3} & 1.142581e^{-2} & 2.345994e^{-2} \\ 5.59115e^{-3} & -1.45771e^{-3} & 1.530454e^{-2} \\ 2.27817e^{-3} & -1.900343e^{-2} & 3.07287e^{-3} \end{pmatrix}$$

This result makes sense since it is in the form of $\mathbf{F} = \begin{pmatrix} c_1\phi_1 & c_2\phi_2 & c_3\phi_3 \end{pmatrix}$, where c_1, c_2, and c_3 are some constants. If we analyze the corresponding $\Delta_{(\mathsf{pd})} = \|\tilde{B}_{X;Y}\Phi\|_F^2 - \|\tilde{B}_{X;Y}\hat{\Phi}_{(\mathsf{pd})}\|_F^2$, we have $\Delta_{(\mathsf{pd})} = 0.00184$. More specifically, if we calculate the percentage change, we have that $\frac{\Delta_{(\mathsf{pd})}}{\|\tilde{B}_{X;Y}\Phi\|_F^2} = 0.6\%$, i.e., this model helps to recover 99.4% of the total information of interest.

5.2 $\tilde{B}_{X_1;X_2}$ is not Positive Definite but $\tilde{B}^{\mathbf{sym}}_{X_1;X_2}$ has Full Rank

In the second scenario, we have the following experiment setup:



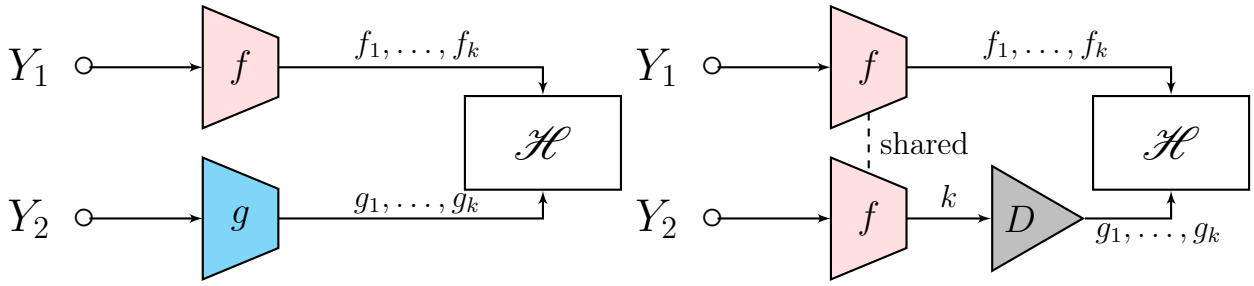
- $|\mathcal{X}| = 4, |\mathcal{Y}| = 8$
- $(y_1^{(i)}, y_2^{(i)})_{i=1}^n$ comes from a homogeneous HMM where the state diagram for MC is shown in Figure 5.2. That is, the self transition probability is $\frac{1}{7}$ while the transition probability between states is $\frac{2}{7}$.

In this scenario, $\tilde{B}^{\text{sym}}_{X_1;X_2}$ is no longer positive definite and has negative eigenvalues. Although $\tilde{B}^{\text{pd}}_{X_1;X_2}$ does not capture all features, $\tilde{B}^{\text{sym}}_{X_1;X_2}$ does. Therefore, we are able to use the following two structures to obtain the features:

More specifically, the right one requires less samples to train and more efficient. If we

(a) (b)

Figure 5.2: (a) State diagram for scenario 2; (b) Transition probabilities for scenario 2.



choose the right one, we have the NN output as follows:

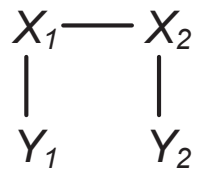
$$\mathbf{F} = \begin{pmatrix} -1.6857454 & -0.49852186 & -1.46100596 \\ -1.53533066 & 0.48129386 & 1.3207045 \\ 1.66136402 & -0.5933844 & 1.30925435 \\ 1.5948398 & 0.57050353 & -1.3588086 \\ -7.822638e^{-2} & -7.464907e^{-2} & -7.787842e^{-2} \\ 4.05275e^{-2} & -6.96193e^{-3} & -6.480303e^{-2} \\ 1.1709695e^{-1} & -2.03253e^{-2} & 4.945641e^{-2} \\ -9.799867e^{-2} & 3.323984e^{-2} & -9.56082e^{-3} \end{pmatrix}$$

This result makes sense since it is again in the form of $\mathbf{F} = \begin{pmatrix} c_1\phi_1 & c_2\phi_2 & c_3\phi_3 \end{pmatrix}$, where c_1, c_2, and c_3 are some constants other than before. If we analyze the corresponding $\Delta_{(\mathsf{s})} = \|\tilde{B}_{X;Y}\Phi\|_F^2 - \|\tilde{B}_{X;Y}\hat{\Phi}_{(\mathsf{sym})}\|_F^2$, we have $\Delta_{(\mathsf{s})} = 0.00454$. More specifically, if we calculate the percentage change, we have that $\frac{\Delta_{(\mathsf{s})}}{\|\tilde{B}_{X;Y}\Phi\|_F^2} = 1.6\%$, i.e., this model helps to recover 98.4% of the total information of interest.

5.3 $\tilde{B}_{X_1;X_2}$ is Rank Deficient

In the last scenario, we have the following experiment setup:

- $|\mathcal{X}| = 4, |\mathcal{Y}| = 8$
- $(y_1^{(i)}, y_2^{(i)})_{i=1}^n$ comes from a homogeneous HMM where the state diagram for MC is

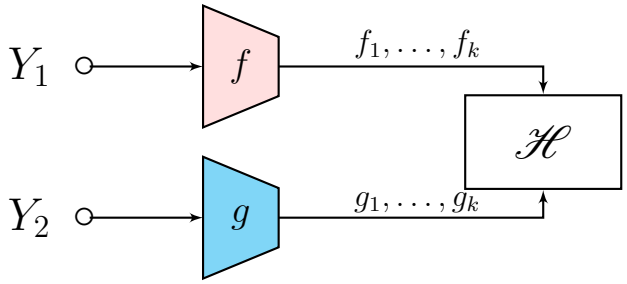


shown in Figure 5.4. That is, the self transition probability is $\frac{1}{3}$ while the transition probability between states is $\frac{1}{3}$, $\frac{1}{4}$, or $\frac{1}{8}$ depending on the initial and ending state.

(a) (b)

Figure 5.4: (a) State diagram for scenario 3; (b) Transition probabilities for scenario 3.

In this scenario, after careful analysis, $\tilde{B}^{\text{sym}}_{X_1;X_2}$ only has 1 non-zero eigenvalue corresponding to 1 mode, while there are 3 modes in total. Although $\tilde{B}^{\text{sym}}_{X_1;X_2}$ does not capture all features, $\tilde{B}_{X_1;X_2}$ does. These features can only be recovered by the first proposed structure:

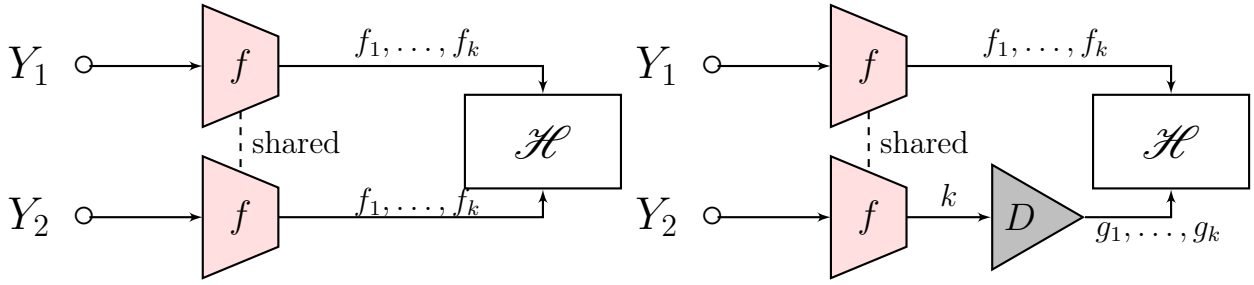


and we have the NN output as follows:

$$\mathbf{F} = \begin{pmatrix} -1.3319606 & -1.2055342 & 1.3556286 \\ -1.3442682 & 1.1980174 & -1.3629394 \\ 1.3481466 & -1.1362327 & -1.4217433 \\ 1.3401704 & 1.1722374 & 1.4313569 \\ 2.2276532e^{-2} & -1.9362740e^{-2} & -1.7726040e^{-2} \\ -7.7401940e^{-3} & -1.3329463e^{-2} & -8.1622172e^{-3} \\ -3.1435837e^{-3} & 1.1085311e^{-3} & 2.1041211e^{-2} \\ -4.0000244e^{-3} & 4.9380683e^{-3} & 9.6413335e^{-3} \end{pmatrix}$$

which is again in the form of $\mathbf{F} = \begin{pmatrix} c_1\phi_1 & c_2\phi_2 & c_3\phi_3 \end{pmatrix}$, where c_1, c_2, and c_3 are some constants other than before. If we analyze the corresponding $\Delta_{(\mathsf{a})} = \|\tilde{B}_{X;Y}\Phi\|_F^2 - \|\tilde{B}_{X;Y}\hat{\Phi}_{(\mathsf{a})}\|_F^2$, we have $\Delta_{(\mathsf{a})} = 0.00602$. More specifically, if we calculate the percentage change, we have that $\frac{\Delta_{(\mathsf{a})}}{\|\tilde{B}_{X;Y}\Phi\|_F^2} = 2.1\%$, i.e., this model helps to recover 97.9% of the total information of interest.

If we use either the second or third structure to obtain the features



only the one corresponding to $\tilde{B}^{\text{sym}}_{X_1;X_2}$ ($\tilde{B}^{\text{pd}}_{X_1;X_2} = \tilde{B}^{\text{sym}}_{X_1;X_2}$ in this example) will be returned:

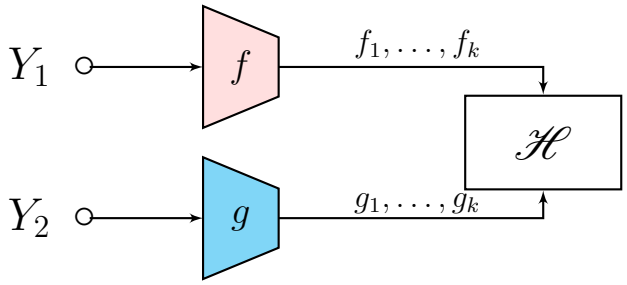
$$\mathbf{F}^{\text{sym}} = \begin{pmatrix} 7.6277673e^{-1} \\ -7.768358e^{-1} \\ 7.794626e^{-1} \\ -7.490486e^{-1} \\ -2.060318e^{-2} \\ -2.821795e^{-2} \\ 4.500071e^{-2} \\ -8.38531e^{-3} \end{pmatrix}, \mathbf{F}^{\text{pd}} = \begin{pmatrix} -3.2981014e^{-1} \\ 3.6776304e^{-1} \\ -3.1002647e^{-1} \\ 3.4087816e^{-1} \\ 1.706952e^{-2} \\ -4.246237e^{-2} \\ -3.434657e^{-2} \\ -1.221869e^{-2} \end{pmatrix}$$

which are both in the form of $\mathbf{F} = \begin{pmatrix} c\phi_2 \end{pmatrix}$ for some constant c. The other 2 modes (ϕ_1 and ϕ_3) will be neglected. If we analyze the corresponding $\Delta_{(\mathsf{s})} = \|\tilde{B}_{X;Y}\Phi\|_F^2 - \|\tilde{B}_{X;Y}\hat{\Phi}_{(\mathsf{sym})}\|_F^2$, we have $\Delta_{(\mathsf{s})} = 0.200$. If we analyze the corresponding $\Delta_{(\mathsf{pd})} = \|\tilde{B}_{X;Y}\Phi\|_F^2 - \|\tilde{B}_{X;Y}\hat{\Phi}_{(\mathsf{pd})}\|_F^2$, we have $\Delta_{(\mathsf{pd})} = 0.201$. More specifically, if we calculate the percentage change, we have that $\frac{\Delta_{(\mathsf{s})}}{\|\tilde{B}_{X;Y}\Phi\|_F^2} = 69.0\%$ and $\frac{\Delta_{(\mathsf{pd})}}{\|\tilde{B}_{X;Y}\Phi\|_F^2} = 69.3\%$, i.e., the second and third model recovers only 31% and 30.7%of the total information of interest. This is actually close to the theoretical value of $\frac{\sigma_2^2}{\sum_{i=1}^3 \sigma_i^2} = \frac{0.3^2}{0.4^2+0.3^2+0.2^2} \approx 31.0\%$.

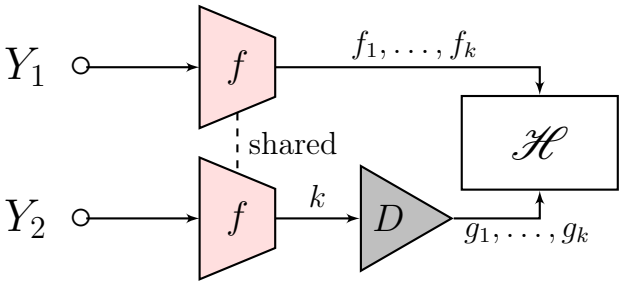
5.4 General Case

From the simulation results, we conclude the application scenario for the three proposed models.

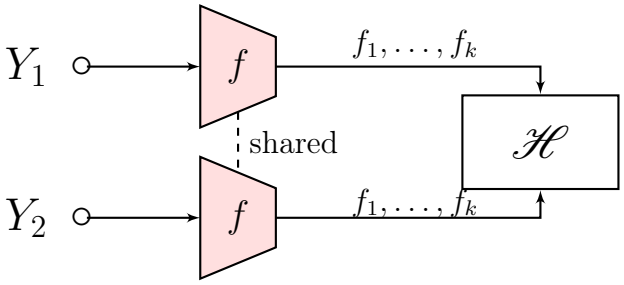
- The following model is able to extract feature functions for a general full rank $\tilde{B}_{X_1;X_2}$. However, the number of parameters in the model is large and thus requires more training samples.



- The following model is able to extract feature functions for full rank $\tilde{B}^{\text{sym}}_{X_1;X_2}$. When there are no modes contained only in the anti-symmetric part, it is better to choose this structure because the number of parameters to learn can be reduced.



- The following model is able to extract feature functions for positive definite part of $\tilde{B}^{\text{sym}}_{X_1;X_2}$. When there is an obvious symmetric structure contained in the problem, it is better to choose this structure because it has the least number of parameters to learn.



Chapter 6

Conclusion and Future Work

In Chapter 2, we introduced the local information geometry to analyze the problem of extracting low-dimensional features from high-dimensional observations. In Chapter 3, We formulated the problem using the HMM, and derived the corresponding optimization target. We also proposed several NN structures to solve this problem. In Chapter 4, we compare them along with the correctness proof and several tail bounds. And finally in Chapter 5, we applied the algorithm on synthesized data and showed the usage of different algorithms for difference cases. This work opens up possibilities for enhancing the feature extraction for Big Data by leveraging information-centric ideas, encouraging further research in this direction.

Figure 6.1: Transition Probability Diagram of X

Future work includes extending the current Two-Tap HMM, illustrated in Figure 3.2, to a multi-tap scenario. At present, all models are analyzed and solved under the assumption of a homogeneous HMM with the memory of only one step, and this thesis provides detailed guidelines on selecting the appropriate model based on specific requirements. However, this assumption does not always hold in real-world scenarios. A key objective is to develop tests that assist system designers in selecting the optimal tap model and corresponding methods without conducting labor-intensive experiments across multiple models. For instance, factors such as the number of available samples, accuracy requirements, and complexity constraints could be used to identify the most suitable model. While preliminary work has been explored in [102], further research is needed to advance this area.

Another potential direction is to clarify some borderline cases. In this thesis, we assume

that the Markov chain of X does not affect the representation map between X and Y, which is a general case in reality. There are, however, some special cases where $P_{X_2|X_1}$ comes into the game. For example, when the transition probability of X is given as in Figure 6.1, the singular modes of $\tilde{B}_{X;Y}$ are killed, and thus there is no hope to extract features from observations of Y. More theoretical analysis and numerical experiments need to be done to fill this gap.

Appendix A

Solution to Problem (3.6) under Special Cases

$$
\begin{aligned}
& \|B - PMP^T\|_F^2 \\
= \quad & \|\frac{B+B^T}{2} + \frac{B-B^T}{2} - P(M^s + M^a)P^T\|_F^2 \\
= \quad & \|\frac{B+B^T}{2} - PM^sP^T\|_F^2 + \|\frac{B-B^T}{2} - PM^aP^T\|_F^2, \qquad \text{(A.1)}
\end{aligned}
$$

where $M^s \triangleq \frac{M+M^T}{2}$ is the symmetric part of matrix M, and $M^a \triangleq \frac{M-M^T}{2}$ is the anti-symmetric part of matrix M. It can be shown that Problem (3.19) is equivalent to the first term of (A.1) but ignores the second term. Therefore, if we transform a general B into $\frac{B+B^T}{2}$ and then solve the closest mapping problem, we will lose the anti-symmetric component of B.

A.0.1 Special Case: When $k = 1$

In this case, problem (3.6) reduces to the following problem

$$
\min_{a\in\mathbb{R},\phi\in\mathbb{R}^{N\times 1}} \quad \|B - a\phi\phi^T\|_F^2. \qquad \text{(A.2)}
$$

Notice that this is actually in the form of Problem (opt:5), and thus the solution is to set a as the eigenvalue of $\frac{B+B^T}{2}$ that has the largest absolute value and ϕ as the corresponding eigenvector.

A.0.2 Special Case: When B is a symmetric matrix with rank$(B) \geq k$

In this case, we do not have the second term in Eq. (A.1). The problem can be reduced to find the rank-k approximation to a symmetric matrix. Therefore, the solution in this case due to Theorem 3.2.1 is the first k singular vectors of $\frac{B+B^T}{2}$.

A.0.3 Special Case: When B is an anti-symmetric matrix with rank$(B) \geq k$

In this case, we do not have the first term in (A.1). Then, the problem becomes to find a low-rank anti-symmetric matrix approximation to B with the constraint that the approximated matrix has a rank less than or equal to k. The properties of an anti-symmetric matrix are:

Let A be an anti-symmetric matrix, and denote its singular values by $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_n \geq 0$, then

- If n is even, then $\sigma_{2k} = \sigma_{2k-1} \geq 0, k = 1, 2, \ldots, n/2$. If n is odd, then the same relationship holds up to $k = (n-1)/2$ and also $\sigma_n = 0$.
- The eigenvalues $\lambda_j = (-1)^j i\sigma_j, j = 1, 2, \ldots, n$.

That is to say, the singular values of an anti-symmetric matrix come in pairs and thus the rank must be an even number. Moreover, every $2n \times 2n$ real anti-symmetric matrix can be written in the form $B = Q\Sigma Q^T$ where Q is orthogonal and

$$\Sigma = \begin{pmatrix} 0 & \sigma_1 & & & & & & & \\ -\sigma_1 & 0 & & & & & & & \\ & & 0 & \sigma_2 & & & & & \\ & & -\sigma_2 & 0 & & & & & \\ & & & \cdots & \cdots & & & & \\ & & & & & 0 & \sigma_r & & \\ & & & & & -\sigma_r & 0 & & \\ & & & & & & & 0 & \\ & & & & & & & \cdots & \cdots \\ & & & & & & & & 0 \end{pmatrix}$$

In the odd-dimensional case, Σ always has at least one row and column of zeros. Note that $Q\Sigma Q^T$ given here is not the SVD of B. If we want the SVD form, we need to decompose Σ

into

$$\Sigma = \begin{pmatrix} U & & & & \\ & U & & & \\ & & \cdots & & \\ & & & U & \\ & & & & U \end{pmatrix} \times \begin{pmatrix} \sigma_1 & & & & & & & & \\ & \sigma_1 & & & & & & & \\ & & \sigma_2 & & & & & & \\ & & & \sigma_2 & & & & & \\ & & & & \cdots & & & & \\ & & & & & \sigma_r & & & \\ & & & & & & \sigma_r & & \\ & & & & & & & 0 & \\ & & & & & & & & \cdots \\ & & & & & & & & & 0 \end{pmatrix} \times$$

$$\begin{pmatrix} V^T & & & & \\ & V^T & & & \\ & & \cdots & & \\ & & & V^T & \\ & & & & V^T \end{pmatrix}$$

where $U = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}$, and $V = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}$ are both orthogonal matrices.

If we ignore the anti-symmetric constraint at first, instead considering a low-rank (a rank smaller than or equal to $2\lfloor \frac{k}{2} \rfloor$ since when we move back to anti-symmetric matrix space the rank must be an even number) approximation of the anti-symmetric B, it can be shown that the optimal low-rank matrix satisfies the anti-symmetric constraint automatically. Therefore, the solution in this case due to Theorem 3.2.1 is the first $2\lfloor \frac{k}{2} \rfloor$ columns of $Q \times$ the first $\lfloor \frac{k}{2} \rfloor$ diagonal 2×2 block of $\Sigma \times$ the first $2\lfloor \frac{k}{2} \rfloor$ rows of Q^T.

A.0.4 Back to the homogeneous HMM

In our problem, with a limited number of samples, the empirical B matrix $\check{B}_{Y_1;Y_2} = \tilde{B}^T_{X;Y}\tilde{B}_{X_1;X_2}\tilde{B}_{X;Y} + \epsilon$, where ϵ represents the noise introduced by samples. We want a rank-k approximation in the form of PMP^T where $k \geq |\mathcal{X}| - 1 \triangleq \text{rank}(\tilde{B}^T_{X;Y}\tilde{B}_{X_1;X_2}\tilde{B}_{X;Y})$. Again, we decompose this problem,

$$\begin{aligned} & \underset{P \in \mathbb{R}^{|\mathcal{Y}| \times k}, M \in \mathbb{R}^{k \times k}}{\arg\min} \quad \|\check{B}_{Y_1;Y_2} - PMP^T\|_F^2 \\ = & \underset{P \in \mathbb{R}^{|\mathcal{Y}| \times k}, M \in \mathbb{R}^{k \times k}}{\arg\min} \quad \|\tilde{B}^T_{X;Y}\tilde{B}^s_{X_1;X_2}\tilde{B}_{X;Y} + \epsilon^s + \tilde{B}^T_{XY}\tilde{B}^a_{X_1;X_2}\tilde{B}_{X;Y} + \epsilon^a - P(M^s + M^a)P^T\|_F^2 \\ = & \underset{P \in \mathbb{R}^{|\mathcal{Y}| \times k}, M \in \mathbb{R}^{k \times k}}{\arg\min} \quad \|\tilde{B}^T_{X;Y}\tilde{B}^s_{X_1;X_2}\tilde{B}_{X;Y} + \epsilon^s - PM^sP^T\|_F^2 + \|\tilde{B}^T_{X;Y}\tilde{B}^a_{X_1;X_2}\tilde{B}_{X;Y} + \epsilon^a - PM^aP^T\|_F^2. \end{aligned} \tag{A.3}$$

If we consider the two optimization problems separately, we know the optimal solution to $\arg\min_{P\in\mathbb{R}^{|\mathcal{Y}|\times k}} \|\tilde{B}^T_{X;Y}\tilde{B}^s_{X_1;X_2}\tilde{B}_{X;Y}+\epsilon^s-PM^sP^T\|^2_F$ is the truncated singular vector matrix of $\tilde{B}^T_{X;Y}\tilde{B}^s_{X_1;X_2}\tilde{B}_{X;Y}+\epsilon^s$, and the optimal solution to $\arg\min_{P\in\mathbb{R}^{|\mathcal{Y}|\times k}} \|\tilde{B}^T_{X;Y}\tilde{B}^a_{X_1;X_2}\tilde{B}_{X;Y}+\epsilon^a-PM^aP^T\|^2_F$ is the truncated singular vector matrix of $\tilde{B}^T_{X;Y}\tilde{B}^a_{X_1;X_2}\tilde{B}_{X;Y}+\epsilon^a$, which according to matrix perturbation theory [101] should be close to the $\tilde{B}_{X;Y}$ space. However, this is no general solution to the problem (A.3) if $\check{B}_{Y_1;Y_2}$ is not in one of the special forms mentioned above.

Appendix B

Proof of $\Delta_{(\cdot)} \geq 0$

The proof is completed by using the following lemma [103].

Lemma B.0.1. Given an arbitrary $k_1 \times k_2$ matrix A and $k \in \{1, \ldots, \min(k_1, k_2)\}$, we have

$$\max_{M \in \mathbb{R}^{k_2 \times k}: M^T M = I} \|AM\|_F^2 = \sum_{i=1}^{k} \sigma_i^2(A),$$

where $\sigma_1(A) \geq \cdots \geq \sigma_{\min(k_1,k_2)}(A)$ denotes the ordered singular values of A. Moreover, the maximum is achieved by

$$M = \left[\psi_1(A), \ldots, \psi_k(A)\right],$$

with $\psi_i(A)$ denoting the right singular vector of A corresponding to $\sigma_i(A)$, for $i = 1, \ldots, \min(k_1, k_2)$.

References

[1] L. Rabiner and B. Juang, "An introduction to hidden markov models," *IEEE ASSP Magazine*, vol. 3, no. 1, pp. 4–16, 1986.

[2] B. Mor, S. Garhwal, and A. Kumar, "A systematic review of hidden markov models and their applications," *Archives of computational methods in engineering*, vol. 28, pp. 1429–1448, 2021.

[3] L. R. Rabiner, "A tutorial on hidden markov models and selected applications in speech recognition," *Proceedings of the IEEE*, vol. 77, no. 2, pp. 257–286, 1989.

[4] J. Picone, "Continuous speech recognition using hidden markov models," *IEEE Assp magazine*, vol. 7, no. 3, pp. 26–41, 1990.

[5] B. H. Juang and L. R. Rabiner, "Hidden markov models for speech recognition," *Technometrics*, vol. 33, no. 3, pp. 251–272, 1991.

[6] M. Gales, S. Young, *et al.*, "The application of hidden markov models in speech recognition," *Foundations and Trends® in Signal Processing*, vol. 1, no. 3, pp. 195–304, 2008.

[7] A. Elakkiya, K. J. Surya, K. Venkatesh, and S. Aakash, "Implementation of speech to text conversion using hidden markov model," in *2022 6th International Conference on Electronics, Communication and Aerospace Technology*, IEEE, 2022, pp. 359–363.

[8] J. S. Boreczky and L. D. Wilcox, "A hidden markov model framework for video segmentation using audio and image features," in *Proceedings of the 1998 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP'98 (Cat. No. 98CH36181)*, IEEE, vol. 6, 1998, pp. 3741–3744.

[9] P. Chang, M. Han, and Y. Gong, "Extract highlights from baseball game video with hidden markov models," in *Proceedings. International Conference on Image Processing*, IEEE, vol. 1, 2002, pp. I–I.

[10] L. Xie, S.-F. Chang, A. Divakaran, and H. Sun, "Structure analysis of soccer video with hidden markov models," in *2002 IEEE International Conference on Acoustics, Speech, and Signal Processing*, IEEE, vol. 4, 2002, pp. IV–4096.

[11] X. Liu and T. Cheng, "Video-based face recognition using adaptive hidden markov models," in *2003 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2003. Proceedings.*, IEEE, vol. 1, 2003, pp. I–I.

[12] B. U. Toreyin, Y. Dedeoglu, and A. E. Cetin, "Flame detection in video using hidden markov models," in *IEEE International Conference on Image Processing 2005*, IEEE, vol. 2, 2005, pp. II–1230.

[13] Q. Zhang and B. Li, "Relative hidden markov models for video-based evaluation of motion skills in surgical training," *IEEE transactions on pattern analysis and machine intelligence*, vol. 37, no. 6, pp. 1206–1218, 2014.

[14] D. K. D. Haussler and M. Eeckman, "A generalized hidden markov model for the recognition of human genes in dna," in *Proc. int. conf. on intelligent systems for molecular biology, st. louis*, 1996, pp. 134–142.

[15] J. Henderson, S. Salzberg, and K. H. Fasman, "Finding genes in dna with a hidden markov model," *Journal of Computational Biology*, vol. 4, no. 2, pp. 127–141, 1997.

[16] R. Durbin, S. R. Eddy, A. Krogh, and G. Mitchison, *Biological sequence analysis: probabilistic models of proteins and nucleic acids*. Cambridge university press, 1998.

[17] R. J. Boys, D. A. Henderson, and D. J. Wilkinson, "Detecting homogeneous segments in dna sequences by using hidden markov models," *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, vol. 49, no. 2, pp. 269–285, 2000.

[18] T. Koski, *Hidden Markov models for bioinformatics*. Springer Science & Business Media, 2001, vol. 2.

[19] K.-C. Liang, X. Wang, and D. Anastassiou, "Bayesian basecalling for DNA sequence analysis using hidden markov models," *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, vol. 4, no. 3, pp. 430–440, 2007.

[20] Y. Liu, S. C. Reed, C. Lo, A. D. Choudhury, H. A. Parsons, D. G. Stover, G. Ha, G. Gydush, J. Rhoades, D. Rotem, *et al.*, "Finaleme: Predicting DNA methylation by the fragmentation patterns of plasma cell-free DNA," *Nature Communications*, vol. 15, no. 1, p. 2790, 2024.

[21] H. Ringbauer, Y. Huang, A. Akbari, S. Mallick, I. Olalde, N. Patterson, and D. Reich, "Accurate detection of identity-by-descent segments in human ancient dna," *Nature Genetics*, vol. 56, no. 1, pp. 143–151, 2024.

[22] M. R. Hassan and B. Nath, "Stock market forecasting using hidden markov model: A new approach," in *5th international conference on intelligent systems design and applications (ISDA'05)*, IEEE, 2005, pp. 192–196.

[23] J. G. Dias, J. K. Vermunt, and S. Ramos, "Clustering financial time series: New insights from an extended hidden markov model," *European Journal of Operational Research*, vol. 243, no. 3, pp. 852–864, 2015.

[24] R. S. Mamon and R. J. Elliott, *Hidden Markov models in finance*. Springer, 2007, vol. 4.

[25] Y. Zhang, "Prediction of financial time series with hidden markov models," 2004.

[26] D. Roman, G. Mitra, and N. Spagnolo, "Hidden markov models for financial optimization problems," *IMA Journal of Management Mathematics*, vol. 21, no. 2, pp. 111–129, 2010.

[27] P. Nystrup, H. Madsen, and E. Lindström, "Long memory of financial time series and hidden markov models with time-varying parameters," *Journal of Forecasting*, vol. 36, no. 8, pp. 989–1002, 2017.

[28] N. Nguyen, "Hidden markov model for stock trading," *International Journal of Financial Studies*, vol. 6, no. 2, p. 36, 2018.

[29] M. Zhang, X. Jiang, Z. Fang, Y. Zeng, and K. Xu, "High-order hidden markov model for trend prediction in financial time series," *Physica A: Statistical Mechanics and its Applications*, vol. 517, pp. 1–12, 2019.

[30] H. Soltani and M. B. Abbes, "The predictive power of financial stress on the financial markets dynamics: Hidden markov model," *Journal of Economics and Finance*, vol. 47, no. 1, pp. 94–115, 2023.

[31] I. Guyon and A. Elisseeff, "An introduction to feature extraction," in *Feature extraction: foundations and applications*, Springer, 2006, pp. 1–25.

[32] I. Guyon, S. Gunn, M. Nikravesh, and L. A. Zadeh, *Feature extraction: foundations and applications*. Springer, 2008, vol. 207.

[33] Y. Bengio, A. Courville, and P. Vincent, "Representation learning: A review and new perspectives," *IEEE transactions on pattern analysis and machine intelligence*, vol. 35, no. 8, pp. 1798–1828, 2013.

[34] W. Wang, R. Arora, K. Livescu, and J. Bilmes, "On deep multi-view representation learning," in *International conference on machine learning*, PMLR, 2015, pp. 1083–1092.

[35] T. Jebara, Y. Song, and K. Thadani, "Spectral clustering and embedding with hidden markov models," in *Machine Learning: ECML 2007: 18th European Conference on Machine Learning, Warsaw, Poland, September 17-21, 2007. Proceedings 18*, Springer, 2007, pp. 164–175.

[36] S. Adams, P. A. Beling, and R. Cogill, "Feature selection for hidden markov models and hidden semi-markov models," *IEEE Access*, vol. 4, pp. 1642–1657, 2016.

[37] S. Adams and P. A. Beling, "A survey of feature selection methods for gaussian mixture models and hidden markov models," *Artificial Intelligence Review*, vol. 52, pp. 1739–1779, 2019.

[38] C. M. Bishop, "Neural networks and their applications," *Review of scientific instruments*, vol. 65, no. 6, pp. 1803–1832, 1994.

[39] K. Gurney, *An introduction to neural networks*. CRC press, 2018.

[40] B. Müller, J. Reinhardt, and M. T. Strickland, *Neural networks: an introduction*. Springer Science & Business Media, 2012.

[41] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998.

[42] D. Maulud and A. M. Abdulazeez, "A review on linear regression comprehensive in machine learning," *Journal of Applied Science and Technology Trends*, vol. 1, no. 2, pp. 140–147, 2020.

[43] H. Abdi and L. J. Williams, "Principal component analysis," *Wiley interdisciplinary reviews: computational statistics*, vol. 2, no. 4, pp. 433–459, 2010.

[44] I. T. Jolliffe and J. Cadima, "Principal component analysis: A review and recent developments," *Phil. Trans. R. Soc. A.*, 2016.

[45] J. Ye, R. Janardan, and Q. Li, "GPCA: An efficient dimension reduction scheme for image compression and retrieval," ser. KDD '04, New York, NY, USA: Association for Computing Machinery, 2004, pp. 354–363. URL: https://doi.org/10.1145/1014052.1014092.

[46] J. Yang, D. Zhang, A. F. Frangi, and J.-Y. Yang, "Two-dimensional PCA: A new approach to appearance-based face representation and recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 26, no. 1, pp. 131–137, 2004.

[47] J. Ye, "Generalized low rank approximations of matrices," *Mach Learn*, pp. 167–191, 2005.

[48] X. He, D. Cai, and P. Niyogi, "Tensor subspace analysis," Jan. 2005.

[49] L. D. Lathauwer, B. D. Moor, and J. Vandewalle, "A multilinear singular value decomposition," *SIAM J. Matrix Anal. Appl.*, vol. 21, no. 4, pp. 1253–1278, 2000.

[50] L. D. Lathauwer and J. Vandewalle, "Dimensionality reduction in higher-order signal processing and rank-(R1,R2,. . . ,RN) reduction in multilinear algebra," *Linear Algebra Appl.*, vol. 391, pp. 31–55, Nov. 2004.

[51] H. Lu, K. N. Plataniotis, and A. N. Venetsanopoulos, "MPCA: Multilinear principal component analysis of tensor objects," *IEEE Trans. Neural Netw.*, vol. 19, no. 1, pp. 18–39, 2008.

[52] J. Zhai, S. Zhang, J. Chen, and Q. He, "Autoencoder and its various variants," in *2018 IEEE international conference on systems, man, and cybernetics (SMC)*, IEEE, 2018, pp. 415–419.

[53] A. Géron, *Hands-on machine learning with Scikit-Learn, Keras, and TensorFlow*. O'Reilly Media, Inc., 2022.

[54] C.-Y. Liou, J.-C. Huang, and W.-C. Yang, "Modeling word perception using the elman network," *Neurocomputing*, vol. 71, no. 16-18, pp. 3150–3157, 2008.

[55] C.-Y. Liou, W.-C. Cheng, J.-W. Liou, and D.-R. Liou, "Autoencoder for words," *Neurocomputing*, vol. 139, pp. 84–96, 2014.

[56] L. Breiman and J. H. Friedman, "Estimating optimal transformations for multiple regression and correlation," *Journal of the American statistical Association*, vol. 80, no. 391, pp. 580–598, 1985.

[57] A. Makur, F. Kozynski, S.-L. Huang, and L. Zheng, "An efficient algorithm for information decomposition and extraction," in *2015 53rd Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, IEEE, 2015, pp. 972–979.

[58] S.-L. Huang, A. Makur, L. Zheng, and G. W. Wornell, "An information-theoretic approach to universal feature selection in high-dimensional inference," in *2017 IEEE International Symposium on Information Theory (ISIT)*, 2017, pp. 1336–1340.

[59] H. O. Hirschfeld, "A connection between correlation and contingency," *Math. Proc. Cambridge Philos. Soc.*, vol. 31, no. 4, pp. 520–524, 1935.

[60] H. Gebelein, "Das statistische problem der korrelation als variations- und eigenwertproblem und sein zusammenhang mit der ausgleichsrechnung," *ZAMM - Journal of Applied Mathematics and Mechanics / Zeitschrift für Angewandte Mathematik und Mechanik*, vol. 21, no. 6, pp. 364–379, 1941.

[61] A. Rényi, "On measures of dependence," *Acta Mathematica Academiae Scientiarum Hungaricae*, pp. 441–451, 1959.

[62] L. Breiman and J. H. Friedman, "Estimating optimal transformations for multiple regression and correlation," *J. Amer. Statist. Assoc.*, vol. 80, no. 391, pp. 580–598, 1985.

[63] S. P. Borade, "When all information is not created equal," Ph.D. Thesis, Massachusetts Institute of Technology, 2008. URL: http://hdl.handle.net/1721.1/45883.

[64] E. A. Abbe, "Local to global geometric methods in information theory," Ph.D. Thesis, Massachusetts Institute of Technology, 2008. URL: http://hdl.handle.net/1721.1/44404.

[65] S.-L. Huang, "Euclidean network information theory," Ph.D. Thesis, Massachusetts Institute of Technology, 2013. URL: http://hdl.handle.net/1721.1/84888.

[66] A. Makur, "A study of local approximations in information theory," S.M. Thesis, Massachusetts Institute of Technology, 2015. URL: http://hdl.handle.net/1721.1/99789.

[67] A. Makur, "Information contraction and decomposition," Sc. D. Thesis, Massachusetts Institute of Technology, 2019. URL: https://hdl.handle.net/1721.1/122692.

[68] S. Borade and L. Zheng, "Euclidean information theory," in *2008 IEEE International Zurich Seminar on Communications*, 2008, pp. 14–17. DOI: 10.1109/IZS.2008.4497265.

[69] E. Abbe and L. Zheng, "A coordinate system for gaussian networks," *IEEE Trans. Inf. Theory*, vol. 58, no. 2, pp. 721–733, 2012. DOI: 10.1109/TIT.2011.2169536. URL: https://doi.org/10.1109/TIT.2011.2169536.

[70] S.-L. Huang, C. Suh, and L. Zheng, "Euclidean information theory of networks," *IEEE Trans. Inf. Theory*, vol. 61, no. 12, pp. 6795–6814, 2015. DOI: 10.1109/TIT.2015.2484066. URL: https://doi.org/10.1109/TIT.2015.2484066.

[71] A. Makur, F. Kozynski, S.-L. Huang, and L. Zheng, "An efficient algorithm for information decomposition and extraction," in *53rd Annual Allerton Conference on Communication, Control, and Computing, Allerton 2015, Allerton Park & Retreat Center, Monticello, IL, USA, September 29 - October 2, 2015*, IEEE, 2015, pp. 972–979. DOI: 10.1109/ALLERTON.2015.7447113. URL: https://doi.org/10.1109/ALLERTON.2015.7447113.

[72] D. Qiu, A. Makur, and L. Zheng, "Probabilistic clustering using maximal matrix norm couplings," in *56th Annual Allerton Conference on Communication, Control, and Computing, Allerton 2018, Monticello, IL, USA, October 2-5, 2018*, IEEE, 2018, pp. 1020–1027. DOI: 10.1109/ALLERTON.2018.8635939. URL: https://doi.org/10.1109/ALLERTON.2018.8635939.

[73] S.-L. Huang, X. Xu, L. Zheng, and G. W. Wornell, "An information theoretic interpretation to deep neural networks," in *IEEE International Symposium on Information Theory, ISIT 2019, Paris, France, July 7-12, 2019*, IEEE, 2019, pp. 1984–1988. DOI: 10.1109/ISIT.2019.8849720. URL: https://doi.org/10.1109/ISIT.2019.8849720.

[74] L. Wang, J. Wu, S.-L. Huang, L. Zheng, X. Xu, L. Zhang, and J. Huang, "An efficient approach to informative feature extraction from multimodal data," in *The Thirty-Third AAAI Conference on Artificial Intelligence, AAAI 2019*, pp. 5281–5288. DOI: 10.1609/AAAI.V33I01.33015281. URL: https://doi.org/10.1609/aaai.v33i01.33015281.

[75] X. Xu and L. Zheng, "Neural feature learning in function space," *Journal of Machine Learning Research*, vol. 25, no. 142, pp. 1–76, 2024. URL: http://jmlr.org/papers/v25/23-1202.html.

[76] J. Z. H. Chen, C. Wei, A. Gaidon, and T. Ma, "Provable guarantees for self-supervised deep learning with spectral contrastive loss," in *Proceedings of the 35th International Conference on Neural Information Processing Systems*, ser. NIPS '21, Red Hook, NY, USA: Curran Associates Inc., 2024, ISBN: 9781713845393.

[77] N. Young, *An introduction to Hilbert space*. Cambridge university press, 1988.

[78] J. Weidmann, *Linear operators in Hilbert spaces*. Springer Science & Business Media, 2012, vol. 68.

[79] S.-L. Huang, A. Makur, G. W. Wornell, and L. Zheng, "Universal features for high-dimensional learning and inference," *Foundations and Trends in Communications and Information Theory*, vol. 21, no. 1-2, pp. 1–299, 2024, ISSN: 1567-2190. DOI: 10.1561/0100000107. URL: http://dx.doi.org/10.1561/0100000107.

[80] J. W. Demmel, *Applied numerical linear algebra*. SIAM, 1997.

[81] K. Fukushima, "Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position," *Biological cybernetics*, vol. 36, no. 4, pp. 193–202, 1980.

[82] W. Zhang, J. Tanida, K. Itoh, and Y. Ichioka, "Shift-invariant pattern recognition neural network and its optical architecture," in *Proceedings of annual conference of the Japan Society of Applied Physics*, Montreal, CA, vol. 564, 1988.

[83] A. Waibel, T. Hanazawa, and G. Hinton, "Phoneme recognition using time-delay neural networks," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 31, no. 3, 1989.

[84] Y. LeCun, B. Boser, J. S. Denker, D. Henderson, R. E. Howard, W. Hubbard, and L. D. Jackel, "Backpropagation applied to handwritten zip code recognition," *Neural computation*, vol. 1, no. 4, pp. 541–551, 1989.

[85] J. J. Weng, N. Ahuja, and T. S. Huang, "Learning recognition and segmentation of 3-d objects from 2-d images," in *1993 (4th) International Conference on Computer Vision*, IEEE, 1993, pp. 121–128.

[86] W. Rawat and Z. Wang, "Deep convolutional neural networks for image classification: A comprehensive review," *Neural computation*, vol. 29, no. 9, pp. 2352–2449, 2017.

[87] S. Hochreiter, "Long short-term memory," *Neural Computation MIT-Press*, 1997.

[88] F. A. Gers and J. Schmidhuber, "Recurrent nets that time and count," in *Proceedings of the IEEE-INNS-ENNS International Joint Conference on Neural Networks. IJCNN 2000. Neural Computing: New Challenges and Perspectives for the New Millennium*, IEEE, vol. 3, 2000, pp. 189–194.

[89] F. A. Gers, J. Schmidhuber, and F. Cummins, "Learning to forget: Continual prediction with lstm," *Neural computation*, vol. 12, no. 10, pp. 2451–2471, 2000.

[90] F. Gers, "Long short-term memory in recurrent neural networks," Ph.D. dissertation, Verlag nicht ermittelbar, 2001.

[91] F. A. Gers, N. N. Schraudolph, and J. Schmidhuber, "Learning precise timing with lstm recurrent networks," *Journal of machine learning research*, vol. 3, no. Aug, pp. 115–143, 2002.

[92] K. Cho, B. Van Merriënboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio, "Learning phrase representations using rnn encoder-decoder for statistical machine translation," *arXiv preprint arXiv:1406.1078*, 2014.

[93] C. Eckart and G. Young, "The approximation of one matrix by another of lower rank," *Psychometrika*, vol. 1, no. 3, pp. 211–218, 1936.

[94] T. T. Cai and A. Zhang, "Rate-optimal perturbation bounds for singular subspaces with applications to high-dimensional statistics," *Ann. Statist.*, vol. 46, no. 1, pp. 60–89, 2018.

[95] P. Wedin, "Perturbation bounds in connection with singular value decomposition," *BIT*, pp. 99–111, 1972.

[96] G. W. Stewart, *Matrix perturbation theory*, 1990.

[97] H. Weyl, "Das asymptotische verteilungsgesetz der eigenwerte linearer partieller differentialgleichungen (mit einer anwendung auf die theorie der hohlraumstrahlung)," *Mathematische Annalen*, vol. 71, no. 4, pp. 441–479, 1912.

[98] T. Terry, *254A, Notes 3a: Eigenvalues and sums of hermitian matrices*, Retrieved 25 May 2015. URL: https://terrytao.wordpress.com/2010/01/12/254a-notes-3a-eigenvalues-and-sums-of-hermitian-matrices.

[99] J. A. Tropp, "User-friendly tail bounds for sums of random matrices," *Foundations of computational mathematics*, vol. 12, pp. 389–434, 2012.

[100] E. J. Candes and Y. Plan, "A probabilistic and ripless theory of compressed sensing," *IEEE transactions on information theory*, vol. 57, no. 11, pp. 7235–7254, 2011.

[101] Y. Yu, T. Wang, and R. J. Samworth, "A useful variant of the Davis—Kahan theorem for statisticians," *Biometrika*, vol. 102, no. 2, pp. 315–323, 2015. (visited on 03/15/2023).

[102] J. Jin, "An information-centric algorithm for feature extraction in high-dimensional data," Ph.D. dissertation, Massachusetts Institute of Technology, 2021.

[103] R. A. Horn and C. R. Johnson, *Matrix analysis*. Cambridge university press, 2012.