# Statistical Methods with Applications in Epigenomics, Metagenomics and Neuroimaging

by

Xin Xing

(Under the Direction of Wenxuan Zhong)

## Abstract

As the rapid development of biotechnology, more complex data sets are now generated to address extremely complex biological problems. It is challenging to develop new statistical methods to analyze such data. In this thesis, I propose a nonparametric hypothesis test and two statistical learning methods to solve biological problems arising from epigenomics, metagenomics, and neuroimaging. First, the proposed test aims at testing the significance of the interaction in bivariate smoothing spline ANOVA model. The derived asymptotic distribution of the test statistic unveils a new version of Wilks phenomenon, and the power is minimax optimal in the sense of Ingster. The performance of the proposed test was demonstrated on discovering differentially methylated regions in a genome-wide DNA methylation study. Second, I propose a statistical learning method that simultaneously identifies microbial species and estimates their abundances without using reference genomes. I show that the proposed method achieves high accuracy in both simulated data and real metagenomic data related to inflammatory bowel disease (IBD), type-2 diabetes (T2D) and obesity. Third, I develop a model-based dictionary learning (MDL) method which provides an effective and flexible framework for different types of data: continuous, discrete and categorical. It also

provides a general framework to model data with spatial or temporal correlation. The performance of the MDL method was demonstrated in studying the brain connectivity and learning the cell-type specific expression profile through spatial transcriptomic imaging.

INDEX WORDS: Nonparametric inference, Smoothing spline ANOVA, Minimax, Metagenomics, EM algorithm, Dictionary learning

# Statistical Methods with Applications in Epigenomics, Metagenomics and Neuroimaging

by

Xin Xing

B.A., University of Science and Technology of China, China, 2010

M.S., University of Science and Technology of China, China, 2013

A Dissertation Submitted to the Graduate Faculty

of The University of Georgia in Partial Fulfillment

of the

Requirements for the Degree

Doctor of Philosophy

Athens, Georgia

2018

# Statistical Methods with Applications in Epigenomics, Metagenomics and Neuroimaging

by

Xin Xing

Major Professor: Wenxuan Zhong

Committee: Jun Liu
Tianming Liu
Ping Ma
Yin Xu

Electronic Version Approved:

Suzanne Barbour
Dean of the Graduate School
The University of Georgia
May 2018

## Acknowledgments

Foremost, I would like to express my sincere gratitude to my advisor Prof. Wenxuan Zhong for the continuous support of my Ph.D study and research, for her patience, motivation, enthusiasm, and immense knowledge. Her guidance helped me in all the time of researchand writing of this thesis. I could not have imagined having a better advisor and mentor for my Ph.D study.

Besides my advisor, I would like to thank the rest of my thesis committee: Prof. Jun Liu, Prof. Tianming Liu, Prof. Ping Ma, and Prof. Yin Xu, for their encouragement, insightful comments, and hard questions.

I would like to thank my wife for immense support and encourage for my reserch. I especially thankful to my wife and my son for their love throughout my life.

This work was supported in part by NSF DMS-1440037, DMS-1438957, DMS-1440038 and NIH 1R01GM113242-01, 1R01GM122080-01

# Table of Contents

# List of Tables

## List of Figures

# Chapter 1

# The Minimax Hypothesis Testing in Smoothing Spline ANOVA Models: Two-Way Case

## 1.1 Introduction

We consider the following nonparametric regression model,

$$Y_i = f(\mathbf{x}_i) + \epsilon_i, \qquad i = 1, \ldots, n \tag{1.1}$$

where $Y_i$ is the $i$th observation of the response variable, $\mathbf{x}_i = (x_i^{\langle 1 \rangle}, \ldots, x_i^{\langle d \rangle})$ is the $i$th observation of $d$ predictor variables, $f$ is the nonparametric function to be estimated, and $\epsilon_i$s are independently and identically distributed (IID) random errors following a normal distribution with mean zero and variance $\sigma^2$. A popular method for multivariate predictors model-building is the smoothing spline analysis of variance (SS-ANOVA) model [1; 2]. In the SS-ANOVA, we write

$$f(\mathbf{x}) = f_0 + \sum_{j=1}^{d} f_j(x^{\langle j \rangle}) + \sum_{j,k} f_{j,k}(x^{\langle j \rangle}, x^{\langle k \rangle}) + \ldots, \tag{1.2}$$

where $f_0$ is the grand mean, $f_j$s are main effects, $f_{j,k}$s are two-way interactions, and so on. The identifiability of the terms in (1.2) is assured by side conditions through averaging operators [1]. Insignificant higher-order interactions are often excluded to enhance the model interpretability and predictive power; exclusion of all interactions yields the additive models [3].

In practice, many problems, e.g., whether two curves are significantly different up to a constant can be formulated by testing whether an interaction component in a SS-ANOVA

model is zero. We shall present two examples which motivate our theory and methodology development.

**Example** *DNA methylation in case-control study.* DNA methylation is an essential epigenetic mechanism that regulates gene expression. Aberrant DNA methylation can contribute to a number of human diseases including cancer [4]. In a typical case-control study of DNA methylation [5], the DNA samples are extracted from the cells of patients with a disease of interest (case group) and those of normal subjects (control group). The DNA methylation levels are then measured. Of the primary interest is to identify the genome regions that have significantly different methylation levels, i.e., differentially methylated regions (DMRs), between case and control groups. In particular, the data consists of $(M_{ij}, s_i, g_j)$, where $M_{ij}$ is the methylation level at the $i$th genome location of the $j$th subject, $s_i$ denotes the $i$th genome location, and $g_j$ equals to 1 if the $j$th subject is in the case group, equals to 0 if the $j$th subject is in the control group. We assume $M_{ij} = f(s_i, g_j) + \epsilon_{ij}$, where $\epsilon_{ij}$ is the random error. With the SS-ANOVA decomposition,

$$f(s, g) = f_0 + f_1(s) + f_2(g) + f_{1,2}(s, g), \tag{1.3}$$

we aim to test

$$H_0 : f_{1,2}(s, g) = 0 \text{ v.s. } H_1 : f_{1,2}(s, g) \neq 0, \tag{1.4}$$

which is equivalent to testing whether the methylation levels in two groups have same profiles along the genome. □

Motivated by the real example, we consider a special case of general nonparamtric model (1.1). That is the following nonparametric model with two predictors,

$$Y_{ij} = f(x_i^{\langle 1 \rangle}, x_j^{\langle 2 \rangle}) + \epsilon_{ij}, \qquad i = 1, \ldots, n, \quad j = 1, \ldots, a \tag{1.5}$$

where $Y_{ij}$ is the measurement of at the $i$th coordinate of $j$th subject, $x_i^{\langle 1 \rangle} \in \mathcal{X}_1$ is the $i$th coordinate (in time or space) of the measurement, $x_j^{\langle 2 \rangle} \in \mathcal{X}_2$ is the group indicator for the

$j$th subject.The SS-ANOVA model (1.2) thus reduces to

$$f(x^{\langle 1\rangle}, x^{\langle 2\rangle}) = f_0 + f_1(x^{\langle 1\rangle}) + f_2(x^{\langle 2\rangle}) + f_{1,2}(x^{\langle 1\rangle}, x^{\langle 2\rangle}) \tag{1.6}$$

where $f_0$ is the grand mean, $f_1$ is the temporal or spatial effect, $f_2$ is the group effect, and $f_{1,2}$ is the interaction. As evident in the examples, we are interested in testing the significance of the interaction between $x^{\langle 1\rangle}$ and $x^{\langle 2\rangle}$ in model (1.6). That is, we aim to test

$$H_0 : f_{1,2}(x^{\langle 1\rangle}, x^{\langle 2\rangle}) = 0 \text{ v.s. } H_1 : f_{1,2}(x^{\langle 1\rangle}, x^{\langle 2\rangle}) \neq 0. \tag{1.7}$$

However, unlike classical linear ANOVA models having well developed hypothesis testing procedures, the hypothesis testing in SS-ANOVA is still lacking. The major difficulty is that unknown function $f$ resides in an infinite dimensional space. Consequently, some subtle but significant differences between functions $f$ in $H_0$ and those in $H_1$ may be negligible in an infinite dimensional space. Furthermore, a finite sample size makes it extremely difficult if not impossible to distinguish the alternative hypothesis $H_1$ from $H_0$ in an infinite dimensional space. In other words, the statistical power of a testing procedure may be very low. Moreover, Cox et al. [6] showed that for the hypothesis of $f$ being a polynomial versus $f$ being smooth, there is no uniformly most powerful (UMP) test. Another major difficulty of testing the significance of interaction in SS-ANOVA model is that the null limiting distribution of test statistic is had to obtain since the model under null hypothesis is also nonparametric.

There has been considerable recent work on the development of many statistical tests on testing general departures from a parametric model. For example, Azzalini and Bowman [7] proposed a test on checking whether the nonparametric regression model can reduce to a parametric model; Xiang and Wahba [8] developed the symmetrized Kullback-Leibler (SKL) test based on the SKL distance between the "parametric" function estimated under the null hypothesis and the "smooth" function estimated under the alternative; When the model under null hypothesis is parametric, there are several new efforts on developing the minimax optimal test Fan et al. [9] proposed a generalized likelihood ratio test for testing

nonparametric regression models; Shang et al. [10] developed a penalized likelihood ratio test in smoothing spline model. However, there is no general applicable hypothesis testing methods when the model under the null hypothesis is nonparametric.

In this paper, we develop a nonparametric statistical inference framework for testing the interaction in SS-ANOVA model. In particular, we aim to test $\mathcal{H}_0$ which is the functional space only including the functions with main effects versus $\mathcal{H}_1$ which is functional space including functions with both main effects and the interaction. We propose a "Wald Type" test statistics and derive the asymptotic normality of our proposed test statistic. Moreover, we unveil an interesting Wilks phenomenon, i.e., the asymptotic null distribution is free of nuisance parameters.

To study the power of proposed test, we consider a slightly different alternative hypothesis,

$$H_1^* : ||f_{1,2}(x^{\langle 1\rangle}, x^{\langle 2\rangle})||_2 \geq d_n \tag{1.8}$$

where $||\cdot||_2$ is the $L_2$ norm. Compared to the alternative hypothesis in (1.7), the neighborhood within the distance to $f_{1,2} = 0$ smaller than $d_n$ is removed. Here the sequence $d_n$ is called the distinguishable rate (or separation rate) [11; 12]. We prove that the proposed test has the minimax distinguishable rate in the sense of Ingster and Suslina [11].

The article is organized as follows. Section 2 introduces the background of the SS-ANOVA model, and Section 3 formulates the testing problem and constructs the test statistic. The asymptotic distribution and optimal distinguishable rate are also derived in Section 3. Section 4 shows the simulation results of the proposed test statistics. In Section 5, we apply the proposed test on a genome-wide DNA methylation data set and a neuroimaging data set related to Alzheimer's disease. The detailed proofs are given in Appendix.

## 1.2 Fitting SS-ANOVA Models

In this section, we present the penalized least squares for fitting the SS-ANOVA model. The optimization is performed in a tensor product reproducing kernel Hilbert space (RKHS).

### 1.2.1 Penalized least squares

The unknown function $f$ in (1.5) is estimated through minimizing the penalized least squares functional

$$\frac{1}{na}\sum_{i=1}^{n}\sum_{j=1}^{a}(Y_{ij}-f(\mathbf{x}_{ij}))^2+\lambda_n J(f) \tag{1.9}$$

where $\mathbf{x}_{ij}=(x_i^{\langle 1\rangle},x_j^{\langle 2\rangle})$, the quadratic functional $J(f)$ quantifies the roughness of $f$ and the smoothing parameter $\lambda_n$ controls the trade-off between the goodness-of-fit and the roughness of $f$. The minimization of (1.9) shall be performed in a space $\mathcal{H}\in\{f:J(f)<\infty\}$ in which $J(f)$ is a square seminorm. The evaluation functional $[\mathbf{x}]f=f(\mathbf{x})$ over $\mathcal{H}$ is assumed to be continuous. A Hilbert space $\mathcal{H}$ in which the evaluation is continuous is called a reproducing kernel Hilbert space (RKHS) possessing a kernel function $\mathcal{K}(\cdot,\cdot)$, a nonnegative definite symmetric function satisfying

$$\mathcal{K}(\mathbf{x}',\mathbf{x})=\mathcal{K}(\mathbf{x},\mathbf{x}') \text{ and } \langle\mathcal{K}(\mathbf{x},\cdot),f(\cdot)\rangle=f(\mathbf{x}),$$

for any $\mathbf{x},\mathbf{x}'\in\mathcal{X}=\mathcal{X}_1\times\mathcal{X}_2$ and $f\in\mathcal{H}$, where $\langle\cdot,\cdot\rangle$ is the inner product in $\mathcal{H}$. If $\mathcal{K}$ is also a continuous function, Mercer's Theorem [13] guarantees that there exists an orthogonal basis (a.k.a. eigenfunctions) $\{\phi_i\}_{i=1}^{\infty}$ such that

$$\mathcal{K}(\mathbf{x},\mathbf{x}')=\sum_{i=1}^{\infty}\rho_i\phi_i(\mathbf{x})\phi_i(\mathbf{x}')$$

where $\{\rho_i\}_{i=0}^{\infty}$ are non-negative eigenvalues.

### 1.2.2 SS-ANOVA decomposition on a tensor product RKHS

To incorporate SS-ANOVA for the estimation of the bivariate function in (1.9), we construct a RKHS defined on a product domain. The RKHS on a product domain $\mathcal{X}_1 \times \mathcal{X}_2$ is conveniently constructed by taking the tensor product of the RKHSs on marginal domains $\mathcal{X}_1$ and $\mathcal{X}_2$. Let $\mathcal{H}^{\langle 1\rangle}$ and $\mathcal{H}^{\langle 2\rangle}$ denote the RKHSs on the marginal domain $\mathcal{X}_1$ and $\mathcal{X}_2$ respectively. Their tensor product space is $\mathcal{H} = \mathcal{H}^{\langle 1\rangle} \otimes \mathcal{H}^{\langle 2\rangle}$, where $\otimes$ denotes the tensor product of two spaces.

Consider the one-way ANOVA decomposition of $\mathcal{H}^{\langle 1\rangle}$ on the marginal domain $\mathcal{X}_1$,

$$\mathcal{H}^{\langle 1\rangle} = \mathcal{H}_0^{\langle 1\rangle} \oplus \mathcal{H}_1^{\langle 1\rangle},$$

where "parametric" space $\mathcal{H}_0^{\langle 1\rangle} = \{\mathcal{A}^{\langle 1\rangle} f | f \in \mathcal{H}^{\langle 1\rangle}\}$ has the kernel $\mathcal{K}_0^{\langle 1\rangle}$, "nonparametric" space $\mathcal{H}_1^{\langle 1\rangle} = \{(\mathcal{I} - \mathcal{A}^{\langle 1\rangle}) f | f \in \mathcal{H}^{\langle 1\rangle}\}$ has the kernel $\mathcal{K}_1^{\langle 1\rangle}$, $\mathcal{A}^{\langle 1\rangle}$ denotes an averaging operator and $\mathcal{I}$ is the identity operator. Analogously, we have the one-way ANOVA decomposition of

$$\mathcal{H}^{\langle 2\rangle} = \mathcal{H}_0^{\langle 2\rangle} \oplus \mathcal{H}_1^{\langle 2\rangle}$$

on marginal domain $\mathcal{X}_2$ with corresponding kernels $\mathcal{K}_0^{\langle 2\rangle}$ and $\mathcal{K}_1^{\langle 2\rangle}$. Thus, the two-way SS-ANOVA decomposition of $\mathcal{H} = \mathcal{H}^{\langle 1\rangle} \otimes \mathcal{H}^{\langle 2\rangle}$ is obtained through applying the distributive law,

$$\begin{aligned}
\mathcal{H} &= (\mathcal{H}_0^{\langle 1\rangle} \oplus \mathcal{H}_1^{\langle 1\rangle}) \otimes (\mathcal{H}_0^{\langle 2\rangle} \oplus \mathcal{H}_1^{\langle 2\rangle}) \\
&= (\mathcal{H}_0^{\langle 1\rangle} \otimes \mathcal{H}_0^{\langle 2\rangle}) \oplus (\mathcal{H}_0^{\langle 1\rangle} \otimes \mathcal{H}_1^{\langle 2\rangle}) \oplus (\mathcal{H}_1^{\langle 1\rangle} \otimes \mathcal{H}_0^{\langle 2\rangle}) \oplus (\mathcal{H}_1^{\langle 1\rangle} \otimes \mathcal{H}_1^{\langle 2\rangle}) \\
&= \mathcal{H}_{00} \oplus \mathcal{H}_{01} \oplus \mathcal{H}_{10} \oplus \mathcal{H}_{11}, \qquad (1.10)
\end{aligned}$$

where $\mathcal{H}_{00}$ and $\mathcal{H}_{10}$ are "parametric" spaces, $\mathcal{H}_{01}$ and $\mathcal{H}_{11}$ are "nonparametric spaces". The corresponding kernels for the four RKHSs are $\mathcal{K}_{00} = \mathcal{K}_0^{\langle 1\rangle}\mathcal{K}_0^{\langle 2\rangle}$, $\mathcal{K}_{01} = \mathcal{K}_0^{\langle 1\rangle}\mathcal{K}_1^{\langle 2\rangle}$, $\mathcal{K}_{10} = \mathcal{K}_1^{\langle 1\rangle}\mathcal{K}_0^{\langle 2\rangle}$ and $\mathcal{K}_{11} = \mathcal{K}_1^{\langle 1\rangle}\mathcal{K}_1^{\langle 2\rangle}$ respectively (by Theorem 2.5 [1]). For each subspace $\mathcal{H}_\beta$, $\beta = \{00, 01, 10, 11\}$, we use the norm

$$\langle f, f\rangle_{\mathcal{H}_\beta} = \langle f_\beta, f_\beta\rangle_{\mathcal{H}}$$

where $f_\beta$ is the projection of $f$ from $\mathcal{H}$ to $\mathcal{H}_\beta$.

In this paper, we consider a special case of the above general construction. In particular, we consider the tensor product RKHS

$$\mathcal{H} = \mathcal{H}^{\langle 1\rangle} \otimes \mathcal{H}^{\langle 2\rangle} = \mathcal{S}^m(\mathbb{I}) \otimes \mathbb{R}^2,$$

where $\mathcal{H}^{\langle 1\rangle} = \mathcal{S}^m(\mathbb{I})$ is the $m$th order Sobolev space, $\mathcal{H}^{\langle 2\rangle} = \mathbb{R}^2$ is a two-dimensional Euclidean space. In particular, the $\mathcal{S}^m(\mathbb{I})$ is defined as

$$\begin{aligned}\mathcal{S}^m(\mathbb{I}) \equiv &\{f \in L_2(\mathbb{I}) | f^{(j)} \text{ is absolutely continuous,} \\ &\text{for } j = 0, \ldots, m-1 \text{ and } f^{(m)} \in L_2(\mathbb{I})\}.\end{aligned}$$

For simplicity, we consider a homogeneous subspace $S_0^m(\mathbb{I})$ of $S^m(\mathbb{I})$ which has additional constraints $f^{(j)}(0) = f^{(j)}(1)$ for $j = 0, 1, \ldots, m-1$. Our test can be easily generalized to general Soblev space which is discussed in Section 6. We have the following the inner product,

$$\langle f, \widetilde{f}\rangle_{\mathcal{S}_0^m(\mathbb{I})} = \int_0^1 f(u)du \int_0^1 \widetilde{f}(u)du + \int_0^1 f^{(m)}(u)\widetilde{f}^{(m)}(u)du,$$

defined on $S_0^m(\mathbb{I})$. The one-way ANOVA decomposition is $S_0^m(\mathbb{I}) = \{1\} \oplus \{f \in S_0^m(\mathbb{I}) : \int_0^1 f = 0\}$, and corresponding kernels are defined as

$$\mathcal{K}_0^{\langle 1\rangle}(x^{\langle 1\rangle}, \tilde{x}^{\langle 1\rangle}) = 1, \quad \mathcal{K}_1^{\langle 1\rangle}(x^{\langle 1\rangle}, \tilde{x}^{\langle 1\rangle}) = (-1)^{m-1} k_{2m}(x^{\langle 1\rangle} - \tilde{x}^{\langle 1\rangle})$$

where $k_r(\cdot) = B(\cdot)/r!$ is the scaled Bernoulli polynomials [14]. For simplicity, we use $\mathcal{K}^*$ to denote $\mathcal{K}_1^{\langle 1\rangle}$ in the rest of the paper. The inner product of $\mathbb{R}^2$ is defined as $\langle f, \tilde{f}\rangle = f^T\tilde{f}$ for $f \in \mathbb{R}^2$. The one-way ANOVA decomposition is $\mathbb{R}^2 = \{1\} \oplus \{f \in \mathbb{R}^2 : f^T\mathbf{1}_2 = 0\}$ where $\mathbf{1}_2$ is the $2 \times 1$ vector with all components being 1. The corresponding kernels are defined as

$$\mathcal{K}_0^{\langle 2\rangle}(x^{\langle 2\rangle}, \tilde{x}^{\langle 2\rangle}) = 1/2, \quad \mathcal{K}_1^{\langle 2\rangle}(x^{\langle 2\rangle}, \tilde{x}^{\langle 2\rangle}) = \mathbb{1}_{(x^{\langle 2\rangle} = \tilde{x}^{\langle 2\rangle})} - 1/2$$

where $\mathbb{1}$ is an indicator function. The inner product $\langle \cdot, \cdot \rangle_{\mathcal{H}}$ on the tensor product space $S_0^m(\mathbb{I}) \otimes \mathbb{R}^a$ is induced by the $\langle \cdot, \cdot \rangle_{\mathcal{S}_0^m(\mathbb{I})}$ and the standard inner product of Euclidean space.

### 1.2.3 Model fitting

We shall now fill in some details concerning the roughness penalty $J$ used in the penalized least squares (1.9). The roughness penalty is a square seminorm $J(f) = J(f, f)$, where $J(f, g) = \sum_\beta \theta_\beta^{-1} \langle f, g \rangle_{\mathcal{H}_\beta}$, $\beta = 01, 10, 11$, $\theta_\beta$ are nonnegative weights of the roughness penalty to make the inner products comparable. Since $\mathcal{H}_{01}$ is a parametric space of finite dimension, we do not include penalty on this subspace and set $\theta_{01}$ to infinity. Thus we have

$$J(f, g) = \theta_{10}^{-1} \langle f, g \rangle_{\mathcal{H}_{10}} + \theta_{11}^{-1} \langle f, g \rangle_{\mathcal{H}_{11}}.$$

Since the smoothness of $f$ is also controlled by $\lambda_n$, we assume $\theta_{10} + \theta_{11} = 1$ to avoid an over-parametrization. Recall that the kernel functions for $\mathcal{H}_{10}$ and $\mathcal{H}_{11}$ are $\mathcal{K}_{10}$ and $\mathcal{K}_{11}$ respectively. Then, the kernel associated with $J$ is $\mathcal{K}_J = \theta_{10}\mathcal{K}_{10} + \theta_{11}\mathcal{K}_{11}$.

Without loss of generality, we assume $a = 2$ which means that there is one subject in case group and one subject in control group. Now, we will focus on minimizing the penalized least squares functional (1.9). The minimizer of (1.9) can be found by solving a $n$-dimensional convex problem based on the representer theorem [15]. We define the empirical kernel matrix $K_{10} \in \mathbb{R}^{2n \times 2n}$ and $K_{11} \in \mathbb{R}^{2n \times 2n}$ as,

$$K_{10} = \frac{1}{2} \begin{bmatrix} K & K \\ K & K \end{bmatrix}, \quad K_{11} = \frac{1}{2} \begin{bmatrix} K & -K \\ -K & K \end{bmatrix}. \tag{1.11}$$

where the $(i, i')$th entry of $K$ is $\frac{1}{n}\mathcal{K}^*(x_i^{\langle 1 \rangle}, x_{i'}^{\langle 1 \rangle})$. The estimation then reduces to the minimization of

$$\frac{1}{2n} ||\mathbf{y} - S\mathbf{d} - nR\mathbf{c}||_2^2 + n\lambda_n \mathbf{c}^T R \mathbf{c}$$

with respect to $\mathbf{d}$ and $\mathbf{c}$, where $\mathbf{y} = (Y_{11}, \ldots, Y_{n1}, Y_{12}, \ldots, Y_{n2})^T$, $R = \theta_{10}K_{10} + \theta_{11}K_{11}$, $S \in \mathbb{R}^{2n \times 2}$ can be written as

$$S = \begin{bmatrix} \mathbf{1}_n & \mathbf{1}_n \\ \mathbf{1}_n & \mathbf{0} \end{bmatrix}, \tag{1.12}$$

and $\mathbf{1}_n$ is a $n \times 1$ vector with all components being 1. The estimates of $\mathbf{d}$ and $\mathbf{c}$ are (see Chapter 3 of [1] for details),

$$\hat{\mathbf{c}} = \frac{1}{n}[M^{-1} - M^{-1}S(S^TM^{-1}S)^{-1}S^TM^{-1}]\mathbf{y},$$
$$\hat{\mathbf{d}} = (S^TM^{-1}S)^{-1}S^TM^{-1}\mathbf{y},$$

where $M = R + \lambda_n I_{2n}$ and $I_{2n}$ is the $2n \times 2n$ identity matrix. The fitted values of function $f$ are

$$\begin{aligned}\hat{\mathbf{f}} &= S\hat{\mathbf{d}} + nR\hat{\mathbf{c}} \\ &= (RM^{-1} - RM^{-1}S(S^TM^{-1}S)^{-1}S^TM^{-1} + S(S^TM^{-1}S)^{-1}S^TM^{-1})\mathbf{y}.\end{aligned}$$

We can also get the fitted values of each components in the SS-ANOVA (1.6). In particular, the fitted values of main effect $f_1$ are

$$\hat{\mathbf{f}}_1 = n\theta_{10}K_{10}\hat{\mathbf{c}} = \theta_{10}K_{10}(M^{-1} - M^{-1}S(S^TM^{-1}S)^{-1}S^TM^{-1})\mathbf{y},$$

and the fitted values of interaction $f_{1,2}$ are

$$\hat{\mathbf{f}}_{1,2} = n\theta_{11}K_{11}\hat{\mathbf{c}} = \theta_{11}K_{11}(M^{-1} - M^{-1}S(S^TM^{-1}S)^{-1}S^TM^{-1})\mathbf{y}. \tag{1.13}$$

## 1.3 Testing the significance of SS-ANOVA Model

In this section, we construct our "Wald-type" test based on the penalized least squares estimation of the unknown function $f \in \mathcal{H}$ and establish the asymptotic properties for the proposed test.

### 1.3.1 Test statistic

In SS-ANOVA model (1.6), we aim to test the significance of the interaction $f_{1,2}$, i.e. (1.7), or equivalently,

$$H_0 : f \in \mathcal{H}_{00} \oplus \mathcal{H}_{01} \oplus \mathcal{H}_{10} \quad v.s. \quad H_1 : f \in \mathcal{H}/(\mathcal{H}_{00} \oplus \mathcal{H}_{01} \oplus \mathcal{H}_{10})$$

We construct a "Wald-type" test statistic based on the (1.13) i.e. fitted values of $f_{1,2}$,

$$T_{n,\lambda_n} = \frac{1}{n}||\hat{\mathbf{f}}_{1,2}||_2^2 = \frac{\theta_{11}^2}{n}||K_{11}(M^{-1} - M^{-1}S(S^TM^{-1}S)^{-1}S^TM^{-1})\mathbf{y}||_2^2 \tag{1.14}$$

where $||\cdot||_2$ is $l_2$ norm. We have the following vector representation of $\mathbf{y}$,

$$\mathbf{y} = \mathbf{f}_0 + \mathbf{f}_1 + \mathbf{f}_2 + \mathbf{f}_{1,2} + \boldsymbol{\epsilon} \tag{1.15}$$

where $\mathbf{f}_0 = f_0\mathbf{1}$, $\mathbf{f}_1 = (f_1(x_1^{\langle 1\rangle}), \ldots, f_1(x_n^{\langle 1\rangle}), f_1(x_1^{\langle 1\rangle}), \ldots, f_1(x_n^{\langle 1\rangle}))^T$, $\mathbf{f}_2 = (f_2(x_1^{\langle 2\rangle}), \ldots, f_2(x_1^{\langle 2\rangle}), f_2(x_2^{\langle 2\rangle}), \ldots,$
$\mathbf{f}_{1,2} = (f_{1,2}(x_1^{\langle 1\rangle}, x_1^{\langle 2\rangle}), \ldots, f_{1,2}(x_n^{\langle 1\rangle}, x_1^{\langle 2\rangle}), f_{1,2}(x_1^{\langle 1\rangle}, x_2^{\langle 2\rangle}), \ldots, f_{1,2}(x_n^{\langle 1\rangle}, x_2^{\langle 2\rangle}))^T$, and $\boldsymbol{\epsilon} = (\epsilon_{11}, \ldots, \epsilon_{n1}, \epsilon_{12}, \ldots,$
$\epsilon_{n2})^T$. Plugging (1.15) back into (1.14), we have

$$T_{n,\lambda_n} = \frac{\theta_{11}^2}{n}||K_{11}M^{-1}(I_n - S(S^TM^{-1}S)^{-1}S^TM^{-1})(\mathbf{f}_0 + \mathbf{f}_1 + \mathbf{f}_2 + \mathbf{f}_{1,2} + \boldsymbol{\epsilon})||_2^2. \tag{1.16}$$

There exists a $\mathbf{d}_0 \in \mathbb{R}^2$ such that $\mathbf{f}_0 + \mathbf{f}_2 = S\mathbf{d}_0$. Simple algebra yields

$$\begin{aligned} &K_{11}M^{-1}(I_n - S(S^TM^{-1}S)^{-1}S^TM^{-1})(\mathbf{f}_0 + \mathbf{f}_2) \\ &\qquad = K_{11}M^{-1}(I_n - S(S^TM^{-1}S)^{-1}S^TM^{-1})S\mathbf{d}_0 = 0. \end{aligned} \tag{1.17}$$

Furthermore, we have

$$K_{11}M^{-1}(I - S(S^TM^{-1}S)^{-1}S^TM^{-1})\mathbf{f}_1 = 0. \tag{1.18}$$

See Appendix 1.8.1 for detailed derivation of (1.18). The test statistic in (1.16) thus reduces to

$$T_{n,\lambda_n} = \frac{\theta_{11}^2}{n}||K_{11}M^{-1}(I_n - S(S^TM^{-1}S)^{-1}S^TM^{-1})(\mathbf{f}_{1,2} + \boldsymbol{\epsilon})||_2^2. \tag{1.19}$$

### 1.3.2 Asymptotic distribution of the test statistic

Test statistic $T_{n,\lambda_n}$ in (1.19) under the null hypothesis can be expressed as,

$$\begin{aligned} T_{n,\lambda_n} &= \frac{\theta_{11}^2}{n}||K_{11}M^{-1}(I_n - S(S^TM^{-1}S)^{-1}S^TM^{-1})\boldsymbol{\epsilon}||_2^2 \\ &= \frac{\theta_{11}^2}{n}||K_{11}M^{-1}\boldsymbol{\epsilon}||_2^2 + \frac{\theta_{11}^2}{n}||K_{11}M^{-1}S(S^TM^{-1}S)^{-1}S^TM^{-1}\boldsymbol{\epsilon}||_2^2 \\ &\quad - \frac{2\theta_{11}^2}{n}\boldsymbol{\epsilon}^TM^{-1}S(S^TM^{-1}S)^{-1}S^TM^{-1}K_{11}^2M^{-1}\boldsymbol{\epsilon}. \end{aligned} \tag{1.20}$$

We shall now derive the asymptotic distribution of test statistic $T_{n,\lambda_n}$ under the null hypothesis. The primary tool used in the development is the eigenvalue analysis of kernel function and matrix. Since our test statistic $T_{n,\lambda_n}$ involves the empirical kernel matrix $K$, we need to characterize the eigenvalues of $K$. To do so, we need to build a connection between eigenvalues of $K$ and those of the population kernel function $\mathcal{K}^*$. Let $\mu_1 \geq \mu_2 \geq \ldots$ be the eigenvalues for kernel function $\mathcal{K}^*$. The $i$th eigenvalue has the order $\mu_i \asymp i^{-2m}$ [16]. The first few eigenvalues play an important role in determining the distribution of test statistics $T_{n,\lambda_n}$. We define the effective dimension $s_{\lambda_n}$ of the eigenvalues of the population kernels as the number of eigenvalues larger than $\lambda_n$, i.e.,

$$s_{\lambda_n} = \max\{i \mid \mu_i \geq \lambda_n\}. \tag{1.21}$$

We define $\hat{\mu}_1 \geq \hat{\mu}_2 \geq \ldots \hat{\mu}_n \geq 0$ as the empirical eigenvalues of kernel matrix $K$. We then define the effective dimension analogous to (1.21) as,

$$\hat{s}_{\lambda_n} = \max\{i \mid \hat{\mu}_i \geq \lambda_n\}. \tag{1.22}$$

We characterize the relationship between eigenvalues of empirical kernel matrix and the population kernel function on the large eigenvalues under two commonly used designs. The first one is the random design which assumes $x^{\langle 1\rangle}$ is sampled from the probability density $\omega^{\langle 1\rangle}$. The second one is the fixed design with $x^{\langle 1\rangle}$ evenly distributed on $[0, 1]$. We summarize these two designs in the following:

**Assumption 1.** *(a). $X_1^{\langle 1\rangle}, X_2^{\langle 1\rangle}, \ldots, X_n^{\langle 1\rangle}$ are IID with a probability density function $\omega^{\langle 1\rangle}$ subject to*

$$c_1 \leq \omega^{\langle 1\rangle}(t) \leq c_2 \quad \textit{for all} \quad t \in [0, 1]$$

*for positive constants $c_1$ and $c_2$. Such design is referred as quasi-uniform design [17].*

*(b). $X_1^{\langle 1\rangle}, X_2^{\langle 1\rangle}, \ldots, X_n^{\langle 1\rangle}$ are evenly distributed on $[0, 1]$.*

Under Assumption 1(a), by Lemma 1.8.1, we have

$$\hat{s}_{\lambda_n} \asymp s_{\lambda_n}, \tag{1.23}$$

with probability at least $1 - (n^{\frac{2}{2m-1}-2\epsilon} + n^{\frac{1}{2m-1}}) \exp\{-cn^{\frac{2m-3}{2m-1}+2\epsilon}\}$ for $\lambda_n > 1/n$, $m > 3/2$, and any $\epsilon > 0$. Under Assumption 1(b), the eigenvalues could be explicitly calculated since the kernel matrix is a circulant matrix [16]. In Lemma 1.8.2, we prove that (1.23) is satisfied for $m > 1/2$.

The following theorem provides the asymptotic distribution of our test statistic under the null hypothesis.

**Theorem 1.3.1.** *If Assumption 1 holds, as $n \to \infty$, $\lambda_n > \frac{1}{n}$ and $\lambda_n \to 0$, under the null hypothesis $H_0$, we have*

$$\frac{T_{n,\lambda_n} - \mu_{n,\lambda_n}}{\sigma_{n,\lambda_n}} \to N(0,1)$$

*where $\mu_{n,\lambda_n} = \theta_{11}^2 \sigma^2 Tr(\Delta)/n$ and $\sigma_{n,\lambda_n}^2 = 2\theta_{11}^4 \sigma^4 Tr(\Delta^2)/n^2$ with $\Delta = M^{-1} K_{11}^2 M^{-1}$.*

The Wald test in linear regression is constructed using the minimizer of the ordinary least squares estimators. Our test is built based on the minimizer of penalized least squares. In a finite dimensional space, the square norm of the fitted values usually follows a chi-square distribution with finite degrees of freedom. Here we call our test a "Wald-type" test in the sense that we also use the squared $l_2$ of the fitted values. However, in an infinite dimensional space, $T_{n,\lambda}$ has diverging degrees of freedom, and thus it is asymptotically normal distributed. In addition, we notice that the asymptotic distribution is free from the "parametric" nuisance parameters $\mu$, $f_2$ and "nonparametric" nuisance parameters $f_1$. This property is also known as the Wilks phenomenon which is an ideal property in nonparametric hypothesis testing [9; 18].

### 1.3.3 Asymptotic optimality

For hypothesis testing with single parameter, there exists the uniformly most powerful or uniformly most powerful unbiased test [19]. However, the uniformly most powerful or uniformly most powerful unbiased test does not exist when the parameter space is infinite-dimensional [11]. Alternatively, we study the proposed test through an asymptotically minimax approach [11].

Based on the test statistic $T_{n,\lambda_n}$, the decision rule $\Phi_{n,\lambda_n}$ for testing hypothesis (1.7) is

$$\Phi_{n,\lambda_n} = \mathbb{1}(|T_{n,\lambda_n} - \mu_{n,\lambda_n}| \geq z_{1-\alpha/2}\sigma_{n,\lambda_n})$$

where $z_{1-\alpha/2}$ is the $1-\alpha/2$ quantile of the standard normal distribution. Instead of the alternative hypothesis of (1.7), we consider a slightly different alternative hypothesis (1.8). Compared to $f_{1,2}$ in the alternative hypothesis of (1.7), $f_{1,2}$ in the alternative hypothesis (1.8) is not only nonzero, but also $d_n$ distance away from zero. Consequently, the power of the decision rule $\Phi_{n,\lambda_n}$ is

$$\mathbb{P}(\Phi_{n,\lambda_n} = 1 \mid f_{1,2} \in \mathcal{H}_{11}, ||f_{1,2}||_2 \geq d_n).$$

Asymptotically, it is highly desirable that our test power goes to one as the sample size goes to infinity with $d_n$ as small as possible. Under an additional assumption, we show that this is true for our test.

**Assumption 2.** *We assume the function in $\mathcal{H}$ has bounded norm. Without loss of generality, we assume,*

$$||f||_{\mathcal{H}} \leq 1$$

*where $||\cdot||_{\mathcal{H}}$ denotes the norm on the tensor product RKHS $\mathcal{H}$.*

Assumption 2 is commonly used in studying the convergence rate of the kernel estimator [10; 20]. The following theorem states that the our test has asymptotically maximum power when the order of distinguishable rate is at least $d_n$.

**Theorem 1.3.2.** *Suppose Assumptions 1 and 2 are satisfied, as* $n \to \infty$, $\lambda_n \to 0$, *for any* $\epsilon > 0$, *there exists constants* $N > 0$ *such that for any* $n > N$, *we have*

$$\inf_{||f_{1,2}||_2 \succeq d_n} \mathbb{P}(\Phi_{n,\lambda_n} = 1) > 1 - 2\epsilon$$

*where* $d_n = \sqrt{\lambda_n + \sigma_{n,\lambda_n}}$.

In smoothing spline ANOVA model, the convergence rate is a trade-off between the bias and variance of the $\hat{f}$. It has been shown in [21] that $\lambda_n$ and $s_{\lambda_n}/n$ are the bias and variance of $\hat{f}$ respectively. Following the similar derivation, we can obtain the bias and variance of $\hat{f}_{12}$ which have the same order with bias and variance of $\hat{f}$. Through balancing these two terms, as shown in the left panel of Figure 1.1, the optimal estimation rate is achieved when $\lambda_n^\dagger = \mathcal{O}(n^{-2m/(2m+1)})$. In contrast, the distinguishable rate $d_n$ is a trade-off between the bias of the $\hat{f}_{12}$ and the standard deviation of $T_{n,\lambda}$. The next theorem provides the optimal distinguishable rate.

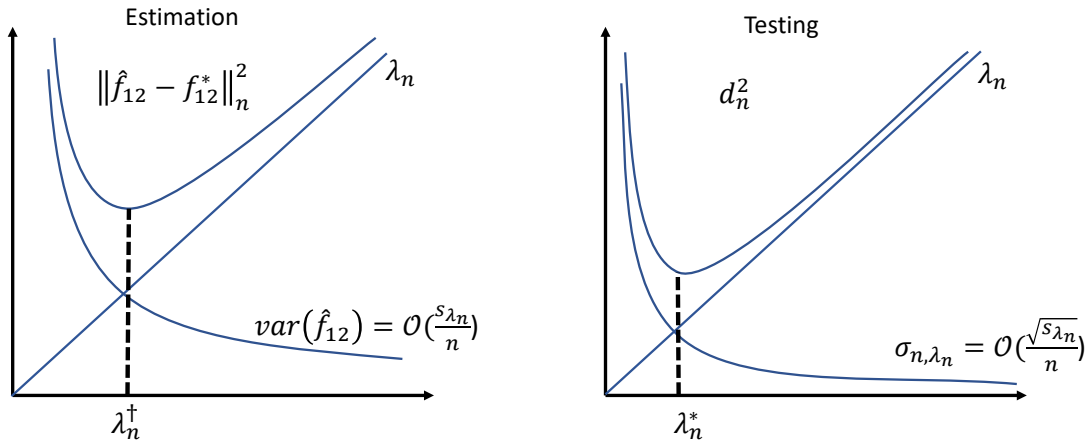



**Figure 1.1:** The left panel shows the trade-off for achieving the optimal estimation rate where $f_{12}^*$ denotes the underlying true interaction. The right panel shows the trade-off for achieving the optimal distinguishable rate.

**Theorem 1.3.3.** *(Optimal distinguishable rate) Suppose Assumptions 2 holds, as* $n \to \infty$, $\lambda_n \to 0$ , *the optimal distinguishable rate is*

$$d_n^* = \mathcal{O}(n^{-2m/(4m+1)})$$

*which is achieved when* $\lambda_n^* = \mathcal{O}(n^{-4m/(4m+1)})$ *with probability at least* $1 - 4\exp(-n^{1/(2m+1)})$ *under Assumption 1(a), and almost surely under Assumption 1(b).*

Theorem 1.3.3 states that the decision rule of proposed test can detect any local alternatives with distinguishable rates no larger than $n^{-2m/(4m+1)}$. In this theorem, we derive that the standard derivation of $T_{n,\lambda_n}$ is $\mathcal{O}(\sqrt{s_{\lambda_n}}/n)$ (see (1.57) for details) which is a decreasing function of $\lambda_n$. As shown in the right panel of Figure 1.1, the optimal distinguishable rate is achieved when $\lambda_n^* = \mathcal{O}(n^{-4m/(4m+1)})$. Furthermore, this optimal distinguish rate, i.e., $d_n^*$, turns out to be the minimax distinguishable rate of the nonparametric test [11]. Thus, we conclude that the proposed "Wald-type" test is minimax optimal.

## 1.4 Simulation Study

To assess the performance of the proposed "Wald-type" test, we carried out extensive analyses on simulated data sets.

### 1.4.1 Power Analysis

We generated samples according to (1.5) for $i = 1, \ldots, n$ and $j = 1, 2$ where $f$ was set to be one of the four functions in settings 1-4 which were plotted in Figure 1.2. We generated i.i.d. $x_1^{\langle 1\rangle}, \ldots, x_n^{\langle 1\rangle}$ from $U(0, 1)$ and set $x_1^{\langle 2\rangle} = 0$ for control group and $x_2^{\langle 2\rangle} = 1$ for case group. We generated i.i.d. $\epsilon_{ij}$s from $N(0, 1)$. The sample size $2n$ was set to be $200, 400, \ldots, 1600, 2000$. We repeated the above procedures to simulate 500 replicated samples.

**Setting 1**.(Case and control have difference in magnitude).

$$f(x^{\langle 1\rangle}, x^{\langle 2\rangle}) = 2.5\sin(3\pi x^{\langle 1\rangle})(1 - x^{\langle 1\rangle})\mathbb{1}([x^{\langle 2\rangle} = 0]) + (2.5 - \delta_1)\sin(3\pi x^{\langle 1\rangle})(1 - x^{\langle 1\rangle})\mathbb{1}([x^{\langle 2\rangle} = 1])$$

where we set $\delta_1$ to be $0.5, 1$ and $1.5$ to characterize the difference in magnitude between the two groups.

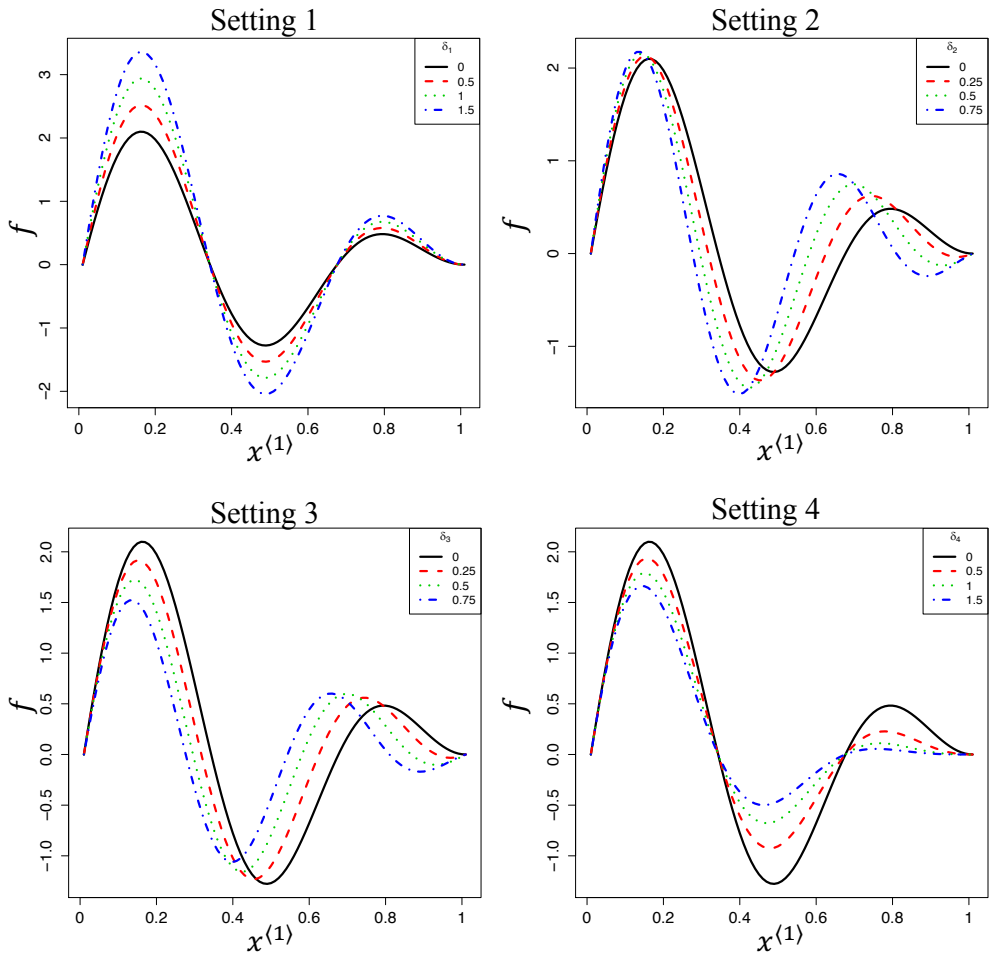



**Figure 1.2:** Curves of functions in Settings 1-4. The solid curve is curve for $x^{\langle 2 \rangle} = 0$. The dashed, dotted and dot-dash lines are curves for $x^{\langle 2 \rangle} = 1$ with distinguishable parameter taken different values.

**Setting 2**.(Case and control have difference in frequency).

$$f(x^{\langle 1 \rangle}, x^{\langle 2 \rangle}) = 2.5 \sin(3\pi x^{\langle 1 \rangle})(1 - x^{\langle 1 \rangle})\mathbb{1}([x^{\langle 2 \rangle} = 0]) + 2.5 \sin((3 - \delta_2)\pi x^{\langle 1 \rangle})(1 - x^{\langle 1 \rangle})\mathbb{1}([x^{\langle 2 \rangle} = 1])$$

where we set $\delta_2$ to be 0.25, 0.5 and 0.75 to characterize the difference in frequency between the two groups.

**Setting 3**.(Case and control have difference in both magnitude and frequency).

$$f(x^{\langle 1\rangle}, x^{\langle 2\rangle}) = 2.5\sin(3\pi x^{\langle 1\rangle})(1-x^{\langle 1\rangle})\mathbb{1}([x^{\langle 2\rangle}=0]) \\ + (2.5-\delta_3)\sin((3-\delta_3)\pi x^{\langle 1\rangle})(1-x^{\langle 1\rangle})\mathbb{1}([x^{\langle 2\rangle}=1])$$

where we set $\delta_3$ to be 0.25, 0.5 and 0.75 to characterize difference in both magnitude and frequency between the two groups.

**Setting 4**.(Case and control have nonlinear difference in magnitude).

$$f(x^{\langle 1\rangle}, x^{\langle 2\rangle}) = 2.5\sin(3\pi x^{\langle 1\rangle})(1-x^{\langle 1\rangle})\mathbb{1}([x^{\langle 2\rangle}=0]) \\ + 2.5\sin(3\pi x^{\langle 1\rangle})(1-x^{\langle 1\rangle})^{(1+\delta_4)}([x^{\langle 2\rangle}=1])$$

where we set $\delta_4$ to be 0.5, 1 and 1.5 to characterize nonlinear difference in magnitude between the two groups.

We applied our proposed test with significance level $\alpha = 0.05$ to the simulated datasets. In addition, we compared the proposed test with the F-test in linear regression model,

$$y_{ij} = \mu + \beta_1 x_i^{\langle 1\rangle} + \beta_2 x_j^{\langle 2\rangle} + \beta_{12} x_i^{\langle 1\rangle} x_j^{\langle 2\rangle} + \epsilon_{ij},$$

where $\mu$ is the grand mean, $\beta_1$ and $\beta_2$ are the coefficients for main effects and $\beta_{12}$ is the coefficient for the interaction. We used the classical F-test of analysis of variance (ANOVA) to test

$$H_0 : \beta_{12} = 0 \quad v.s. \quad H_1 : \beta_{12} \neq 0 \tag{1.24}$$

i.e. testing whether the interaction is significant.

For Setting 1, we plotted in Figure 1.3 the empirical power curves of our proposed test and classical ANOVA test for three different $\delta_1$. The empirical power of our test increases rapidly as sample size increases, and approaches to 1 even for the smallest magnitude ($\delta_1 = 0.5$). In contrast, the empirical power of classical F test increase much slower than our proposed test. When $\delta_1 = 0.5$, the empirical power of the classical F test is still less than 0.2 even when the sample size is as large as 2000.

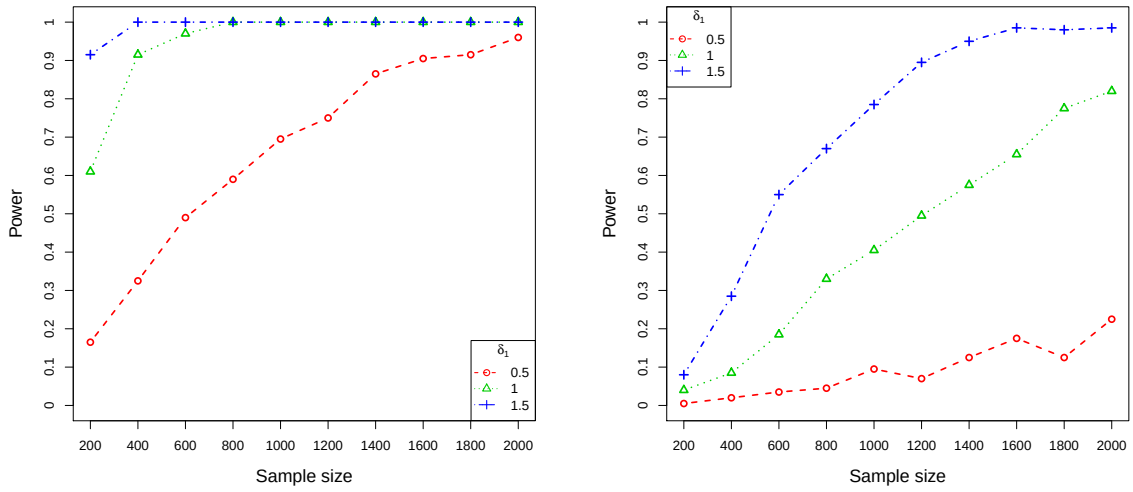



**Figure 1.3:** (Setting 1) Plots of empirical power against the sample size. Red dashed, green dotted and blue dot-dash lines denote different magnitude parameters $\delta_1 = 0.5, 1, 1.5$ respectively. *Left*: The empirical power of our proposed test. *Right*: The empirical power of classical ANOVA test.

For setting 2, as shown in Figure 1.4, the empirical power of our proposed test converges to 1 as the sample size increases for all three cases with $\delta_2 = 0.2$, 0.4 and 0.6. In contrast, the F-test has very small power to detect the difference in frequency. The empirical power is still lower than 0.2 when the sample size is as large as 2000.

Compared with setting 1, setting 3 included additional frequency differences. As shown in Figure 1.5, the empirical power of our proposed test increases for all the three cases with $\delta_3 = 0.25, 0.5, 0.75$. In contrast, the power of classical ANOVA test is smaller than setting 1.

For setting 4, we consider that there is a nonlinear difference in magnitude along the $x^{\langle 1 \rangle}$ between the two groups. As shown in Figure 1.6, F-test has nearly no power to detect the nonlinear magnitude changes even when $\delta_4 = 1.5$. In contrast, the empirical power of our proposed converges to 1 rapidly as the sample size increases.

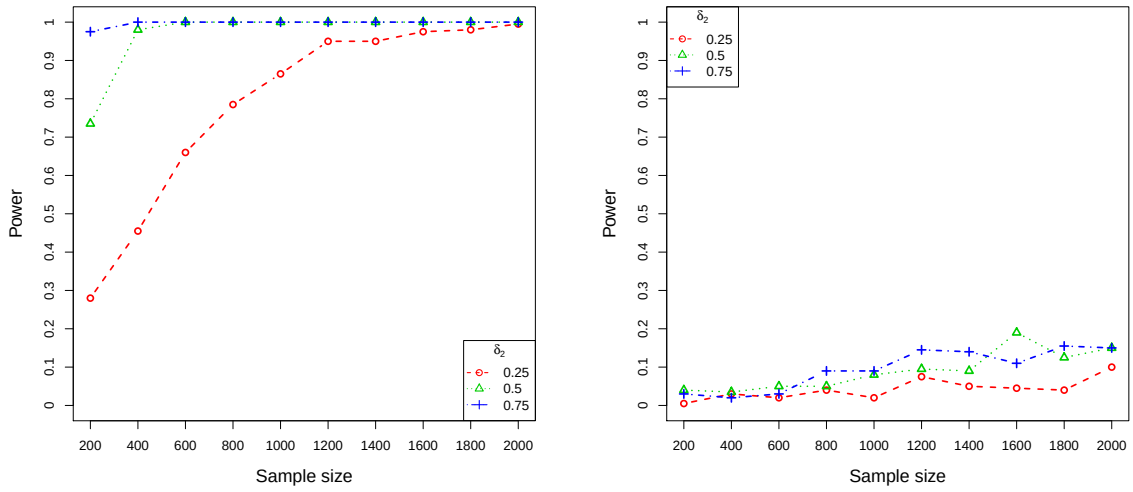



**Figure 1.4:** (Setting 2) Plots of empirical power against the sample size. Red dashed, green dotted and blue dot-dash lines denote different frequency parameters $\delta_2 = 0.25, 0.5, 0.75$ respectively. *Left*: The empirical power of proposed test. *Right*: The empirical power of classical ANOVA test.

### 1.4.2 Significance Levels

To examine the approximation of significance levels, we keep the function form of control group same with the previous section. We only added a parallel shift for the case group, i.e., the model does not have the interaction term. We generated data from (1.5) with function $f$ specified in Setting 5, and rest of parameters were set as before.

**Setting 5**.

$$f(x^{\langle 1\rangle}, x^{\langle 2\rangle}) = 2.5\sin(3\pi x^{\langle 1\rangle})(1 - x^{\langle 1\rangle}) + \delta_5 I_{[x^{\langle 2\rangle}=1]}$$

where we set $\delta_5$ to be 0, 0.5 and 1 to characterize different level parallel difference in the two groups.

Figure 1.7 plots the empirical sizes of our proposed test and ANOVA F-test under Setting 5. We varied $\delta$ from 0 to 1 to model different magnitudes of main effect. As shown in Figure

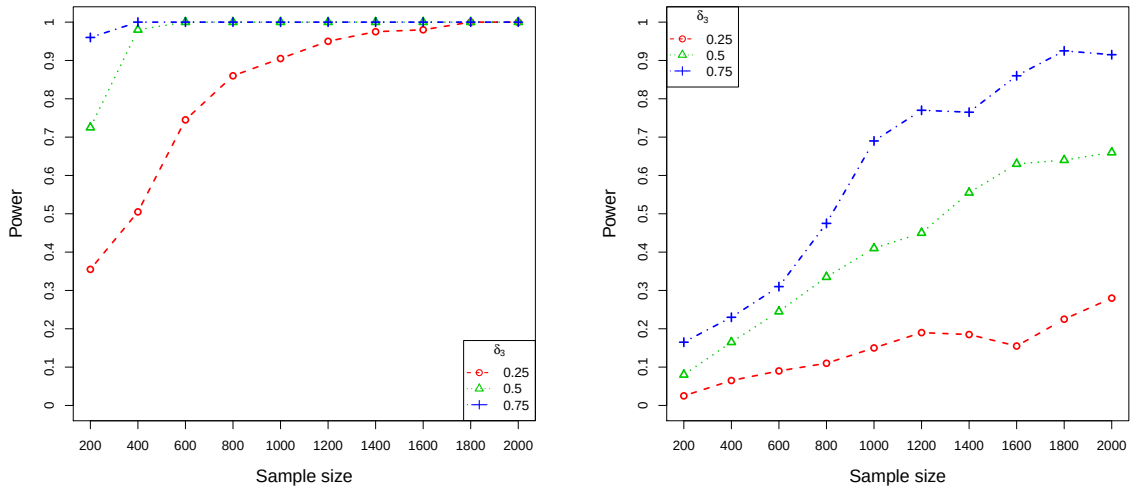



**Figure 1.5:** (Setting 3) Plots of empirical power against the sample size. Red dashed, green dotted and blue dot-dash lines denote different magnitude-frequency parameters $\delta_3 = 0.25, 0.5, 0.75$ respectively. *Left*: The empirical power of our proposed test. *Right*: The empirical power of classical ANOVA test.

5, the empirical size of our proposed test approaches to 0.05 as the sample size increases for different values of $\delta$. However the empirical sizes of ANOVA F-test are all significantly lower than 0.05.

## 1.5 Real Data Examples

We now apply the technique to analyze a couple of real datasets.

### 1.5.1 DNA methylation in chronic lymphocytic leukemia

Recently, Filarsky et al. [5] reported a DNA methylation study for chronic lymphocytic leukemia (CLL) patients. In the study, the DNA samples were extracted from CD19+ cells from 12 CLL patients and B cells from 6 normal subjects. The DNA methylation is profiled

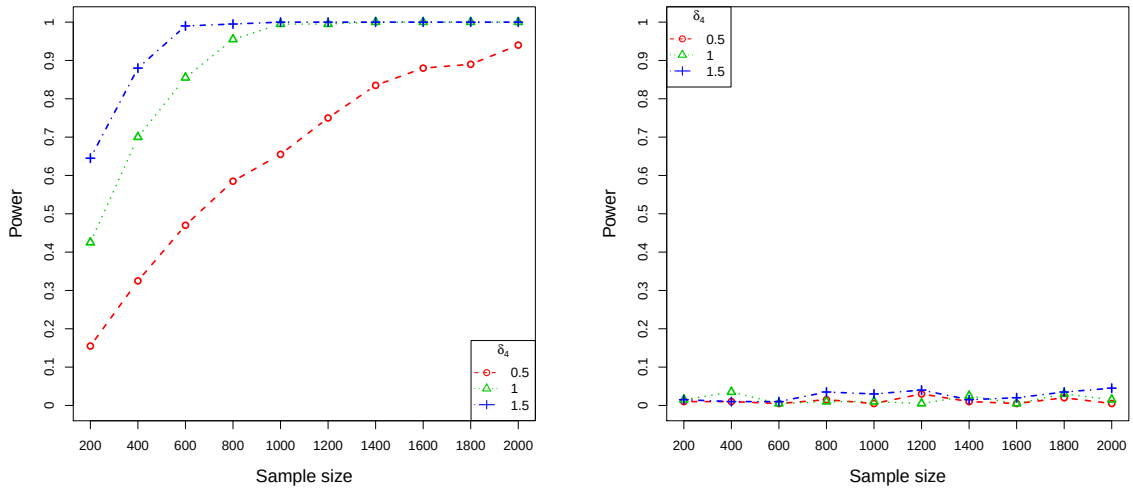



**Figure 1.6:** (Setting 4) Plots of empirical power against the sample size. Red dashed, green dotted and blue dot-dash lines denote different nonlinear magnitude parameters $\delta_4 = 0.5, 1, 1.5$ respectively. *Left*: The empirical power of our proposed test. *Right*: The empirical power of classical ANOVA test.

by the whole-genome tiling array technique. The goal is to identify differentially methylated regions (DMRs) between CLL patients and normal subjects.

To achieve this goal, we compiled the DNA methylation intensities within the $-3.8$ to $+1.8$ kb of transcription start sites (TSS) for each gene. We used the M-value suggested by [22] as methylation level at each site and as our response variable. We fitted the nonparametric model in (1.3), and we tested the hypothesis in (1.4) on 10383 regions.

Through control of the FDR $< 0.01$ using Benjamini-Hochberg Procedure [23], we selected 613 DMRs. We conducted gene ontology analysis on the 613 genes corresponding 613 identified DMRs using the GSEA [24]. Among all genes, 79 genes participate the lipid metabolic process which plays an important role in the development of CLL [25]. This biological process contributes to apoptosis resistance in CLL cells. Furthermore, 78 and 61 genes

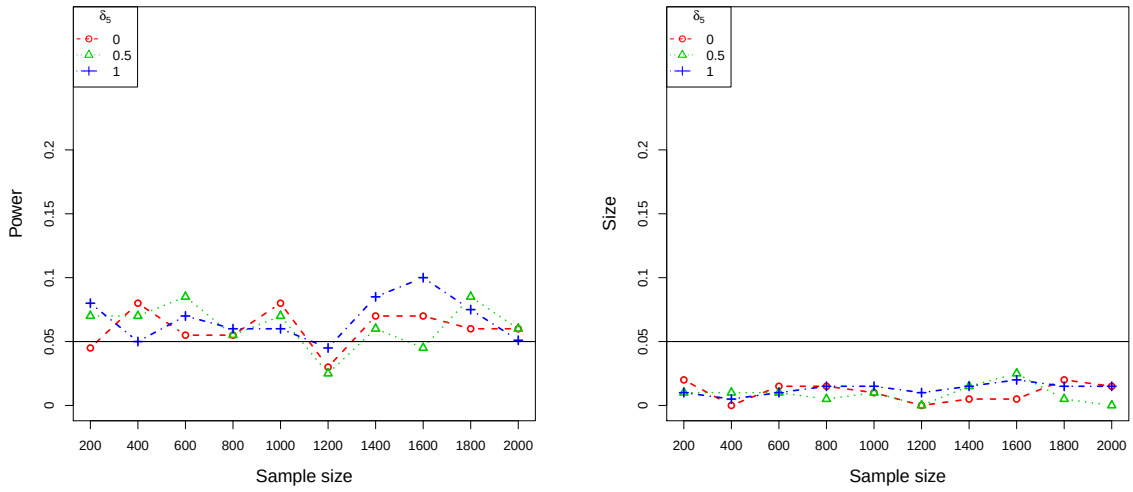
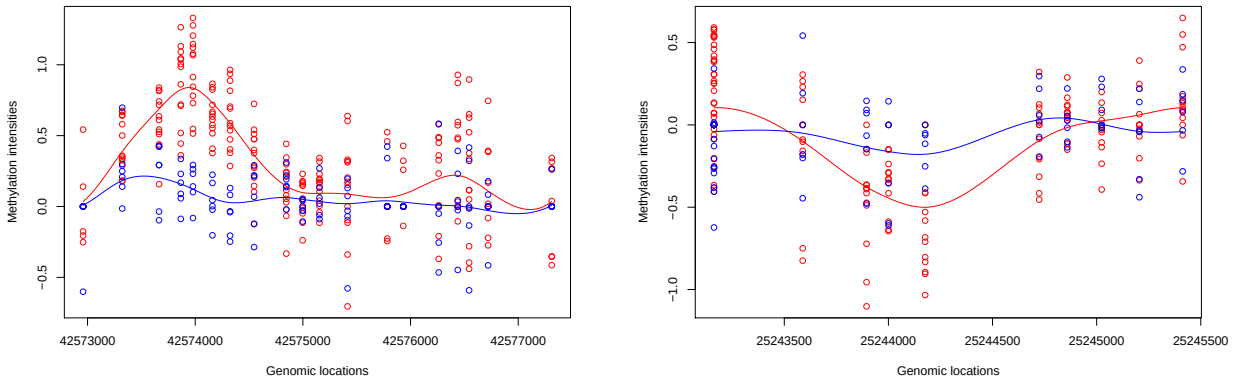



**Figure 1.7:** (Setting 5) Plots of empirical size against the sample size. Red dashed, green dotted and blue dot-dash lines denote $\delta_5 = 0.5, 1, 1.5$ respectively. *Left*: The empirical size of our proposed test. *Right*: The empirical size of classical ANOVA test.

participate the immune related biological process "Immune system process" and "Regulation of immune system process" respectively, which indicates that the aberrant DNA methylation has the potential impact on the immune system.

We highlighted two DMRs with significant interaction in Figure 1.8. The focal hypermethylation at 42574000 and 42576500 are observed on the promoter region of gene MTA3. It was reported in [26] that MTA3 signaling pathway is a potential bio-marker for CLL and shows significantly altered gene expression. Our test also identified that the methylation pattern of MTA3 gene has significantly altered, which has the potential prognostic value. In the promoter region of DNMT3, we observed significantly hypomethylation at 25244500 genomic location. DNMT3 is a family of DNA methyltransferases that could methylate hemimethylated and unmethylated CpG sites at the same rate [27]. Since the global hypomethylation




(*a*) MTA3 (*b*) DNMT3A

**Figure 1.8:** The promoter regions of two genes, (a) MTA3 and (b) DNMT3A. The horizontal axis is the genomic location and the y axis is the M-value representing the methylation intensities. The red and blue line are the fitted curves for the case and control groups respectively.

are observed, the aberrant methylation pattern of this DNA methylatransferases may have influence on this global trend.

## 1.6 Extensions

In some applications, some signals might have strong linear pattern or quadratic pattern. In these situation, the polynomial spline can better fit this situation. The proposed test can be easily generalized to the polynomial spline kernels. Here we consider $\mathcal{H}^{\langle 1\rangle} = S^{(m)}$ and $\mathcal{H}^{\langle 2\rangle} = \mathbb{R}^2$. For the general $m$th order Sobolev space $S^m(\mathbb{I})$, we use the inner product

$$(f, \widetilde{f})_{\mathcal{S}^m(\mathbb{I})} = \sum_{\nu=0}^{m-1} \int_0^1 f^{(\nu)}(u)du \int_0^1 \widetilde{f}^{(\nu)}(u)du + \int_0^1 f^{(m)}(u)\widetilde{f}^{(m)}(u)du,$$

which is associated with a reproducing kernel constructed by Bernoulli polynomials [1]. For example, when $m = 2$, one has the tensor sum decomposition of $S^2(\mathbb{I})$ as

$$\begin{aligned} S^2(\mathbb{I}) =& \{f : f \propto 1\} \oplus \{f : f \propto x - 0.5\} \oplus \{f : \int_0^1 f = \int_0^1 \dot{f} = 0, \ddot{f} \in L_2(\mathbb{I})\} \\ =& \mathcal{H}_{00}^{\langle 1 \rangle} \oplus \mathcal{H}_{01}^{\langle 1 \rangle} \oplus \mathcal{H}_{1}^{\langle 1 \rangle}. \end{aligned}$$

More details of this decomposition are shown in Example 2.5 [1]. Thus, the ANOVA decomposition on the tensor product space, $S^2(\mathbb{I}) \otimes \mathbb{R}^2$, can be written as

$$\begin{aligned} \mathcal{H} =& (\mathcal{H}_{00}^{\langle 1 \rangle} \oplus \mathcal{H}_{01}^{\langle 1 \rangle} \oplus \mathcal{H}_{1}^{\langle 1 \rangle}) \otimes (\mathcal{H}_{0}^{\langle 2 \rangle} \oplus \mathcal{H}_{1}^{\langle 2 \rangle}) \\ =& (\mathcal{H}_{00}^{\langle 1 \rangle} \otimes \mathcal{H}_{0}^{\langle 2 \rangle}) \oplus (\mathcal{H}_{01}^{\langle 1 \rangle} \otimes \mathcal{H}_{0}^{\langle 2 \rangle}) \oplus (\mathcal{H}_{1}^{\langle 1 \rangle} \otimes \mathcal{H}_{0}^{\langle 2 \rangle}) \oplus (\mathcal{H}_{00}^{\langle 1 \rangle} \otimes \mathcal{H}_{1}^{\langle 2 \rangle}) \\ & (\mathcal{H}_{01}^{\langle 1 \rangle} \otimes \mathcal{H}_{1}^{\langle 2 \rangle}) \oplus (\mathcal{H}_{1}^{\langle 1 \rangle} \otimes \mathcal{H}_{1}^{\langle 2 \rangle}) \\ =& \mathcal{H}_{00,0} \oplus \mathcal{H}_{01,0} \oplus \mathcal{H}_{1,0} \oplus \mathcal{H}_{00,1} \oplus \mathcal{H}_{01,1} \oplus \mathcal{H}_{1,1}. \end{aligned}$$

Correspondingly, for any bivariate function in $f \in S^2(\mathbb{I}) \otimes \mathbb{R}^2$, we have the following decomposition,

$$f = f_{00,0} + f_{01,0} + f_{1,0} + f_{00,1} + f_{01,1} + f_{1,1}.$$

where $f_{01,1}$ and $f_{1,1}$ representing the linear and nonlinear interaction between $x^{\langle 1 \rangle}$ and $x^{\langle 2 \rangle}$ . First we test whether the linear interaction of $x^{\langle 1 \rangle}$ and $x^{\langle 2 \rangle}$ exsits as follows:

$$H_0 : f_{01,1} = 0 \quad v.s \quad H_1 : f_{01,1} \neq 0.$$

Since $f_{01,1}$ is the linear interaction of $x^{\langle 1 \rangle}$ and $x^{\langle 2 \rangle}$, it can be tested using the standard parametric test. If failing to reject the null hypothesis, we will continue to test the $f_{1,1}$ which is the nonlinear interaction between $x^{\langle 1 \rangle}$ and $x^{\langle 2 \rangle}$.

$$H_0 : f_{1,1} = 0 \quad v.s \quad H_1 : f_{1,1} \neq 0$$

Notice that, $f_{01,0}$ are linear in terms of $x^{\langle 1 \rangle}$. Applying the Theorem 1 in [28], we have

$$||\hat{f}_{01,0} - f_{01,0}||_2^2 = O(1/n),$$

which implies a faster convergence rate than the nonparametric term. Replacing $\mathbf{y}$ by $\mathbf{y}^* = \mathbf{y} - \hat{\mathbf{f}}_{01,0}$ and plugging to (1.14), we perform the proposed test by assuming

$$f^* \in \mathcal{H}_{00,0} \oplus \mathcal{H}_{1,0} \oplus \mathcal{H}_{00,1} \oplus \mathcal{H}_{1,1}.$$

where $f^* = f - \hat{f}_{01,0}$. Then the test statistics can be constructed using (1.14) by setting $M = R + \lambda_n I$ where $R = \theta_{00,1} K_{00,1} + \theta_{1,1} K_{1,1}$ is the weighted sum of kernels for $\mathcal{H}_{00,1}$ and $\mathcal{H}_{1,1}$ with positive weight $\theta_{00,1}$ and $\theta_{1,1}$.

## 1.7 Discussion

The hypothesis testing in SS-ANOVA is a very difficult problem. In this paper, we developed a "Wald-type" test for testing the significance of the interaction in two-way SS-ANOVA model. The optimality of the proposed test was justified by the minimax distinguishable rate. The extensive empirical studies suggests that proposed test has a superior performance. Even though we only discuss the test of the significance of the interaction in two-way SS-ANOVA model, the test of the significance of main effects can be developed. In higher order SS-ANOVA model, the test of significance of each term can be developed parallel to our development.

## 1.8 Technical Proofs

This section collects some detailed derivation, proofs of the lemmas and theorems. In the theoretical derivation, we only focus on single replicate case. If we have $w_j$ subjects for the $j$th group, it is equivalent to say that there are $w_j$ replicated measurements. The proof can be easily generalized to this situation by using the penalized weighted least squares (see [1] section 3.2.4).

### 1.8.1 Proof of Theorem 3.1

Before deriving the proof of Theorem 1.3.1, we need to derive several quantities and prove Lemma 1.8.1, Lemma 1.8.2 and Lemma 1.8.3.

#### Derivation of Equation (1.18)

Write matrix $R$ as

$$R = \theta_{01}K_{01} + \theta_{11}K_{11} = \frac{1}{2}\begin{bmatrix} K & \theta_d K \\ \theta_d K & K \end{bmatrix},$$

where $\theta_d = \theta_{01} - \theta_{11}$. The inverse of $M$ can be written as

$$\begin{aligned} M^{-1} &= \begin{bmatrix} \frac{1}{2}K + \lambda_n I_n & \frac{\theta_d}{2}K \\ \frac{\theta_d}{2}K & \frac{1}{2}K + \lambda_n I_n \end{bmatrix}^{-1} \triangleq \begin{bmatrix} A & B \\ B & A \end{bmatrix}^{-1} \\ &= \begin{bmatrix} A^{-1} + A^{-1}B(A - BA^{-1}B)^{-1}BA^{-1} & -A^{-1}B(A - BA^{-1}B)^{-1} \\ -A^{-1}B(A - BA^{-1}B)^{-1} & (A - BA^{-1}B)^{-1} \end{bmatrix}, \end{aligned}$$

where $A = \frac{1}{2}K + \lambda_n I_n$, $B = \frac{\theta_d}{2}K$, and $I_n$ denotes the $n \times n$ identity matrix. Note that $S$ is an $2n \times 2$ matrix defined in (1.12). We thus have

$$S^T M^{-1} S = \begin{bmatrix} a & b \\ b & c \end{bmatrix},$$

where

$$\begin{aligned} a &= \mathbf{1}^T A^{-1}\mathbf{1} + \mathbf{1}^T A^{-1}B(A - BA^{-1}B)^{-1}BA^{-1}\mathbf{1} + 2b - c, \\ b &= -\mathbf{1}^T BA^{-1}(A - BA^{-1}B)^{-1}\mathbf{1} + \mathbf{1}^T(A - BA^{-1}B)^{-1}\mathbf{1}, \\ c &= \mathbf{1}^T(A - BA^{-1}B)^{-1}\mathbf{1}. \end{aligned}$$

Consequently,

$$S(S^TM^{-1}S)^{-1}S^T = \frac{1}{ac-b^2}\begin{bmatrix} c\mathbf{1}\mathbf{1}^T & (c-b)\mathbf{1}\mathbf{1}^T \\ (c-b)\mathbf{1}\mathbf{1}^T & (a+c-2b)\mathbf{1}\mathbf{1}^T \end{bmatrix}$$
$$= \frac{1}{ac-b^2}\begin{bmatrix} c\mathbf{1}\mathbf{1}^T & (c-b)\mathbf{1}\mathbf{1}^T \\ (c-b)\mathbf{1}\mathbf{1}^T & c\mathbf{1}\mathbf{1}^T \end{bmatrix},$$

where the second equality is due to the fact $a - 2b = 0$ by Woodbury matrix identity.

Note that $\mathbf{f}_1 = (f_1(x_1^{\langle 1\rangle}), \ldots, f_1(x_n^{\langle 1\rangle}), f_1(x_1^{\langle 1\rangle}), \ldots, f_1(x_n^{\langle 1\rangle}))$. Let

$$(f_1(x_1^{\langle 1\rangle}), \ldots, f_1(x_n^{\langle 1\rangle}) \triangleq \mathbf{h}^T.$$

Therefore, we have

$$K_{11}M^{-1}(I_n - S(S^TM^{-1}S)^{-1}S^TM^{-1})\mathbf{f}_1$$
$$= \frac{1}{2}\begin{bmatrix} K & -K \\ -K & K \end{bmatrix}(M^{-1} - \frac{1}{ac-b^2}M^{-1}\begin{bmatrix} c\mathbf{1}\mathbf{1}^T & (c-b)\mathbf{1}\mathbf{1}^T \\ (c-b)\mathbf{1}\mathbf{1}^T & c\mathbf{1}\mathbf{1}^T \end{bmatrix}M^{-1})\begin{bmatrix} \mathbf{h} \\ \mathbf{h} \end{bmatrix}.$$

Since both $M^{-1}$ and $\frac{1}{ac-b^2}M^{-1}\begin{bmatrix} c\mathbf{1}\mathbf{1}^T & (c-b)\mathbf{1}\mathbf{1}^T \\ (c-b)\mathbf{1}\mathbf{1}^T & c\mathbf{1}\mathbf{1}^T \end{bmatrix}M^{-1})$ are symmetric matrices and their diagnonal entries are identical, we have

$$\begin{bmatrix} \mathbf{h}^* \\ \mathbf{h}^* \end{bmatrix} \triangleq (M^{-1} - \frac{1}{ac-b^2}M^{-1}\begin{bmatrix} c\mathbf{1}\mathbf{1}^T & (c-b)\mathbf{1}\mathbf{1}^T \\ (c-b)\mathbf{1}\mathbf{1}^T & c\mathbf{1}\mathbf{1}^T \end{bmatrix}M^{-1})\begin{bmatrix} \mathbf{h} \\ \mathbf{h} \end{bmatrix}.$$

Simple algebra yields (1.18). □

**Lemma 1.8.1.** *Under Assumption 1(a), for $\lambda_n > 1/n$, $m > 3/2$, and any $\epsilon > 0$, we have*

$$\hat{s}_{\lambda_n} \asymp s_{\lambda_n}$$

*with probability at least* $1 - (n^{\frac{2}{2m-1}-2\epsilon} + n^{\frac{1}{2m-1}})\exp\{-cn^{\frac{2m-3}{2m-1}+2\epsilon}\}$.

*Proof.* Under Assumption 1(a), $X_1^{\langle 1\rangle},\ldots, X_n^{\langle n\rangle}$ are i.i.d with distribution $\omega^{\langle 1\rangle}$. Therefore, by Theorem 3 in [29], for $1 \le i \le n$ and $i \le r \le n$, simple algebra yields

$$\mathbb{P}(|\hat{\mu}_i - \mu_i| \le c_m \mu_i + \mu_r + \Lambda_r) \ge 1 - r(r+1)\exp\{-\frac{nc_m^2}{2C^4 r^2}\},$$

where $\Lambda_r = \sum_{i=r+1}^{\infty} \mu_i$, $C$ is an absolute constant, and $c_m$ is a constant depending solely on $m$. Since the eigenvalue $\mu_i$ has the polynomial decay rate $i^{-2m}$, we have

$$\Lambda_r \asymp \sum_{i=r+1}^{\infty} i^{-2m}.$$

For $m > 1/2$,

$$\sum_{i=r+1}^{\infty} i^{-2m} \le \int_r^{\infty} x^{-2m} dx = \frac{r^{1-2m}}{2m-1} = \mathcal{O}(r^{1-2m}).$$

Let $r = n^{1/(2m-1)-\epsilon}$, we have $\Lambda_r + \mu_r = \mathcal{O}(n^{2\epsilon m - 1 - \epsilon}) = o(\mu_i)$ for $i = 1, \ldots, n^{1/2m-\epsilon}$. Next, we have, for any $i = 1, \ldots, n^{\frac{1}{2m}-\epsilon}$, the empirical eigenvalue $\hat{\mu}_i$ satisfies

$$|\hat{\mu}_i - \mu_i| \le c_m \mu_i$$

with probability at least

$$1 - (n^{\frac{2}{2m-1}-2\epsilon} + n^{\frac{1}{2m-1}})\exp\{-cn^{\frac{2m-3}{2m-1}+2\epsilon}\} \tag{1.25}$$

where $c = \frac{c_m^2}{2C^4}$, $c_m$ is a constant only related to $m$, and $M$ is an absolute constant. To ensure the probability in (1.25) goes to 1, we further require $m > 3/2$. Thus, we have, for $\lambda_n > 1/n$ and $m > 3/2$,

$$\hat{s}_{\lambda_n} \asymp s_{\lambda_n}$$

with probability at least $1 - (n^{\frac{2}{2m-1}-2\epsilon} + n^{\frac{1}{2m-1}})\exp\{-cn^{\frac{2m-3}{2m-1}+2\epsilon}\}$. □

**Lemma 1.8.2.** *Under Assumption 1(b), for $m > 1/2$, we have*

$$\hat{s}_{\lambda_n} \asymp s_{\lambda_n}$$

*for any $\lambda_n > 0$.*

*Proof.* The kernel function $\mathcal{K}^*$ can be explicitly written as,

$$\mathcal{K}^*(x,y) = 2\sum_{i=1}^{\infty} \frac{cos(2\pi k(x-y))}{(2\pi k)^{2m}}.$$

Under Assumption 1(b), we have the $X_1^{\langle 1\rangle}, \ldots, X_n^{\langle 1\rangle}$ evenly distributed on $[0,1]$. Without loss of generality, we assume that $X_1^{\langle 1\rangle} < \cdots < X_n^{\langle 1\rangle}$. Therefore, the kernel matrix $K$ is given by $[\mathcal{K}^*(x_i^{\langle 1\rangle}, x_{i'}^{\langle 1\rangle})]_{1\le i,i'\le n}$ which is a symmetric circulant matrix of order $n$ [16] with eigenvalues

$$\hat{\mu}_i^* = \begin{cases} \sum_{k=1}^{\infty} \frac{1}{[2\pi(kn-i)]^{2m}} + \sum_{k=0}^{\infty} \frac{1}{[2\pi(kn+i)]^{2m}} & \text{if } 1 \le i \le n-1 \\ 2\sum_{k=1}^{\infty} \frac{1}{(2\pi kn)^{2m}} & \text{if } i = n \end{cases}. \tag{1.26}$$

Note that $\hat{\mu}_i^*$ is a re-arrangement of $\hat{\mu}_i$. When $m > 1/2$, simple calculation yields

$$\begin{aligned} \frac{1}{[2\pi(n-i)]^{2m}} + \frac{1}{(2\pi i)^{2m}} + 2\underline{c}_m(2\pi n)^{-2m} \le \hat{\mu}_i^* \\ \le \frac{1}{[2\pi(n-i)]^{2m}} + \frac{1}{(2\pi i)^{2m}} + 2\bar{c}_m(2\pi n)^{-2m}, \end{aligned} \tag{1.27}$$

for $i = 1, \ldots, n-1$, and

$$\hat{\mu}_n^* = 2\bar{c}_m(2\pi n)^{-2m},$$

where $\underline{c}_m := \sum_{k=1}^{\infty} k^{-2m}$, and $\bar{c}_m = \sum_{k=2}^{\infty} k^{-2m}$. By (1.27), we have $\hat{\mu}_i^* \asymp \mu_i$ for $1 \le i \le \frac{n}{2}$ and $\hat{\mu}_i^* \asymp \mu_{n-i}$ for $\frac{n}{2} \le i \le n$. Since $\{\hat{\mu}\}_{i=1}^n$ are obtained by ordering $\{\hat{\mu}_i^*\}_{i=1}^n$ decreasingly, we have $\mu_i \asymp \hat{\mu}_i$, and consequently,

$$s_{\lambda_n} \asymp \hat{s}_{\lambda_n}.$$

for any $\lambda_n > 0$. □

**Lemma 1.8.3.** *If Assumptions 1 and 2 hold, for $\Delta = M^{-1}K_{11}^2M^{-1}$ defined in Theorem 1.3.1, we have*

$$\frac{4\hat{s}_{\lambda_n}}{9} \le Tr(\Delta) \le \frac{4}{(1-\theta_d)^2}(\hat{s}_{\lambda_n} + \frac{1}{2\lambda_n}\sum_{i=\hat{s}_{\lambda_n}+1}^{n} \hat{\mu}_i). \tag{1.28}$$

*Proof.* Note that the kernel matrix $K$ in (1.11) has the spectral decomposition $K = UDU^T$, where the eigenvector matrix $U$ is a $n \times n$ unitary matrix and the eigenvalue matrix $D =$

$Diag\{\hat{\mu}_i\}$ is a diagonal matrix with eigenvalues $\hat{\mu}_1 \geq \hat{\mu}_2 \geq \cdots \geq \hat{\mu}_n$. Correspondingly, we have the following decomposition,

$$K_{11} = \frac{1}{2}\begin{bmatrix} U & 0 \\ 0 & U \end{bmatrix}\begin{bmatrix} D & -D \\ -D & D \end{bmatrix}\begin{bmatrix} U^T & 0 \\ 0 & U^T \end{bmatrix},$$

$$M = \frac{1}{2}\begin{bmatrix} U & 0 \\ 0 & U \end{bmatrix}\begin{bmatrix} D + 2\lambda_n I_n & \theta_d D \\ \theta_d D & D + 2\lambda_n I_n \end{bmatrix}\begin{bmatrix} U^T & 0 \\ 0 & U^T \end{bmatrix},$$

where $I_n$ is the $n \times n$ identity matrix, and $\theta_d = \theta_{01} - \theta_{11}$. Letting $E = D + 2\lambda_n I_n = Diag\{\hat{\mu}_i + 2\lambda_n\}$ and $F = \theta_d D = Diag\{\theta_d \hat{\mu}_i\}$, we have

$$K_{11}M^{-1} = \begin{bmatrix} U & 0 \\ 0 & U \end{bmatrix}\begin{bmatrix} D & -D \\ -D & D \end{bmatrix}\begin{bmatrix} E & F \\ F & E \end{bmatrix}^{-1}\begin{bmatrix} U^T & 0 \\ 0 & U^T \end{bmatrix}.$$

Using the inverse of block matrix, we have

$$\begin{bmatrix} D & -D \\ -D & D \end{bmatrix}\begin{bmatrix} E & F \\ F & E \end{bmatrix}^{-1} = \begin{bmatrix} V_{11} & V_{12} \\ V_{21} & V_{22} \end{bmatrix}$$

where

$$\begin{aligned} V_{11} &= DE^{-1} + (D + DE^{-1}F)(E - FE^{-1}F)^{-1}FE^{-1}, \\ V_{12} &= -(DE^{-1}F + D)(E - FE^{-1}F)^{-1}, \\ V_{21} &= -V_{12}, \qquad (1.29) \\ V_{22} &= -V_{11}. \qquad (1.30) \end{aligned}$$

Consequently,

$$\Delta = M^{-1}K_{11}^2M^{-1} = \begin{bmatrix} U & 0 \\ 0 & U \end{bmatrix}\begin{bmatrix} V_{11} & V_{12} \\ V_{21} & V_{22} \end{bmatrix}^T\begin{bmatrix} V_{11} & V_{12} \\ V_{21} & V_{22} \end{bmatrix}\begin{bmatrix} U^T & 0 \\ 0 & U^T \end{bmatrix}.$$

We thus have

$$Tr(\Delta) = Tr(M^{-1}K_{11}^2M^{-1}) = Tr(\begin{bmatrix} V_{11} & V_{12} \\ V_{21} & V_{22} \end{bmatrix}^T \begin{bmatrix} V_{11} & V_{12} \\ V_{21} & V_{22} \end{bmatrix})$$
$$= Tr \begin{bmatrix} V_{11}^TV_{11} + V_{21}^TV_{21} & V_{11}^TV_{12} + V_{21}^TV_{22} \\ V_{12}^TV_{11} + V_{22}^TV_{21} & V_{12}^TV_{12} + V_{22}^TV_{22} \end{bmatrix}. \quad (1.31)$$

By (1.29) and (1.30), we have

$$V_{11}^TV_{11} + V_{21}^TV_{21} = V_{12}^TV_{12} + V_{22}^TV_{22}.$$

Simple algebra yields

$$V_{11}^TV_{11} + V_{21}^TV_{21} = 2V_{11}^TV_{11}.$$

Therefore, we have

$$Tr(\Delta) = 4Tr(V_{11}^TV_{11}). \quad (1.32)$$

Notice that $D$, $E$, $F$ are diagonal matrices, we have

$$Tr(\Delta) = 4Tr(V_{11}^TV_{11}) \geq 4Tr(D^2E^{-2}).$$

Since

$$D^2E^{-2} = Diag\{\frac{\hat{\mu}_i^2}{(\hat{\mu}_i + 2\lambda_n)^2}\},$$

we have

$$Tr(\Delta) \geq 4\sum_{i=1}^{n} \frac{\hat{\mu}_i^2}{(\hat{\mu}_i + 2\lambda_n)^2} \geq 4\sum_{i=1}^{\hat{s}_{\lambda_n}} \frac{\hat{\mu}_i^2}{(\hat{\mu}_i + 2\lambda_n)^2}, \quad (1.33)$$

where $\hat{s}_{\lambda_n}$ is the effective dimension for kernel matrix $K_{11}$. By the definition of $\hat{s}_{\lambda_n}$ in (1.22), for the any $i < \hat{s}_{\lambda_n}$, we have $\frac{\hat{\mu}_i}{\hat{\mu}_i + 2\lambda_n} > \frac{1}{3}$. Thus we have

$$Tr(\Delta) \geq \frac{4}{9}\hat{s}_{\lambda_n}.$$

Now we shall prove the upper bound for $Tr(\Delta)$. Since $Tr(\Delta)$ has the expression (1.32), we expand $V_{11}$ as

$$V_{11} = DE^{-1} + DE^{-1}(F(E - FE^{-1}F)^{-1}FE^{-1} + (E - FE^{-1})^{-1}F).$$

The $i$th diagonal entry of $F(E-FE^{-1}F)^{-1}FE^{-1}$ is

$$Diag_i(F(E-FE^{-1}F)^{-1}FE^{-1}) = \frac{\theta_d^2\hat{\mu}_i^2}{(\hat{\mu}_i+2\lambda_n-\frac{\theta_d^2\hat{\mu}_i^2}{\hat{\mu}_i+2\lambda_n})(\hat{\mu}_i+2\lambda_n)}$$
$$\leq \frac{\theta_d^2}{1-\theta_d^2}, \tag{1.34}$$

and the $i$th diagonal entry of $(E-FE^{-1})^{-1}F$ is

$$Diag_i((E-FE^{-1})^{-1}F) = \frac{\theta_d\hat{\mu}}{\hat{\mu}_i+2\lambda_n-\frac{\theta_d^2\hat{\mu}_i^2}{\hat{\mu}_i+2\lambda_n}} \leq \frac{\theta_d}{1-\theta_d^2}. \tag{1.35}$$

Combining (1.34) and (1.35), we have the $i$th diagonal entry of $V_{11}$

$$Diag_i(V_{11}) \leq (1+\frac{\theta_d^2}{1-\theta_d^2}+\frac{\theta_d}{1-\theta_d^2}Diag_i(DE^{-1}) = \frac{1}{1-\theta_d}Diag_i(DE^{-1}).$$

Since the lower diagonal block is identical to the upper diagonal block, we only need to bound the trace of $DE^{-1}$. We have

$$\begin{aligned}
Tr(D^2E^{-2}) =& \sum_{i=1}^{\hat{s}_{\lambda_n}} \frac{\hat{\mu}_i^2}{(\hat{\mu}_i+2\lambda_n)^2} + \sum_{i=\hat{s}_{\lambda_n}+1}^{n} \frac{\hat{\mu}_i^2}{(\hat{\mu}_i+2\lambda_n)^2} \\
\leq& \sum_{i=1}^{\hat{s}_{\lambda_n}} \frac{\hat{\mu}_i}{\hat{\mu}_i+2\lambda_n} + \sum_{i=\hat{s}_{\lambda_n}+1}^{n} \frac{\hat{\mu}_i}{\hat{\mu}_i+2\lambda_n} \\
\leq& \hat{s}_{\lambda_n} + \frac{1}{2\lambda_n}\sum_{i=\hat{s}_{\lambda_n}+1}^{n}\hat{\mu}_i.
\end{aligned}$$

Thus we have $Tr(\Delta) \leq \frac{4}{(1-\theta_d)^2}(\hat{s}_{\lambda_n}+\frac{1}{2\lambda_n}\sum_{i=\hat{s}_{\lambda_n}+1}^{n}\hat{\mu}_i)$. □

Proof of Theorem 1.3.1

*Proof.* We define the three terms on the right-hand side of equation (1.20) as $T_1$, $T_2$ and $T_3$, i.e.,

$$\begin{aligned}
T_1 &= \frac{1}{n}\boldsymbol{\epsilon}^T\Delta\boldsymbol{\epsilon}, \\
T_2 &= \frac{1}{n}\boldsymbol{\epsilon}^T M^{-1}S(S^TM^{-1}S)^{-1}S^T\Delta S(S^TM^{-1}S)^{-1}S^TM^{-1}\boldsymbol{\epsilon}, \\
T_3 &= \frac{1}{n}\boldsymbol{\epsilon}^T M^{-1}S(S^TM^{-1}S)^{-1}S^T\Delta\boldsymbol{\epsilon}.
\end{aligned}$$

We now show $T_2$ and $T_3$ are in smaller order compared to $T_1$. First, we analyze the second term $T_2$ in (1.20). We have

$$\begin{aligned}
\mathbb{E}[T_2] =& \frac{1}{n}\mathbb{E}[\boldsymbol{\epsilon}^T M^{-1}S(S^T M^{-1}S)^{-1}S^T\Delta S(S^T M^{-1}S)^{-1}S^T M^{-1}\boldsymbol{\epsilon}] \\
=& \frac{\sigma^2}{n}Tr(M^{-1}S(S^T M^{-1}S)^{-1}S^T\Delta S(S^T M^{-1}S)^{-1}S^T M^{-1}) \\
\leq& \frac{2\sigma^2}{n}\lambda_{\max}(\Delta)\lambda_{\max}(M^{-1}S(S^T M^{-1}S)^{-1}S^T S(S^T M^{-1}S)^{-1}S^T M^{-1}) \\
\leq& \frac{2\sigma^2}{n}\lambda_{\max}(\Delta),
\end{aligned}$$

where $\lambda_{\max}$ denotes the largest eigenvalue. Since all eigenvalues of $\Delta$ are less than 1, we have $\mathbb{E}[T_2] \leq \frac{2\sigma^2}{n}$. Analogously, we can derive the variance inequality of $T_2$. Combining the results together and using the Chebyshev inequality, we have

$$T_2 = \mathcal{O}_p(\frac{1}{n}). \tag{1.36}$$

Second, we analyze the third term $T_3$ in (1.20). We apply the Cauchy-Schwarz inequality and have

$$|T_3| \leq \sqrt{T_2}\sqrt{T_1}. \tag{1.37}$$

Finally, we derive the magnitude of $T_1$. We first consider the testing consistency of $T_1$ conditional on $X$. Denote $\mathbb{E}_\epsilon$ as the expectation with respect to $\epsilon$, and define $\mathbb{V}ar_\epsilon$ as the variance with respect to $\epsilon$. Note that

$$\mathbb{E}_\epsilon[\boldsymbol{\epsilon}^T\Delta\boldsymbol{\epsilon}] = \sigma^2 Tr(\Delta), \quad \mathbb{V}ar_\epsilon[\boldsymbol{\epsilon}^T\Delta\boldsymbol{\epsilon}] = 2\sigma^4 Tr(\Delta^2).$$

Let $Z = (\boldsymbol{\epsilon}^T\Delta\boldsymbol{\epsilon} - \sigma^2 Tr(\Delta))/\sigma^2\sqrt{2Tr(\Delta^2)}$ and $t \in (-1/2, 1/2)$. Then the log-characteristic function of $Z$ can be written as

$$\begin{aligned}
&\log \mathbb{E}_\epsilon[\exp(itZ)] \\
=& \log \mathbb{E}_\epsilon[\exp(it\boldsymbol{\epsilon}^T\Delta\boldsymbol{\epsilon}/\sigma^2\sqrt{2Tr(\Delta^2)})] - itTr(\Delta)/\sigma^2\sqrt{2Tr(\Delta^2)} \\
=& -\frac{1}{2}\log\det\{I_{2n} - 2it\Delta/\sigma^2\sqrt{2Tr(\Delta^2)}\} - itTr(\Delta)/\sigma^2\sqrt{2Tr(\Delta^2)}.
\end{aligned} \tag{1.38}$$

Through Taylor expansion, one has

$$-\frac{1}{2}\log\det\{I_{2n}-2it\Delta/\sigma^2\sqrt{2Tr(\Delta^2)}\}$$
$$=it\frac{Tr(\Delta)}{\sigma^2\sqrt{2Tr(\Delta^2)}}-t^2\frac{Tr(\Delta^2)}{2\sigma^4Tr(\Delta^2)}+\mathcal{O}(t^3\frac{Tr(\Delta^3)}{\sigma^6[Tr(\Delta^2)]^{3/2}}). \quad (1.39)$$

Combining Equations (1.38) and (1.39), we have

$$\log\mathbb{E}_\epsilon[\exp(itZ)]=-\frac{t^2}{2\sigma^4}+\mathcal{O}(t^3\frac{Tr(\Delta^3)}{\sigma^6[Tr(\Delta^2)]^{3/2}}). \quad (1.40)$$

Since all eigenvalues of $\Delta$ are less than 1, we have $\frac{Tr(\Delta^3)}{Tr(\Delta^2)}\leq 1$. Analogous to (1.33), we have

$$Tr(\Delta^2)\geq\frac{16}{81}\hat{s}_{\lambda_n}. \quad (1.41)$$

Under Assumption 1(a), we have $Tr(\Delta^2)\to\infty$ as $\lambda_n\to 0$ with probability approaching 1 by Lemma 1.8.1 and (1.41). Hence, the second term on the right-hand side of Equation (1.40) is $o_p(1)$. We thus conclude that

$$\mathbb{E}_\epsilon[\exp(itZ)]\xrightarrow{P}\exp(-\frac{t^2}{2\sigma^4}).$$

Next, we show that

$$\mathbb{E}[\exp(itZ)]=\mathbb{E}_X\big[\mathbb{E}_\epsilon[\exp(itZ)]\big]\to\exp(-t^2/(2\sigma^4))$$

for $t\in(-\frac{1}{2},\frac{1}{2})$. If not, there exists a subsequence of r.v $X_{nk}^{\langle 1\rangle}$, such that for $\forall\varepsilon>0$, $|\mathbb{E}_{X_{nk}^{\langle 1\rangle}}\mathbb{E}_\epsilon\exp(itZ)-\exp(-t^2/(2\sigma^4))|>\varepsilon$. On the other hand, since $\mathbb{E}_\epsilon\exp(itZ(X_{nk}^{\langle 1\rangle}))\xrightarrow{P}\exp(-t^2/(2\sigma^4))$, which is bounded, there exists a sub-sub sequence $\{X_{nkl}^{\langle 1\rangle}\}$, such that $\mathbb{E}_\epsilon\exp(itZ(X_{nkl}^{\langle 1\rangle}))\xrightarrow{a.s}\exp(-t^2/(2\sigma^4))$. Then by dominate convergence theorem, $\mathbb{E}_{X_{nkl}^{\langle 1\rangle}}\mathbb{E}_\epsilon\exp(itZ)\to\exp(-t^2/(2\sigma^4))$, which is a contradiction.

Under Assumption 1(b), we can easily obtain $\mathbb{E}[\exp(itZ)]\to\exp(-\frac{t^2}{2\sigma^4})$ by Lemma 1.8.2 and (1.41).

Thus $Z$ is asymptotically Gaussian distributed, and

$$\frac{T_1-\sigma^2Tr(\Delta)/n}{\sigma^2\sqrt{2Tr(\Delta^2)/n^2}}\xrightarrow{d}N(0,1). \quad (1.42)$$

Combining (1.36), (1.37) and (1.42), the theorem follows. □

### 1.8.2 Proof of Theorem 3.2

In SS-ANOVA (1.6), we denote $g^* = f_1 + f_{1,2}$. We shall now estimate $g^*$ using the noise-free data. That is, we consider the following penalized least squares,

$$\frac{1}{2n}||\mathbf{f} - S\mathbf{d} - nR\mathbf{c}||_2^2 + n\lambda_n \mathbf{c}^T R\mathbf{c} \tag{1.43}$$

where $\mathbf{f} = (f(x_1^{\langle 1\rangle}, x_1^{\langle 2\rangle}), \ldots, f(x_n^{\langle 1\rangle}, x_n^{\langle 2\rangle}))^T$. The minimizer of (1.43) is seen to be

$$\begin{aligned}\tilde{\mathbf{c}} &= \frac{1}{n}[M^{-1} - M^{-1}S(S^T M^{-1} S)^{-1} S^T M^{-1}]\mathbf{f},\\ \tilde{\mathbf{d}} &= (S^T M^{-1} S)^{-1} S^T M^{-1}\mathbf{f}.\end{aligned}$$

The estimator of $g^*$ is $\tilde{g}^* = \xi^T \tilde{\mathbf{c}}$, where the kernel vector $\xi$ has the $i$th entry of $\theta_{01}\mathcal{K}_{01}(x_i, \cdot) + \theta_{11}\mathcal{K}_{11}(x_i, \cdot)$. We thus have

$$\tilde{\mathbf{g}}^* = R[M^{-1} - M^{-1}S(S^T M^{-1} S)^{-1} S^T M^{-1}]\mathbf{f},$$

where $\tilde{\mathbf{g}}^* = (\tilde{g}^*(x_1), \ldots, \tilde{g}^*(x_1))^T$ and $\mathbf{g}^* = (g^*(x_1), \ldots, g^*(x_1))^T$. By (1.17), we have

$$\tilde{\mathbf{g}}^* = R[M^{-1} - M^{-1}S(S^T M^{-1} S)^{-1} S^T M^{-1}]\mathbf{g}^*, \tag{1.44}$$

**Lemma 1.8.4.** *If Assumption 2 holds, as $n \to \infty$, $\lambda_n \to 0$ and $\lambda_n \geq n^{-1}$, we have,*

$$||\tilde{g}^* - g^*||_n^2 \leq c\lambda_n,$$

*where $c$ is a constant.*

*Proof.* By Assumption 2, we have

$$||f_1 + f_{1,2}||_{\mathcal{H}_{10}\oplus\mathcal{H}_{11}}^2 \leq ||f||_{\mathcal{H}}^2 < 1. \tag{1.45}$$

For any function $g$ in $\mathcal{H}_{10}\oplus\mathcal{H}_{11}$, we can represent it as $g = \xi^T\tilde{\mathbf{c}} + \zeta(\cdot)$, where $\zeta(\cdot) \in \mathcal{H}_{10}\oplus\mathcal{H}_{11}$ is orthogonal to $\xi$. Moreover,

$$\begin{aligned}||f_1 + f_{1,2}||_{\mathcal{H}_{10}\oplus\mathcal{H}_{11}}^2 =& ||\xi^T\tilde{\mathbf{c}}||_{\mathcal{H}_{10}\oplus\mathcal{H}_{11}}^2 + ||\zeta(\cdot)||_{\mathcal{H}_{10}\oplus\mathcal{H}_{11}}^2\\ \geq& n\tilde{\mathbf{c}}^T R\tilde{\mathbf{c}} = \frac{1}{n}(n\tilde{\mathbf{c}}^T R)R^{-1}(nR\tilde{\mathbf{c}})\\ =& \frac{1}{n}\mathbf{g}^{*T} R^{-1}\mathbf{g}^*.\end{aligned} \tag{1.46}$$

Combining (1.45) and (1.46), we have

$$\frac{1}{n}\mathbf{g}^{*T}R^{-1}\mathbf{g}^{*} < 1. \tag{1.47}$$

By (1.44), we have

$$\begin{aligned}||\tilde{g}^{*} - g^{*}||_n^2 &= \frac{1}{n}||\mathbf{g}^{*} - RM^{-1}\mathbf{g}^{*} + RM^{-1}S(S^TM^{-1}S)^{-1}S^TM^{-1}\mathbf{g}^{*}||_2^2 \\ &= \frac{1}{n}\mathbf{g}^{*T}(I - RM^{-1})^2\mathbf{g}^{*} + \frac{1}{n}||RM^{-1}S(S^TM^{-1}S^T)^{-1}S^TM^{-1}\mathbf{g}^{*}||_2^2.\end{aligned}$$

Noting that $M = R + \lambda_n I_n$, the eigenvalues of $I_n - R(R + \lambda_n I_n)^{-1}$ are all smaller than 1, and the rank of $RM^{-1}S(S^TM^{-1}S^T)^{-1}S^TM^{-1}$ is 2, we have

$$\begin{aligned}||\tilde{g}^{*} - g^{*}||_n^2 \leq& \frac{1}{n}\mathbf{g}^{*}(I - R(R + \lambda_n I)^{-1})\mathbf{g}^{*} + \mathcal{O}(\frac{1}{n}) \\ \leq& \lambda_n + \mathcal{O}(\frac{1}{n}),\end{aligned}$$

where the last inequality holds by applying Woodbury matrix identity,

$$(R + \lambda_n I_n)^{-1} = R^{-1} - R^{-1}(\frac{1}{\lambda_n}I_n + R^{-1})^{-1}R^{-1} \geq R^{-1} - \lambda_n R^{-2},$$

and (1.47). The proof is thus completed. □

Remark for empirical norm and $L_2$ norm

In this section, we discuss the relationship between the empirical norm and $L_2$ norm under Assumption 1. Recall the definition of empirical norm and $L_2$ norm as,

$$||f||_n^2 = \frac{1}{2n}\sum_{j=1}^{2}\sum_{i=1}^{n} f^2(x_i^{\langle 1\rangle}, x_j^{\langle 2\rangle}) \quad \text{and} \quad ||f||_2^2 = \sum_{x^{\langle 2\rangle}=0}^{1}\int_0^1 f^2(x^{\langle 1\rangle}, x^{\langle 2\rangle})d\omega_1.$$

In the following lemma, we will establish their relationship.

**Lemma 1.8.5.** *Under Assumption 1, for $f : \mathcal{X}_1 \times \mathcal{X}_2 \to \mathbb{R}$ and a positive constant $c$, we have*

$$||f||_2 \leq c||f||_n$$

*i.e. the empirical norm of $f$ dominates the $L_2$ norm.*

*Proof.* Under Assumption 1(a), Theorem 3.1 of Eggermont et al. [17](page 384, [17]) implies that $||\cdot||_{\omega^{\langle 1\rangle}mh}$ norm is equivalent to $||\cdot||_n$ for any fixed $x^{\langle 2\rangle}$. The $||\cdot||_{\omega^{\langle 1\rangle}hm}$ is defined as $||f(x^{\langle 1\rangle},0)||^2_{\omega^{\langle 1\rangle}mh} = ||f(x^{\langle 1\rangle},0)||^2_{L_2(\omega^{\langle 1\rangle})} + h^{2m}||f(x^{\langle 1\rangle},0)^{(m)}||^2$ which trivially dominates the $||\cdot||_{L_2(\omega^{\langle 1\rangle})}$ norm. Since $x^{\langle 2\rangle}$ can only take the value of 0 or 1, we have that $||\cdot||_n$ dominate $||\cdot||_2$, i.e. there exists a positive constant $c$ such that $||f||_2 \leq c||f||_n$.

Under assumption 1(b), Lemma 2.27 in Eggermont et al. [17] states that the $||\cdot||_n$ dominates $||\cdot||_2$ for $x^{\langle 1\rangle}$ satisfying Assumption 1(b).

□

Poof of Theorem 1.3.2

*Proof.* Under the alternative hypothesis, the statistic $T_{n,\lambda_n}$ in (1.19) can be decomposed into three terms,

$$T_{n,\lambda_n} = \frac{1}{n}||H\boldsymbol{\epsilon}||_2^2 + \frac{1}{n}||H\mathbf{f}_{1,2}||_2^2 + \frac{2}{n}\mathbf{f}_{1,2}^T H^T H\boldsymbol{\epsilon}. \tag{1.48}$$

where $H = \theta_{11}K_{11}M^{-1}(I - S(S^TM^{-1}S)^{-1}S^TM^{-1})$. Let $W_1 = \frac{1}{n}||H\boldsymbol{\epsilon}||_2^2$, $W_2 = \frac{1}{n}||H\mathbf{f}_{1,2}||_2^2$, and $W_3 = \frac{2}{n}\mathbf{f}_{1,2}^T H^T H\boldsymbol{\epsilon}$ denote corresponding three terms on the right-hand side of equation (1.48).

We now derive a lower bound for $W_2$. By Lemma 1.8.4, we have

$$\begin{aligned}\frac{1}{n}||H\mathbf{f}_{1,2} - \mathbf{f}_{1,2}||_2^2 &\leq \frac{1}{n}||H\mathbf{f}_{1,2} - \mathbf{f}_{1,2}||_2^2 + \frac{1}{n}||H\mathbf{f}_1 - \mathbf{f}_1||_2^2 \\ &= \frac{1}{n}||H\mathbf{f}_1 + H\mathbf{f}_{1,2} - \mathbf{f}_1 - \mathbf{f}_{1,2}||_2^2 \\ &= ||\tilde{g}^* - g^*||_n^2 \leq c\lambda_n. \end{aligned} \tag{1.49}$$

Let $c' = \sqrt{c}$, we consider the distinguishable rate

$$\frac{1}{n}||\mathbf{f}_{1,2}||_2^2 = ||f_{1,2}||_n^2 > c'^2 d_n^2 = c(\lambda_n + \sigma_{n,\lambda_n}). \tag{1.50}$$

where the inequality is satisfied since $||\cdot||_n$ dominates $||\cdot||_2$ by Lemma 1.8.5. The lower bound of $W_2$ is thus,

$$\begin{aligned} W_2 = \frac{1}{n}||H\mathbf{f}_{1,2}||_2^2 =& \frac{1}{n}||\mathbf{f}_{1,2}||_2^2 - \frac{1}{n}||\mathbf{f}_{1,2} - H\mathbf{f}_{1,2}||_2^2 \\ \geq& cd_n^2 - c\lambda_n \\ \geq& c\sigma_{n,\lambda_n}. \end{aligned} \tag{1.51}$$

where the first inequality is obtained by (1.49) and the second inequality is obtained through pluging in (1.50)

For the third term $W_3$, it is seen that $\mathbb{E}W_3 = 0$. It is easy to verify that the eigenvalues of $HH^T$ are all less than 1. Moreover,

$$\begin{aligned} \mathbb{E}W_3^2 &= \frac{4}{n^2}\mathbb{E}[\mathbf{f}_{1,2}^T H^T H\boldsymbol{\epsilon}\boldsymbol{\epsilon}^T H^T H\mathbf{f}_{1,2}] \\ &= \frac{4}{n^2}(H\mathbf{f}_{1,2})^T HH^T(H\mathbf{f}_{1,2}) \\ &\leq \frac{4}{n^2}(H\mathbf{f}_{1,2})^T(H\mathbf{f}_{1,2}) = \frac{4}{n}W_2. \end{aligned}$$

By Chebyshev's inequality, for any $\epsilon > 0$, we have

$$\mathbb{P}(|W_3| \geq \frac{2\epsilon^{-\frac{1}{2}}W_2^{\frac{1}{2}}}{\sqrt{n}}) \leq \frac{n\mathbb{E}W_3^2}{4\epsilon^{-1}W_2} \leq \epsilon.$$

Consequently, there exists an $N_1$, for any $n > N_1$, we have

$$\mathbb{P}\{|W_3| > \frac{1}{2}W_2\} \leq \mathbb{P}(|W_3| \geq \frac{2\epsilon^{-\frac{1}{2}}W_2^{\frac{1}{2}}}{\sqrt{n}}) \leq \epsilon. \tag{1.52}$$

By the triangle inequality, we have

$$\begin{aligned} |\frac{W_1 - \mu_{n,\lambda_n}}{\sigma_{n,\lambda_n}} + \frac{W_2 + W_3}{\sigma_{n,\lambda_n}}| &\geq |\frac{W_2 + W_3}{\sigma_{n,\lambda_n}}| - |\frac{W_1 - \mu_{n,\lambda_n}}{\sigma_{n,\lambda_n}}| \\ &\geq |\frac{W_2}{\sigma_{n,\lambda_n}}| - |\frac{W_3}{\sigma_{n,\lambda_n}}| - |\frac{W_1 - \mu_{n,\lambda_n}}{\sigma_{n,\lambda_n}}|. \end{aligned} \tag{1.53}$$

If $\frac{|W_1 - \mu_{n,\lambda_n}|}{\sigma_{n,\lambda_n}} \leq C_\epsilon$, and $|W_3| \leq \frac{1}{2}W_2$ hold, in view of (1.53) and (1.51), we have

$$|\frac{W_1 - \mu_{n,\lambda_n}}{\sigma_{n,\lambda_n}} + \frac{W_2 + W_3}{\sigma_{n,\lambda_n}}| \geq \frac{1}{2}c - C_\epsilon.$$

Noting that $W_1$ is identical to (1.20), by Theorem 1.3.1, we have $\frac{|W_1-\mu_{n,\lambda_n}|}{\sigma_{n,\lambda_n}} = \mathcal{O}_p(1)$. That is for a $C_\epsilon > 0$, there exists an $N_2$, for any $n > N_2$, we have

$$\mathbb{P}(\frac{|W_1 - \mu_{n,\lambda_n}|}{\sigma_{n,\lambda_n}} > C_\epsilon) \leq \epsilon. \tag{1.54}$$

Setting $c \geq 2(C_\epsilon + z_{1-\frac{\alpha}{2}})$ and $N = \max(N_1, N_2)$, for any $n > N$, we have

$$\begin{aligned}
\mathbb{P}(\Phi_{n,\lambda_n} = 1) =& \mathbb{P}\{\frac{|W_1 + W_2 + W_3 - \mu_{n,\lambda_n}|}{\sigma_{n,\lambda_n}} \geq z_{1-\frac{\alpha}{2}}\} \\
\geq& \mathbb{P}\{\frac{|W_1 - \mu_{n,\lambda_n}|}{\sigma_{n,\lambda_n}} \leq C_\epsilon, |W_3| \leq \frac{1}{2}W_2\} \\
\geq& 1 - \mathbb{P}\{\frac{|W_1 - \mu_{n,\lambda_n}|}{\sigma_{n,\lambda_n}} > C_\epsilon\} - \mathbb{P}\{|W_3| > \frac{1}{2}W_2\} \\
\geq& 1 - 2\epsilon,
\end{aligned}$$

where the second inequality is due to Boole's inequality and the last inequality is obtained through plugging (1.51) and (1.54). The proof is completed.

□

### 1.8.3 Proof of Theorem 3.3

In order to find the optimal distinguishable rate, we need to bound the tail sum of the eigenvalues of the empirical kernel matrix. We state the following two lemma which gives an upper bound for the tail sum of the eigenvalues of the empirical kernel matrix under Assumption 1(a) and 1(b) respectively.

**Lemma 1.8.6.** *(Liu et al. [21]) If $1/n < \lambda_n \to 0$ and Assumption 1(a) is satisfied, then with probability at least $1 - 4e^{-s_{\lambda_n}}$,*

$$\sum_{i=\hat{s}_{\lambda_n}+1}^{n} \hat{\mu}_i \leq C s_{\lambda_n} \mu_{s_{\lambda_n}},$$

*where $C > 0$ is an absolute constant.*

**Lemma 1.8.7.** *If $\lambda_n > 0$ and Assumption 1(b) is satisfied, we have*

$$\sum_{i=\hat{s}_{\lambda_n}+1}^{n} \hat{\mu}_i \leq C s_{\lambda_n} \mu_{s_{\lambda_n}},$$

*where $C > 0$ is an absolute constant.*

*Proof.* Under Assumption 1(b), the empirical eigenvalues could be calculated by (1.26). By the definition of $\hat{s}_{\lambda_n}$ in (1.22), we have

$$\sum_{i=\hat{s}_{\lambda_n}+1}^{n} \hat{\mu}_i = \sum_{\{i|\hat{\mu}_i^* < \lambda_n\}} \hat{\mu}_i^*. \tag{1.55}$$

Since the population eigenvalues are $\{(2\pi i)^{-2m}\}_{i=1}^{\infty}$, we calculate the population efficient dimension as $s_{\lambda_n} = (\lambda_n)^{-1/2m}/2\pi$. Through the inequalities in (1.27), we have $\hat{\mu}_i^* \geq \lambda_n$ for $i = 1, \ldots, s_{\lambda_n}$ or $i = n - s_{\lambda_n}, \ldots, n$. We can bound the term in (1.55)

$$\sum_{\{i|\hat{\mu}_i^* < \lambda_n\}} \hat{\mu}_i^* \leq \sum_{i=s_{\lambda_n}}^{n-s_{\lambda_n}} \hat{\mu}_i^*$$

By the upper bound of $\hat{\mu}_i^*$ given in (1.27), we have

$$\sum_{i=s_{\lambda_n}}^{n-s_{\lambda_n}} \hat{\mu}_i^* \leq C s_{\lambda_n} \mu_{s_{\lambda_n}}$$

which completes the proof. □

Proof of Theorem 1.3.3

*Proof.* The distinguishable rate is

$$d_n = \sqrt{\lambda_n + \sigma_{n,\lambda_n}},$$

where $\sigma_{n,\lambda_n}^2 = 2\theta_{11}^4 \sigma^4 Tr(\Delta^2)/n^2$. We now derive the order of $\sigma_{n,\lambda_n}^2$. Since the eigenvalues of $\Delta$ are less than 1, and by Lemma 1.8.3, we have

$$Tr(\Delta^2) \leq Tr(\Delta) \leq \frac{4}{(1-\theta_d)^2}(\hat{s}_{\lambda_n} + \frac{1}{2\lambda_n} \sum_{i=\hat{s}_{\lambda_n}+1}^{n} \hat{\mu}_i).$$

Under Assumption 1(a), we apply the results in Lemma A.1 and have

$$Tr(\Delta^2) \lesssim \frac{4}{(1-\theta_d)^2}(\hat{s}_{\lambda_n} + \frac{1}{2\lambda_n}\lambda_n s_{\lambda_n})$$

satisfied with probability at least $1 - 4e^{-s_{\lambda_n}}$. Combining with the lower bound of $Tr(\Delta^2)$ in (1.41) and Lemma 1.8.1, we have

$$Tr(\Delta^2) = \mathcal{O}(s_{\lambda_n}). \tag{1.56}$$

with probability at least $1 - 4e^{-s_\lambda} - (n^{\frac{2}{2m-1}-2\epsilon} + n^{\frac{1}{2m-1}}) \exp\{-cn^{\frac{2m-3}{2m-1}+2\epsilon}\}$. Similarly, we have (1.56) satisfied under Assumption 1(b) by applying the results in Lemma 1.8.2 and Lemma 1.8.7.

By using (1.56), we have

$$\sigma^2_{n,\lambda_n} \asymp \lambda_n^{-\frac{1}{2m}} n^{-2} \asymp s_{\lambda_n} n^{-2}. \tag{1.57}$$

By Cauchy-Schwartz inequality, the distinguishable rate $d_n = \sqrt{\lambda_n + \sigma_{n,\lambda_n}}$ is minimized when $\lambda_n = \sigma_{n,\lambda_n}$, i.e.,

$$\lambda_n \asymp n^{-4m/(4m+1)}$$

Thus we have the minimum distinguishable rate

$$d_n^* = \mathcal{O}(n^{-2m/(4m+1)}),$$

By Lemma 1.8.5, this optimal distinguishable rate is achieved in the sense of $L_2$ norm. □

## Reference


[1] Chong Gu. *Smoothing spline ANOVA models*. Springer, 2013.

[2] Grace Wahba. *Spline models for observational data*. Siam, 1990.

[3] Trevor Hastie and Robert Tibshirani. *Generalized additive models*. Wiley Online Library, 1990.

[4] Dirk Stach, Oliver J Schmitz, Stephan Stilgenbauer, Axel Benner, Hartmut DoÈhner, Manfred Wiessler, and Frank Lyko. Capillary electrophoretic analysis of genomic dna methylation levels. *Nucleic Acids Research*, 31(2):e2–e2, 2003.

[5] Katharina Filarsky, Angela Garding, Natalia Becker, Christine Wolf, Manuela Zucknick, Rainer Claus, Dieter Weichenhan, Christoph Plass, Hartmut Döhner, Stephan Stilgenbauer, Peter Lichter, and Daniel Mertens. Krüppel-like factor 4 (klf4) inactivation in chronic lymphocytic leukemia correlates with promoter dna-methylation and can be reversed by inhibition of notch signaling. *haematologica*, 101(6):e249, 2016.

[6] Dennis Cox, Eunmee Koh, Grace Wahba, and Brian S Yandell. Testing the (parametric) null model hypothesis in (semiparametric) partial and generalized spline models. *The Annals of Statistics*, pages 113–119, 1988.

[7] Adelchi Azzalini and Adrian Bowman. On the use of nonparametric regression for checking linear relationships. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 549–557, 1993.

[8] Dong Xiang and Grace Wahba. Testing the generalized linear model null hypothesis versus smoothalternatives. *Rapport technique*, (953), 1995.

[9] Jianqing Fan, Chunming Zhang, and Jian Zhang. Generalized likelihood ratio statistics and wilks phenomenon. *Annals of statistics*, pages 153–193, 2001.

[10] Zuofeng Shang, Guang Cheng, et al. Local and global asymptotic inference in smoothing spline models. *The Annals of Statistics*, 41(5):2608–2638, 2013.

[11] Yuri Ingster and Irina A Suslina. *Nonparametric goodness-of-fit testing under Gaussian models*, volume 169. Springer Science & Business Media, 2012.

[12] Evarist Giné and Richard Nickl. *Mathematical foundations of infinite-dimensional statistical models*, volume 40. Cambridge University Press, 2015.

[13] James Mercer. Functions of positive and negative type, and their connection with the theory of integral equations. *Philosophical transactions of the royal society of London. Series A, containing papers of a mathematical or physical character*, 209:415–446, 1909.

[14] Milton Abramowitz and Irene A Stegun. *Handbook of mathematical functions: with formulas, graphs, and mathematical tables*. Courier Corporation, 1964.

[15] Bernhard Schölkopf, Ralf Herbrich, and Alex J Smola. A generalized representer theorem. In *International Conference on Computational Learning Theory*, pages 416–426. Springer, 2001.

[16] Zuofeng Shang and Guang Cheng. Computational limits of a distributed algorithm for smoothing spline. *Journal of Machine Learning Research*, 18(1):3809–3845, 2017.

[17] Paulus Petrus Bernardus Eggermont, Vincent N LaRiccia, and VN LaRiccia. *Maximum penalized likelihood estimation*. Springer, 2001.

[18] Jianqing Fan and Jian Zhang. Sieve empirical likelihood ratio tests for nonparametric functions. *Annals of Statistics*, pages 1858–1907, 2004.

[19] Erich L Lehmann and Joseph P Romano. *Testing statistical hypotheses*. Springer, 2006.

[20] Yun Yang, Mert Pilanci, and Martin J Wainwright. Randomized sketches for kernels: Fast and optimal non-parametric regression. *arXiv preprint arXiv:1501.06195*, 2015.

[21] Meimei Liu, Zuofeng Shang, and Guang Cheng. Nonparametric testing under random projection. *arXiv preprint arXiv:1802.06308*, 2018.

[22] Rafael A Irizarry, Christine Ladd-Acosta, Benilton Carvalho, Hao Wu, Sheri A Brandenburg, Jeffrey A Jeddeloh, Bo Wen, and Andrew P Feinberg. Comprehensive high-throughput arrays for relative methylation (charm). *Genome research*, 18(5):780–790, 2008.

[23] Yoav Benjamini and Yosef Hochberg. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the royal statistical society. Series B*, pages 289–300, 1995.

[24] Aravind Subramanian, Pablo Tamayo, Vamsi K Mootha, Sayan Mukherjee, Benjamin L Ebert, Michael A Gillette, Amanda Paulovich, Scott L Pomeroy, Todd R Golub, Eric S Lander, , , and Jill P Mesirov. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences*, 102(43):15545–15550, 2005.

[25] CP Pallasch, J Schwamb, S Königs, A Schulz, S Debey, D Kofler, JL Schultze, M Hallek, A Ultsch, and CM Wendtner. Targeting lipid metabolism by the lipoprotein lipase inhibitor orlistat results in apoptosis of b-cell chronic lymphocytic leukemia cells. *Leukemia*, 22(3):585–592, 2008.

[26] Martin Bilban, Daniel Heintel, Theresa Scharl, Thomas Woelfel, Michael M Auer, Edit Porpaczy, Birgit Kainz, Alexander Kröber, Vincent J Carey, Medhat Shehata, et al. Deregulated expression of fat and muscle genes in b-cell chronic lymphocytic leukemia with high lipoprotein lipase expression. *Leukemia*, 20(6):1080–1088, 2006.

[27] Masaki Okano, Shaoping Xie, and En Li. Cloning and characterization of a family of novel mammalian dna (cytosine-5) methyltransferases. *Nature genetics*, 19(3):219, 1998.

[28] Nancy E Heckman. Spline smoothing in a partly linear model. *Journal of the Royal Statistical Society. Series B*, pages 244–248, 1986.

[29] Mikio L Braun. Accurate error bounds for the eigenvalues of the kernel matrix. *Journal of Machine Learning Research*, 7(Nov):2303–2328, 2006.

# Chapter 2

## MetaGen: reference-free learning with multiple metagenomic samples

### 2.1 Background

Due to the rapid advancement of high-throughput sequencing technologies, metagenomics, which investigates the genetic contents of the entire collection of microbial species in a set of environmental samples, is becoming a major tool for studying microbial ecology, evolution, and diversity, as well as linking microbial features to the surrounding environment or human health [1–3].

In the past decade, many methods have been proposed to estimate microbial compositions from metagenomic sequencing data, with a majority focused on those targeted sequencing data that only provide information on a few selected genes such as 16s rRNA [4]. Because the targeted approach requires the sequencing of only a limited number of genes instead of hundreds of microbial genomes, it is cheap and computationally efficient. The trade-offs are that it can only reach a fairly high taxonomy rank, i.e. having a relatively low resolution in differentiating distinct species, and that it cannot provide information regarding other important genomic components. Moreover, statistical estimation based on targeted sequencing data can be biased because the PCR primers used for amplifying the targeted genes, such as 16s rRNA, have different levels of sensitivity in different species [5].

Because of the drastic cost reduction in next-generation sequencing technologies and the disadvantages of targeted-gene based approaches, the genome-wide shotgun sequencing has become a dominant technique in metagenomic studies. The genomic fragments obtained from metagenomic samples are binned into different species or taxonomical bins either according

to the fragments' similarities to some known reference genomes or according to the sequence composition similarities (e.g., similarities between $k$-mer distributions [6] or oligonucleotide frequencies [7]). This class of approaches is referred to as "binning methods". Reference-based binning methods such as MAGEN [8], MetaPhyler [9], Kraken [10] and CLARK [11] require us to know the reference genomes of the interested microbial species, which can be a serious limitation. In contrast, the $k$-mer or the oligonucleotide-frequency based methods are reference-free. However, the binning accuracy of $k$-mer based method can be significantly compromised because the $k$-mer distributions estimated from short contigs (e.g., <10kb) can be far from their corresponding whole-genome $k$-mer distributions. Meanwhile, the effectiveness of $k$-mer based methods is also diminished when the microbial community under consideration contains organisms with moderate to high sequence similarities. In order to improve $k$-mer based approaches, coverage-based methods such as CONCOCT [12], MaxBin [13], MetaBAT [14], Groopm [15] and VizBin [16] are developed to integrate the coverage information (i.e., the average number of short-reads covering each base pair of a contig after alignment) with the sequence composition information. Although integrating coverage information can significantly improve the binning accuracy, how to balance the $k$-mer information with the coverage information is by no means a banal development. Our simulation studies suggest that most of the existing coverage-based methods still fail in distinguishing genetically similar species. Moreover, the coverage estimate is biased when the species does not have adequately coverage or when the sequence bias is high.

In this article, we propose a reference-free and distribution-free binning method, MetaGen, which makes use of the relative abundance information from multiple samples to cluster contigs into different species bins and relies on the Bayesian information criterion (BIC) to determine the number of species in the samples. Since MetaGen solely uses the cross-sample abundance patterns for binning, we recommend that the number of samples in consideration should be larger than 10 (or 5% of the total number of specie). Compared to

existing unsupervised binning methods, MetaGen not only clusters short contigs accurately for samples with low coverage but also has the ability to distinguish species with high sequence similarities. In addition, MetaGen can estimate the relative abundance of cultured and uncultured species simultaneously, which provides a way to study distributional changes in microbial colonies dynamically and spatially. Moreover, MetaGen is not susceptible to sequencing biases, which is an important advantage compared with many existing methods. MetaGen is computationally efficient and can easily handle large data sets with more than $500,000$ contigs.

## 2.2 Results

### 2.2.1 Multi-sample reference-free binning: An overview

We consider metagenomic sequencing data consisting of short reads from the genomes of the organisms in the samples. The first step in almost all analysis methods is to connect overlapping short reads from the pooled sample into longer sequences, termed as "contigs". The $k$-mer based reference-free methods proceed to "bin" (i.e., cluster) these contigs, regarding them as coming from the same or similar species, according to similarities among the $k$-mer distributions of these contigs. Our proposed method, MetaGen, however, uses the relative abundance information of the contigs across multiple samples to cluster them. Thus, whereas the $k$-mer based methods need to assume that contigs derived from the same species have similar $k$-mer distributions, MetaGen assumes that abundances of different species vary across multiple samples.

Since each contig is composed of many short reads coming from all samples, we define each contig's *sample profile* as the vector of percentages of short reads mapped from different samples. As the genome of a species can be thought of as the longest possible contig, we refer to the similarly-defined short-read percentage vector as the species's *sample profile*. In theory, a contig's *sample profile* should be the same as the *sample profile* of the species that contains

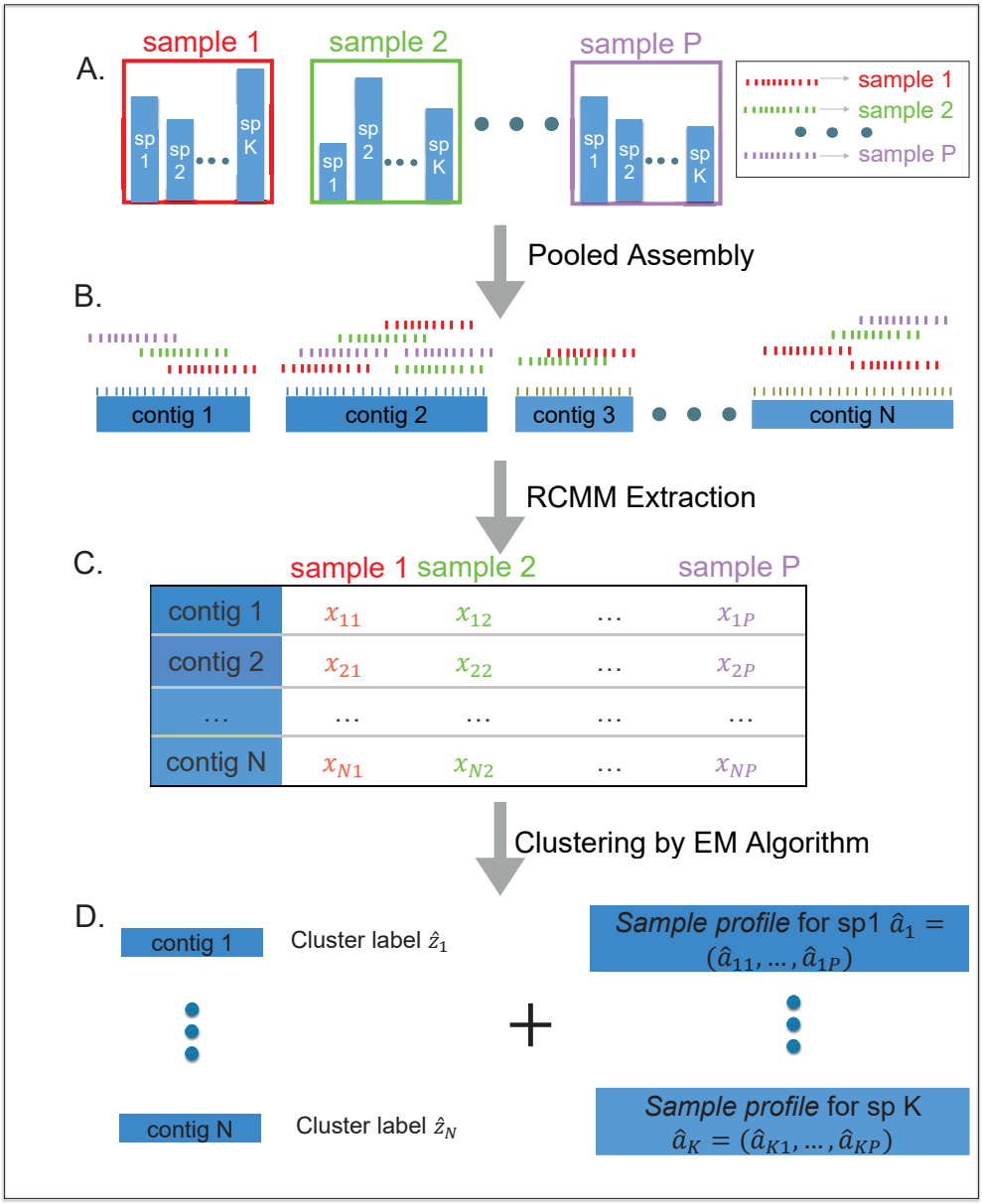



**Figure 2.1: MetaGen Pipline: A.** Sequencing the DNA of $P$ metagenomic samples. **B.** Pooled assembly for multiple samples. **C.** Constructing the RCMM (Read Counts Mapping Matrix). **D.** Clustering the contigs and estimating the *sample profile* by the EM Algorithm.

the contig (if we assume that the contig is long enough for a unique mapping). Thus, if two contigs have similar *sample profiles*, they are likely derived from the same genome. MetaGen

models the mapped short-read counts of each contig by a mixture of multinomial distributions, with each of its mixture components representing a distinct species. The limitation of MetaGen is that if two species have nearly proportional abundances in all the samples, their corresponding contigs will tend to have highly correlated *sample profiles*, which makes it difficult for MetaGen to differentiate the two species. As shown by our simulation studies, however, this difficulty can be alleviated by increasing the sequencing depth.

### 2.2.2 Statistical deconvolution of metagenomic samples

As explained previously, if two contigs have very similar *sample profiles*, they are likely part of the same species' genome. Let us assume that $N$ contigs were obtained from $P$ metagenomic samples, with a total of $K$ species involved. The extracted *read counts mapping matrix* (RCMM) has $N$ rows and $P$ columns, with its $(i, j)$th entry recording the read count from the $j$th sample mapped on to the $i$th contig, as shown in Step C of Fig. 2.1. Thus, each row of RCMM is proportional to the *sample profile* of a contig. A direct clustering of the rows of RCMM provides information about the number of species and their distributions in the samples.

Let $\mathbf{X}_i$, $i = 1, \ldots, N$, denote the row vectors of the RCMM, each of length $P$, and let $Z_i$ take values in $\{1, \ldots, K\}$, indicating from which species contig $i$ is derived. We assume that the $Z_i$'s are independent, and $P(Z_i = k) = \pi_k$, with the probability vector $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_K)$. Furthermore, we assume that given the species label $Z_i$, $\mathbf{X}_i$ follows the multinomial distribution:

$$\Pr(\mathbf{X}_i = \mathbf{x}_i | Z_i = k) = \frac{n_i}{x_{i1}! \ldots x_{iP}!} a_{k1}^{x_{i1}} \ldots a_{kP}^{x_{iP}} \tag{2.1}$$

where $\mathbf{a}_k = (a_{k1}, \ldots, a_{kP})$, $\sum_{j=1}^{P} a_{kj} = 1$, is the *sample profile* of the $k$th species, $n_i = \sum_{j=1}^{P} x_{ij}$ is the total number of mapped reads on $i$th contig. Let $A$ denote the $K \times P$ *sample profile* matrix constructed by stacking up the $\mathbf{a}_k$'s, and let $\theta = (\boldsymbol{\pi}, A)$. Treating $Z_i$ as missing

data, we have the complete-data likelihood function as

$$L(\theta; \mathbf{x}_1, \ldots, \mathbf{x}_N, z_1, \ldots, z_N) =$$
$$\prod_{i=1}^{N} \sum_{k=1}^{K} \pi_k \mathbf{1}(z_i = k) \frac{n_i}{x_{i1}! \ldots x_{iP}!} a_{k1}^{x_{i1}} \ldots a_{kP}^{x_{iP}} \tag{2.2}$$

where $\mathbf{1}(\cdot)$ is an indicator function. The maximum likelihood estimate (MLE) of $\theta$ can be obtained by the Expectation-Maximization (EM) algorithm [17], which iterates the following two steps:

**E-step**: Calculate $Q(\theta|\theta^{(t)})$, the expectation of the complete-data log-likelihood function based on the parameter fixed at $\theta^{(t)}$:

$$Q(\theta \mid \theta^{(t)}) = \sum_{i=1}^{N} \sum_{k=1}^{K} \hat{q}_{ik}^{(t)} \left[ \log \pi_k + \sum_{j=1}^{P} x_{ij} \log(a_{kj}) \right], \tag{2.3}$$

where $\hat{q}_{ik}^{(t)} = \pi_k^{(t)} a_{k1}^{(t)\, x_{i1}} \ldots a_{kP}^{(t)\, x_{iP}} \Big/ \left[ \sum_{l=1}^{K} \pi_l^{(t)} a_{l1}^{(t)\, x_{i1}} \ldots a_{lP}^{(t)\, x_{iP}} \right]$.

**M-step**: Find $\hat{\theta}$ that maximizes the function $Q(\theta|\theta^{(t)})$. This leads to

$$\pi_k^{(t+1)} \propto \sum_{i=1}^{N} \hat{q}_{ik}^{(t)}; \quad \text{and} \quad a_{kj}^{(t+1)} \propto \sum_{i=1}^{N} \hat{q}_{ik}^{(t)} x_{ij} \tag{2.4}$$

**Initialization and final clustering**: Although each EM iteration increases the observed-data likelihood function, the algorithm does not guarantee to converge to the global maximum. We thus employed the following initialization strategy: we first select the $10\% \sim 30\%$ contigs with the largest number of mapped reads and cluster the selected contigs into $K$ species using hierarchical clustering with their pairwise distance defined by

$$d(\mathbf{x}_1, \mathbf{x}_2) = 1 - \frac{\sum_{j=1}^{P} x_{1j} x_{2j}}{\sqrt{\sum_{j=1}^{P} x_{1j}^2 \sum_{j=1}^{P} x_{2j}^2}}.$$

The class mean of species $k$ is then used as the starting values $a_{kj}^{(0)}$, $j = 1, \cdots, P$. With the MLE $\hat{\theta}$ obtained by the EM-algorithm, we assign each $\mathbf{x}_i$ to the species with the highest posterior probability, i.e., we set $\hat{z}_i = \text{argmin}_k\, \hat{q}_{ik}$, $i = 1, \ldots, N$.

### 2.2.3 Determining the number of species in the samples

Since the number of species is generally unknown in most applications, we employed the Bayesian information criteria (BIC) [18; 19] to select the number of species. The BIC score for our model with $K$ species is defined as

$$\mathrm{BIC}(K) = -2 \log L(\hat{\theta}; \mathbf{x}_1, \ldots, \mathbf{x}_N) + (K * P + K) \log(N) \tag{2.5}$$

We determine the number of species $\tilde{K}$ by minimizing this score, i.e.,

$$\tilde{K} = \underset{K}{\operatorname{argmin}} \, \mathrm{BIC}(K) \tag{2.6}$$

In practice, we gradually increase the number of species and stop when BIC score begins to increase. Our simulation studies showed that the criterion worked satisfactorily in accurately determining the number of species involved in the studies.

### 2.2.4 Comparison with coverage-based metagenomic binning methods

There are two types of information contained in metagenomic data: the sequence content information and the sequence quantity information (i.e., numbers of mapped reads of constructed contigs). The sequence content information has been extensively used in existing metagenomic binning methods, whereas the sequence quantity information is much less used. A few exceptions such as CONCOCT, Maxbin and MetaBAT bin contigs together if their sequencing coverages (the average number of reads that can be aligned to a reference base) are similar. These methods intrinsically assume that no fragment of any involved genome in the sample has positional bias. They work well for GC-neutral or GC-rich species, in which the regional GC-bias is not a serious issue. As shown in [20–22], however, the sequencing coverage can be highly variable along the genome, especially for species with a low GC content. For example, it was shown in [21] that *Beta vulgaris BAC ZR-47B15* has nearly 7 times more coverage in GC-rich regions than in GC-poor regions. Consequently, binning contigs

based on their coverage similarities is highly susceptible to sequencing bias. In contrast, MetaGen is less susceptible to sequencing bias since it bins contigs based on the ratio of the mapped-reads counts (i.e., the *sample profile*). Sequencing biases do not affect the *sample profile* because these biases are the same across samples and thus can be canceled out. In other words, two contigs from the same species can still be binned together even if their observed coverage is very different due to positional biases.

Another unique feature of MetaGen is that it does not use the sequence (content) information in binning, because the information gain is offset by undesirable sequencing biases and high computational costs, especially when one deals with short contigs produced from data with relatively low sequencing coverages. As reviewed previously, short contigs are more susceptible to positional and sequencing biases. As shown in our simulation studies, for contigs shorter than 5000 bps, including the sequence information did not increase the binning accuracy, but greatly increased the computational complexity. Another reason for not using the sequence information in MetaGen is that features summarized from the sequence information and those from sequencing coverages are usually at different scales. An *ad hoc* combination of the two types of information can make the computation unstable since one type may completely dominate the other. A potential remedy is to weigh the sequence features and sequencing coverage information properly so that the contribution from each sources is on the same scale [14]. However, choosing a data-driven weight significantly increases the computational burden without bringing much improvement most of the time.

Finally, MetaGen directly models short-read counts rather than their transformations as proposed in some recent papers. Thus, it does not need to add deliberately a small "pseudo-count" to zero coverage values when calculating their logarithmic transformations as suggested in CONCOCT. Moreover, MetaGen avoids using inappropriate Gaussian distributions for non-negative zero-inflated observations as in MetaBAT, which can be important especially for low-coverage data.

### 2.2.5 Simulation studies

To investigate how the binning accuracy was affected by other parameters such as the sequencing depth, the sample size and the number of species, we conducted extensive simulations to compare MetaGen with three state-of-the-art reference-free binning methods: CONCOCT [12], MaxBin [13] and MetaBAT [14]; and one reference-based method, CLARK [11]. The names of the species (or sub-strains) used for all the setups are given in Table S1-3. All the algorithms compared here were implemented on a computer configured with 2× Intel Xeon E5-2670 and 8× 32GB RAM. Under all the simulation setups, MetaGen is at least 10 times faster than other reference-free binning methods (Figure S1).

#### How binning accuracy is affected by sequencing depth

First, we examined three sequencing depths for the pooled sample: 80**x** (1**x** per sample), 120**x** (1.5**x** per sample), and 160**x** (2**x** per sample). Short reads from 100 species mixed at a randomly generated proportional distribution was independently simulated for each of the 80 samples. Because all the methods except MetaGen can be significantly impaired for contigs shorter than 1000bps, we used only the subset of contigs with a length longer than 1000bps for CONCOCT, MetaGen, MaxBin and CLARK. For MetaBAT we used contigs longer than 1500bps, which is the default minimum length for contigs that can be used in MetaBAT. As shown in Figure 2.2, MetaGen performed well at all sequence depths by all three measures: precision, recall, and the adjusted Rand index (ARI, a combination of the precision and recall measurements), especially for data with very low sequencing depth. For example, in the case of 1**x** per sample, MetaGen achieved 0.88 ARI, whereas CONCOCT, MaxBin, and MetaBat had only 0.59, 0.14, and 0.66 ARI, respectively.

It is clear that CLARK outperformed almost all reference-free methods, especially when the sequence depth is low, because we give a significant advantage to CLARK by assuming that all the reference genomes are known (unrealistic, though). It was also shown in Figure

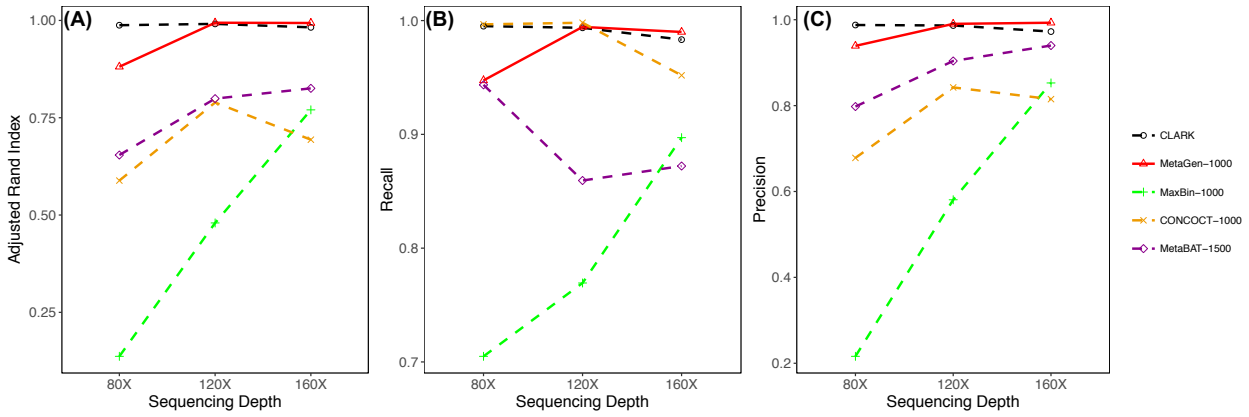



**Figure 2.2:** **(A)**.Adjusted Rand Index, **(B)**.Precision and **(C)**.Recall of CLARK, MetaGen, MaxBin, CONCOCT, and MetaBAT are evaluated under different sequencing depth, 80 samples and 100 species.

2.2 that the benefit of knowing the reference genome is not so significant when the sequence depth is high enough (say, 1.5**x** per sample). In fact, the binning accuracy for CLARK is worse than for MetaGen by a tiny margin at 2**x** per sample due to the alignment error generated by quickly approximating the similarities between contigs and the reference genomes using CLARK. The accuracy of reference-based binning methods can be improved by using BLAST, but the computational cost would be intolerably high.

### How binning accuracy is affected by sample size

In this experiment, we let the sample size vary from 20, 40 to 80 for 100 species with the pooled sequencing depth at 120**x**. We followed the same rule as used in the first experiment to generate each metagenomic samples and select subsets of contigs. Note that the per sample sequencing depth in this experiment decreased as we increased the sample size. Since the pooled sequencing depth was fixed, a contig's coverage in a single sample decreased with the increase in the sample size. As shown in Figure 2.3, the binning accuracy decreased for

all the existing coverage-based binning methods because the approximated distribution of the log-transformation of the sequencing coverage, which was used to bin contigs, performs badly if the per sample coverage is low (near zero, for example).

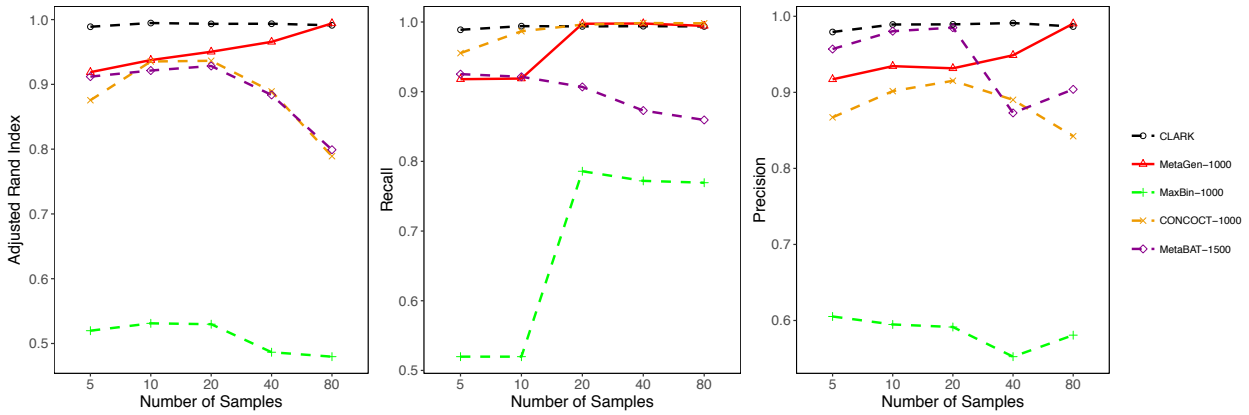



**Figure 2.3: (A)**.Adjusted Rand Index, **(B)**.Recall and **(C)**.Precision of CLARK, MetaGen, MaxBin, CONCOCT, and MetaBAT are evaluated under different number of samples, **120x** sequence depth and 100 species.

However, increasing sample size is a blessing for MetaGen, as the larger the sample size is, the higher the discrimination power of the ratio is and the higher the binning specificity is. As shown in our simulation studies, the precision increased from 0.93 to 0.99 as we increased the sample size, which in turn led to the increases in ARI.

### How binning accuracy is affected by number of species

Here we increased the number of species from 50 to 100 and 150, with the pooled sequencing depth fixed at 120**x** and the sample size fixed at 80. Again, due to the fixed pooled sequencing depth, contigs tend to be shorter for a larger number of species. Thus, increasing the number of species can lead to a higher binning error rate for all methods except MetaGen, because all other methods use $k$-mer distribution similarities for binning and consequently suffer from high binning errors, especially for contigs from genetically similar species.

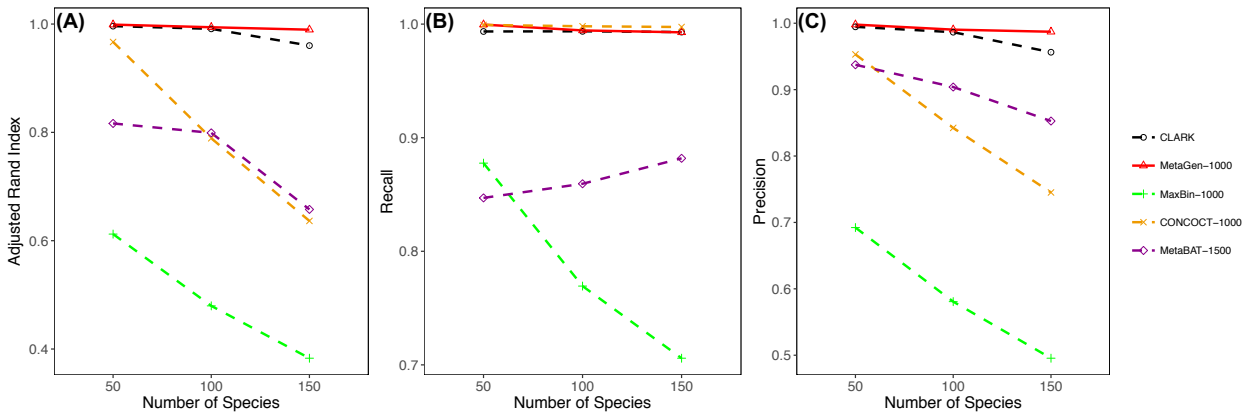



**Figure 2.4: (A)**.Adjusted Rand Index, **(B)**. Recall and **(C)**. Precision of CLARK, MetaGen, MaxBin, CONCOCT, and MetaBAT are evaluated under different number of species, **120x** sequence depth and 80 samples.

Compared to all the methods that use sequencing information, MetaGen only uses the abundance variation across samples and is consequently less susceptible to the lengths of contigs and more robust for data with a large number of species. As illustrated in Figure 2.3, the binning accuracy of MetaGen did not change significantly as we increased the number of species.

### How binning accuracy are affected by sequence similarity

Because MetaGen does not use the sequence information, the binning accuracy is not significantly affected when some of the species are highly similar in their sequences. But MetaGen requires that the distribution of species in different samples be distinguishable. For example, as shown in Figure 2.5, *Cupriavidus metallidurans* CH34 (green) and *Ralstonia eutropha* JMP134 (white), two species that are highly similar in their sequence, are successfully separated by MetaGen but mistakenly binned together in MaxBin, CONCOCT, and MetaBAT.

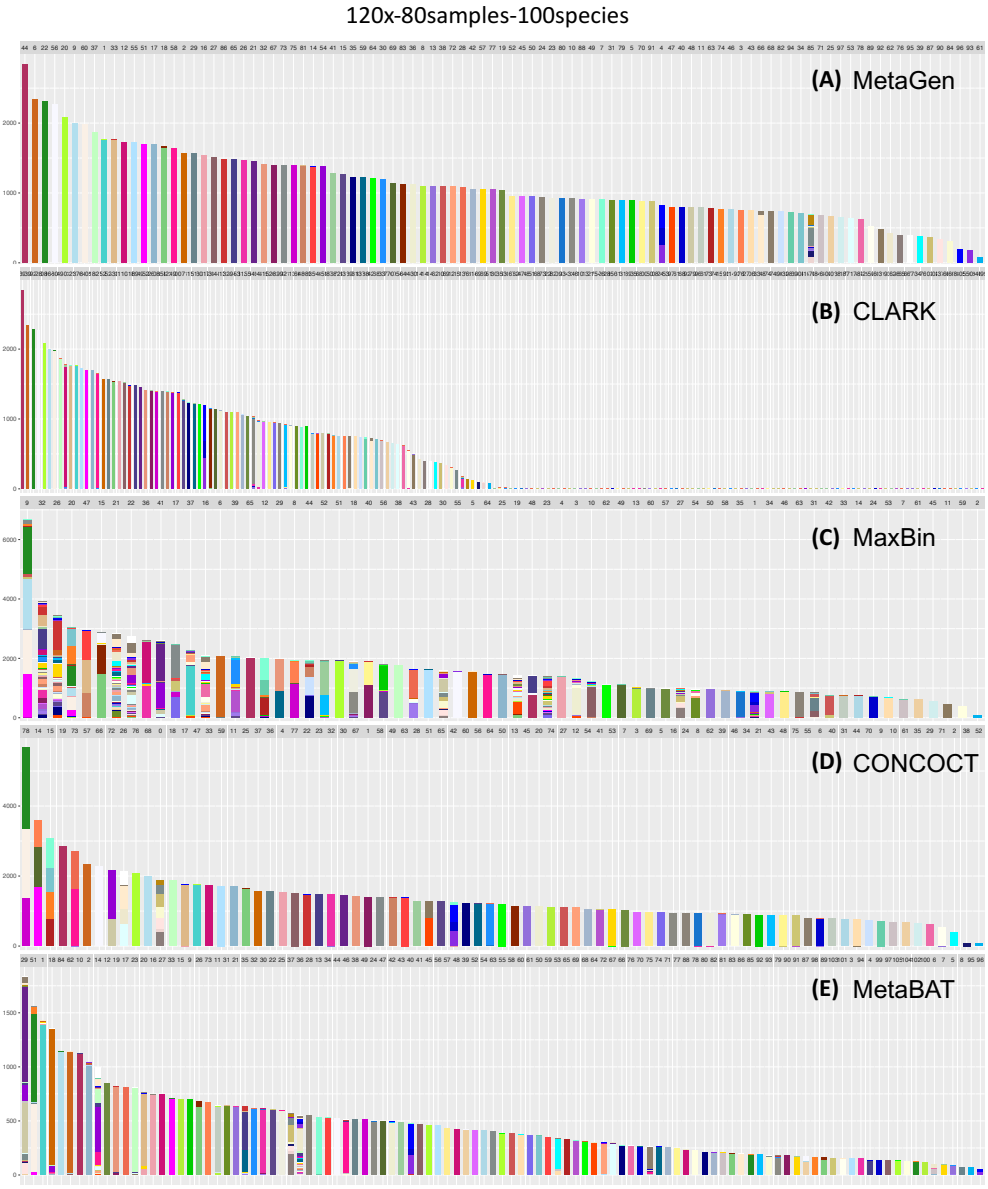



**Figure 2.5:** Binning results of CLARK(**A**), MetaGen(**B**), MaxBin(**C**), CONCOCT(**D**) and MetaBAT(**E**) under **120x** sequencing depth, 80 samples and 100 species (represented by different colors). Each bar represents one bin obtained using the corresponding binning method. The color of a bin should be identical if there is no binning error.

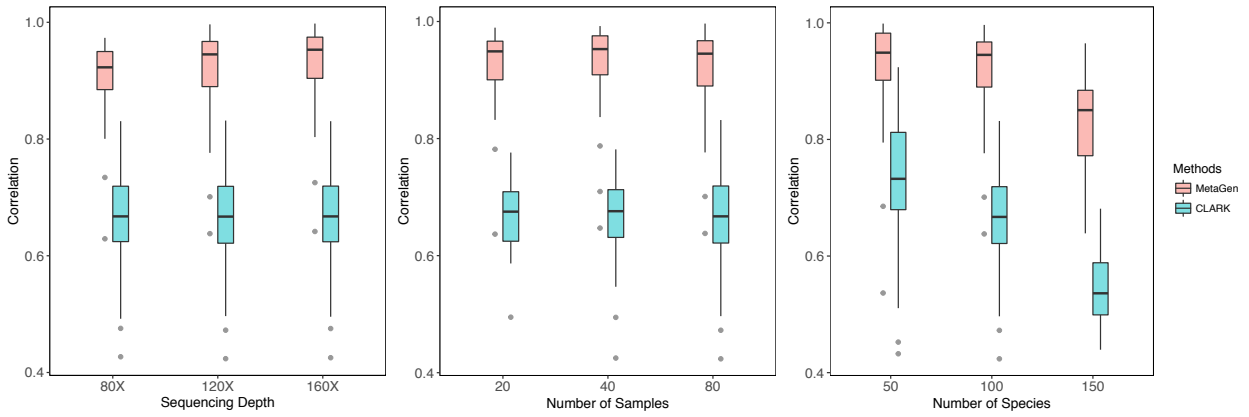



**Figure 2.6:** The boxplot of the Pearson correlation coefficient between the estimated relateive abundance(cross different species within one sample) and the underlying truth. **(A)**.The comparsion is applied to the metagenomic data sets under different sequencing depth, 80 samples and 100 species. **(B)**. The comparsion is applied to the metagenomic data sets under **120x** sequencing depth, different number of samples and 100 species. **(C)**. The comparsion is applied to the metagenomic data sets under **120x** sequencing depth, 80 samples and different number of species.

### Strain-level profiling

We studied the performance of MetaGen in distinguishing microbial strains using a mock data set with 57 *E. coli* strains and 91 circular elements. The data set we generated using MetaSim contains 40 metagenomic samples, each with 2 million paired-end reads. MetaGen outperformed other reference-free binning methods we considered including CONCOCT, MetaBat, and MaxBin, as well as the reference-based method, CLARK, in strain-level discrimination. More specifically, the ARI for MetaGen was 0.50 which is significantly higher than that for CONCOCT (0.16). CLARK assigned all the contigs to one bin because the lowest taxonomy rank that CLARK can reach is at the species level. MetaBat and MaxBin also failed in strain-level profiling by binning all 57 *E. coli* strains into one bin (MetaBat) or two bins (MaxBin). The comparison results are summarized in Table 2.1.

**Table 2.1:** Adjusted Rand Index, Precision and Recall of CLARK, MetaGen, MaxBin, CONCOCT, and MetaBAT are evaluated on the simulated metagenomic community with 57 *E. coli* substrains.

| | MetaGen | MaxBin | CONCOCT | MetaBAT | CLARK |
|---|---|---|---|---|---|
| ARI | 0.50 | 0.01 | 0.16 | 0.00 | 0.00 |
| Recall | 0.65 | 0.86 | 0.80 | 1.00 | 1.00 |
| Precision | 0.81 | 0.16 | 0.48 | 0.13 | 0.12 |

Moreover, we found that MetaGen also outperformed the most popular strain-level profiling tool, ConStrains, using the data simulated under the same settings as [23]. The modified Jenson-Shannon divergence, a measure proposed in [23] to justify the profiling error, was 0.04 for MetaGen and 0.26 for ConStrains. We did not compare MetaGen with ConStrains in distinguishing the 57 *E. coli* strains and 91 circular elements because ConStrains requires 10**x** coverage in at least one sample. This requirement was not satisfied by the 57 *E. coli* strains mock data set, which had only about 1.5**x** average coverage.

### Binning results for a complex community

To investigate the effectiveness of MetaGen for analyzing complex metagenomic communities with a limited number of samples, we simulated 10 metagenomic samples, each with 545 genomes and 439 circular elements based on the most abundant species identified by CLARK in the 269 gut metagenomic samples from [24; 25]. The relative abundance of each species in the ten samples was generated by the CLARK-estimated relative abundance from 10 randomly selected samples in [24; 25] to mimic the real relative abundance. Summarized in Table 2.2 are the adjusted Rand indexes for MetaGen, CONCOCT, MaxBin, MetaBat, and CLARK. MetaGen achieved a higher binning accuracy compared to all the reference-free binning methods in comparison, but a lower accuracy compared to the reference-based method, CLARK.

**Table 2.2:** Adjusted Rand Index, Precision and Recall of CLARK, MetaGen, MaxBin, CONCOCT, and MetaBAT are evaluated on the complex metagenomic community with 545 genomes and 439 circular elements.

| | MetaGen | MaxBin | CONCOCT | MetaBAT | CLARK |
|---|---|---|---|---|---|
| ARI | 0.67 | 0.51 | 0.42 | 0.07 | 0.86 |
| Recall | 0.89 | 0.73 | 0.86 | 0.79 | 0.96 |
| Precision | 0.76 | 0.65 | 0.53 | 0.40 | 0.90 |

### Reference-free estimation of relative abundances

MetaGen provided an estimate of the relative abundance of the microbial species in each sample without utilizing any reference information. Compared to those reference-based methods, which estimate the relative abundance of each species using the proportion of reads from its genome showing up in each sample [26; 27], MetaGen estimates the relative abundance using the estimated *sample profile* for each bin (See Eq. (7)). To compare the relative abundance estimated by each tool, we used Pearson correlation coefficients [28] to characterize the overall relationship between the estimated relative abundance (across different species within one sample) and the underlying truth. We did the comparisons for all nine simulated data sets with varying sequencing depths, number of samples, and number of species. As shown in Figure 2.6, the accuracy of estimated relative abundance by MetaGen is significantly higher than those estimated by CLARK. Even for data with a very low sequencing depth (1**x** per sample), MetaGen demonstrated a high accuracy with an average correlation of 0.908 between the estimated relative abundance and the truth.

### Some other factors relevant to estimation accuracies

Some minor but essential issues were also considered in our simulation. We first compared the binning accuracy of MetaGen to other candidate methods when some species were missing in

certain samples. In this simulation, only 50 or 75 out of 100 species were randomly selected for each sample and the binning accuracy is plotted in Figure S11 indicating that MetaGen was not affected by missing species.

We then tested how the binning accuracy is affected by using different genome assemblers, such as MegaHIT [29] and Ray [30]. Figure S14 plots the ARI, recall and precision of all five binning methods under consideration for Ray and MegaHIT respectively. Clearly, CLARK, MetaGen and MetaBAT performed marginally better using Ray and CONCOCT performs marginally better using MegaHIT. The binning accuracy of MaxBin was significantly better for MegaHIT compared to Ray. Compared to the other methods, MetaGen was least affected by the use of different assemblers.

### 2.2.6 Metagenomic analysis of inflammatory bowel disease

Inflammatory bowel disease (IBD) is an idiopathic disease caused by humans' dysregulated immune responses to their intestinal microbiota. IBD can cause abdominal cramps, bloody diarrhea, fever and weight loss, and may also increase the risk of colon cancer. Each year, about six hundred thousand Americans suffer from one of the two IBD subtypes: ulcerative colitis (UC) and Crohns disease (CD). It was recently shown in [31] that IBD is closely related to aberrant interactions between gut microbial species and the host's immune system.

Qin et al. [24] collected gut microbial DNA samples from 124 European individuals, including 25 IBD patients. The DNA samples were sequenced using Illumina Genome Analyzer with $576.7Gb$ paired-end reads generated. Using MetaGen, we inferred that at least $2,150$ clusters/species (See Figure S24) were presented in the samples, much more than the 155 species identified in [24] using a reference-based method. The significant difference between the two results is mainly caused by the limited availability of reference bacterial species. In fact, only 6.54% of the total contigs can find a closely matched reference genome in the NCBI nucleotide database. The scale of the number of species predicted by MetaGen

is also consistent with the conjecture made in [24]. For the contigs that can be mapped to reference genomes, we found that MetaGen achieved a high binning accuracy with precision 0.937 and recall 0.753. We did not compare our method to other reference-free binning methods for this study because the dataset was too large for other methods to obtain results using the computing resources we had access to.

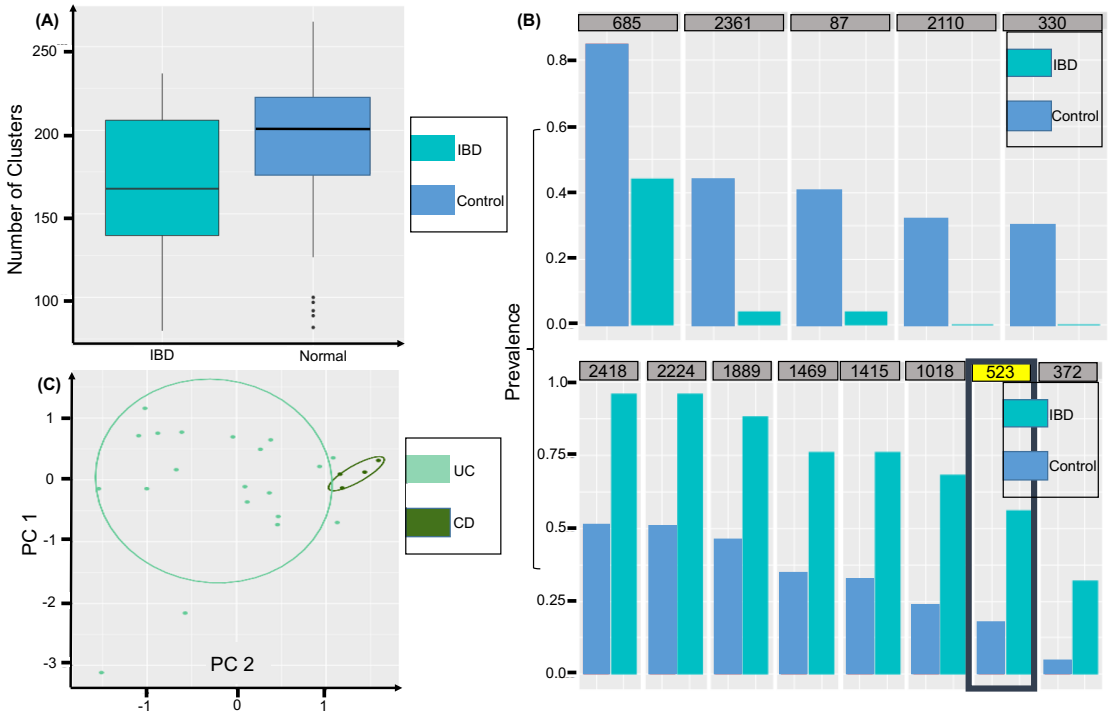



**Figure 2.7: (A)** Boxplots of the number for significant species in each individual of the IBD and control groups, respectively. **(B)**(Upper panel) Prevalences of the 5 highly enriched species in the individuals in the control group relative to the IBD patients. (Lower panel) Prevalences of the 8 highly enriched species in the IBD patients relative to the individuals in the control group. **(C)** The projection of the 4 CD patients and 21 UC patients along the first two principal component directions of the relative abundances of their microbial species.

Figure 2.7(A) shows boxplots of the number of *significant microbial species* (See **Materials and Methods** for definition) found in each individual in the IBD and control groups, respectively, indicating that the biodiversity of microbita in IBD patients is significantly lower than that in individuals in the control group (p-value=0.03). This phenomena was also observed in [25] and [32]. By testing the 561 microbial species that were shared by at least 10

individuals, we found that five species were significantly less common and eight species were significantly more common in IBD patients with the false discovery rate (FDR) controlled at under 5% [33]. Among the 8 species that are more commonly seen in IBD patients, we found that 13 of 25 contigs in one bin (highlighted by the black box in the lower panel of Figure 2.7(B)) could be mapped to an antibiotic resistance bacterial strain *Bacteroides fragilis* HMW 615 with more than 99% identity. Among the 13 contigs, 6 were mapped to *Bacteroides fragilis* HMW 615 with 100% identity.

Based on large scale metagenomic data sets, predictive models using machine learning tools have revealed good predictive capabilities for different phenotypes, such as disease state [34], plant productivity [35], and environmental factors [36]. To investigate whether the microbial composition estimated by MetaGen can be used for disease prediction, we built a logistic regression model with LASSO penalty [37] to classify the IBD and control subjects using relative abundance (See **Materials and Methods** (2.7)) of the clusters inferred by MetaGen as features. The leave-one-out cross validation (CV) procedure was used to assess the classification accuracy. The overall prediction power of the logistic regression model is quite significant, with a leave-one-out CV misclassification rate of 0.1129; the number of misclassifications for the IBD group was 12 and for the control group was 2. We further zoomed in to investigate the difference in gut microbiota between two types of patients, CD and UC, which are not readily separable using existing medical techniques [38]. Fig. 2.7(C) shows the projection of the 25 IBD subjects onto the space formed by their first and second principal components, which shows a clear separation between the two IBD subtypes.

### 2.2.7 Metagenomic analysis of Type 2 diabetes

Type 2 diabetes (T2D) is the most prevalent endocrine disease, which involves a long term metabolic disorder influenced by both genetic and environmental factors [39]. Qin et al. [25] sequenced gut microbial DNA samples from 71 Chinese T2D patients and 74 Chinese

individuals unaffected by T2D using Illumina Genome Analyzer and obtained $3.3M$ genes based on the $378.4Gb$ pair-end reads. They could not obtain taxonomy assignments and the corresponding microbial distribution estimations using a reference-based binning method because only 8.89% of contigs can be mapped to reference genomes. We re-analyzed this data set using MetaGen and identified $2,450$ species clusters (See Figure S25).

Using Fisher's exact test with FDR controlled at 5%, we found that two clusters were more abundant in the control group than in the T2D group (See Figure S26). The majority of contigs in one of the clusters can be mapped to the butyrate-producing bacteria, *Roseburia intestinalis*, which has been shown in [40] to have an immuno-metablic effect and is thus significantly less abundant in T2D patients. This finding also validates the conjecture made in [25] that beneficial bacteria are universally lost in the T2D gut. We also tested to differentiate T2D patients from the control group by building a classifier using the subjects' microbial distributions and the LASSO-logistic regression method used in the previous section. We observed that the leave-one-out CV classification error rate was 0.248. We further validated the classification accuracy using an independent dataset from [25] with 98 T2D patients and 99 controls, and obtained a misclassification error at 0.345, which is highly significant. Although the prediction accuracy is not yet ideal, our study of the T2D metagenomic data showed that an individuals microbial composition estimated in a reference-free way can be significantly predictive of the individuals disease status.

### 2.2.8 Metagenomic analysis of obesity

Obesity is a growing epidemic worldwide and has a significant negative impact on human health. Obese people have significantly higher risks for various diseases, such as high blood pressure, stroke, heart disease, diabetes, cancer, gallstones, etc. Despite its clinical importance, causes for obesity and possible therapeutic options for curing obesity remain poorly understood. Recent studies have found that some bacteria in the human gut can disrupt

metabolic/energy homeostasis [2; 41], and the bacteria's interactions with the host's genes [42] are closely associated with the host's obesity level. It is thus expected that understanding bacterial compositions of metagenomic samples from human guts may be key to understanding obesity.

In [1], DNA samples were extracted from feces of 18 human subjects belonging to 6 families, each of which includes a pair of twins and their maternal parent. After pre-processing (See SI), we obtained $25,383$ contigs. For each contig, we searched the NCBI nucleotide database and used TAXAassign (*https://github.com/umerijaz/TAXAassign*) to assign it to a taxonomic group. Only 29% of the contigs could be assigned at the species level and 54% could be assigned at the phylum level. Roughly 46% of contigs could not be mapped to any reference genomes even at the phylum level. Thus, reference-free binning methods are highly desirable for this data.

Using MetaGen, we identified 56 bins/species (Figure S26) and estimated their relative abundances across samples. For the contigs that have species-level reference genomes, we compared MetaGen with CONCOCT using the reference-based binning results as a gold-standard. We observed that the results of MetaGen were closer to the reference-based binning results (with adjusted Rand index of 0.746) than those of CONCOCT are (with adjusted Rand index 0.592). In Figure 2.8(A), we compared the estimated relative abundances to those published in [2] at phylum level (more details in **SI Note**). MetaGen can accurately estimate relative abundances of the four most enriched phylums: *Firmicutes*, *Bacteroidetes*, *Actinobacteria* and *Verrucomicrobia*. Figure 2.8(B) provides a more detailed relative abundance estimate at species level, an estimate that could not be obtained in [2] due to the limitations of the reference based binning methods.

Figure 2.8(C) shows all pairwise Pearson correlations of the relative abundance for the 18 individuals. Using hierarchical clustering, we obtained two major clusters: Group I includes three families, in which all mothers were obese although the children were either obese or lean.

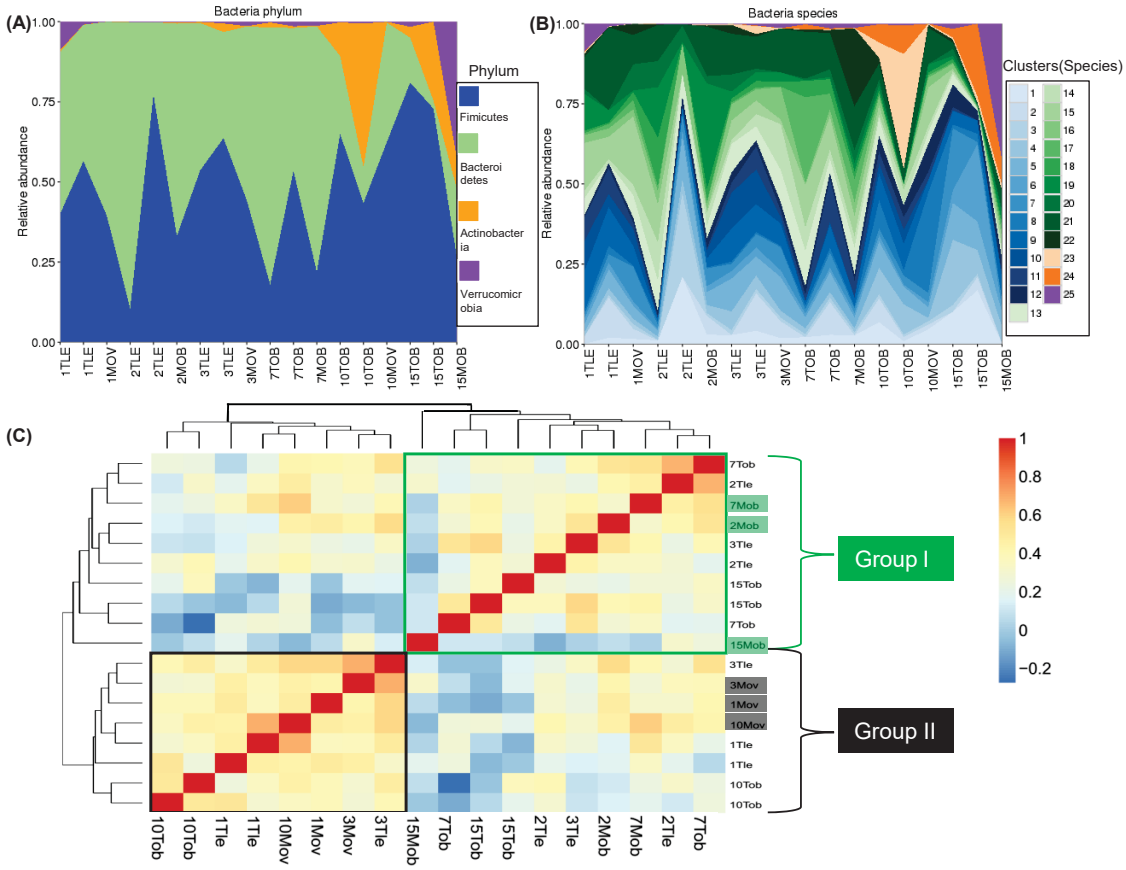



**Figure 2.8: (A)** Relative abundances of *Firmicutes*,*Bacteroidetes*, *Actinobacteria* and *Verrucomicrobia* phylums estimated by MetaGen. **(B)** Relative abundances of the 25 species mapped to one of the four aforementioned phylums. Cluster 1 to 12 are species in *Firmicutes*, cluster 13 to 22 are species in *Bacteroidetes*, cluster 23 to 24 are species in *Actinobacteria* and cluster 25 is a species in *Verrucomicrobia*. **(C)** Heatmap of the correlation of the relative abundance for the 18 individuals (samples). The samples are clustered by hierarchical clustering using complete linkage functions. In all the plots, a subject's ID can be parsed into three parts: the family ID (1-6), twin or mother (T, M), and BMI (LEan, OVerweight, or OBese).

In contrast, all mothers in group II were overweight. Only members of one family were split into the two clusters. The correlation analysis suggests that the microbial distribution of the mother is associated with her BMI status and also plays a key role in shaping up the microbial distribution of her children. To test the predictive power of the microbial distribution of the

identified species for a individual's BMI status, we fitted a LASSO-logistic regression model using the relative abundances as predictors and the individual's BMI status as the response. The leave-one-out cross-validation error rate of the resulting model was 0.33.

## 2.3 Discussion

We proposed a new method, MetaGen, for estimating species compositions in multiple metagenomic samples without any prior knowledge of either reference microbial genomes or the actual microbial distributions of the samples. MetaGen is thus a completely reference-free metagenomic procedure and is especially useful for analyzing new and foreign microbial samples. As demonstrated by our simulation studies, MetaGen can handle data with fairly low sequencing coverage, which can be extremely challenging with the currently available methods for metagenomic analysis. When a reference genome is available for some of the microbial species, we recommend the use of MetaGen together with reference-based methods as a safeguard against possible false positives.

As a trade-off for having no reference genomes, MetaGen requires multiple samples (preferably $\geq 10$) and imposes a key *differential abundance* assumption, i.e., the abundance patterns of microbial species across multiple samples should vary appreciably. This assumption is clearly confounded with sequencing depth in the study: by increasing the sequencing depth, one can recognize more species as is true for all other available methods. The differential abundance assumption can be satisfied in most metagenomic studies related to human health, such as the study of microbial distributions in the human gut and the study of human pathogens in a bio-threat attack. When the number of bacterial species is extremely large, many low abundance species will have low coverage and cannot be detected. This limitation can be overcome by performing a screening step to trim the contigs with very low coverage.

## 2.4 Materials and Methods

### 2.4.1 Connection with Non-negative Matrix Factorization

The sample profile-based binning problem can also be solved by a non-negative matrix factorization (NMF) algorithm, of which the EM algorithm can be viewed as a principled generalization. In situations where the information is strong enough so that random errors and fluctuation can be ignored, the $(i, j)$th entry of RCMM, $x_{ij}$, is just the *theoretical* amount of reads that are mapped to contig $i$ in sample $j$, which should be equal to the number of short-reads that one copy of contig $i$ can produce multiplied by the number of copies of contig $i$ in $j$th sample.

If we assume that the contig is long enough so that it only belongs to one species, we can rewrite the RCMM $\mathbf{X}$ as the product of a signature matrix $M$ and the total abundance matrix $E$, where $(i, k)$th entry of $M$ is the number of reads that a single copy of contig $i$ in species $k$ can produce (it is zero if the $k$th species does not contain contig $i$), and the $(k, j)$th entry of $E$ represents the amount of species $k$ in sample $j$. Thus, we can obtain an estimate of both $M$ and $E$ simultaneously by minimizing $||X - ME||_F$, where $|| \cdot ||_F$ denotes the Frobenius matrix norm. Note that if we normalize each row of $E$ to sum to one, it gives rise to the *sample profile* matrix $A$, i.e., $E = DA$, where $D$ is a diagnal matrix with $d_{ii}$ indicating the total number of counts for contig $i$ in the pooled sample. Based on extensive simulations, we observed that the NMF algorithm and the EM algorithm lead to very similar results empirically for given $K$. However, this NMF approach cannot account for the estimation uncertainty and also does not provide a principled way to determine the number of species $K$.

### 2.4.2 Normalization to compare microbial distributions across samples

To compare microbial distributions across samples, we need to normalize the *sample profiles* of different species to control the between-sample library size (sequence depth) variation and the genome length variation. Motivated by the definition of RPKM, which has been commonly used to normalize RNAseq data across samples and across genes, we first rescale the number of mapped reads for species $k$ in sample $j$, i.e., $\hat{a}_{kj} \sum_{\{i:\hat{z}_i=k\}} n_i$, where $n_i$ is the total number of mapped reads on contig $i$, by a factor reflecting sample $j$'s library size, i.e., the total reads $T_j$ in sample $j$, and by another factor estimating the genome length of each species, i.e.,the sum of length of all contigs for species $k$, say $L_k$. To set the number in a comfortable range, we multiply the rescaled number by a constant $10^9$ and denote it by $\hat{b}_{kj}$:

$$\hat{b}_{kj} = 10^9 \times \frac{\hat{a}_{kj} \sum_{\{i:\hat{z}_i=k\}} n_i}{L_k T_j}, \tag{2.7}$$

where $\hat{a}_{kj}$ and $\hat{z}_i$ are obtained using our algorithm. We refer to $\hat{b}_{kj}$ as *relative abundance* of species $k$ in sample $j$. To compare the relative abundance in each sample, we recommend to add an additional step to correct the GC bias by using GCcorrect (R package) [43]. When a species has relative abundance $\hat{b}_{kj} \geq 0.1\% \sum_{k=1}^{K} \hat{b}_{kj}$, we define the species to be a *significant microbial species* for sample $j$. Here, we use 0.1% as a convenient cut off because the relative abundances that are lower than 0.1% may suffer from a much higher estimation error and thus be unreliable.

### 2.4.3 Evaluating the binning results

In order to evaluate the estimated bins with true taxonomical groups, we define two groupings, $\mathbf{x} = (x_1, \cdots, x_r)$ and $\mathbf{y} = (y_1, \cdots, y_s)$ where $r$ and $s$ are the number of clusters for groupings $\mathbf{x}$ and $\mathbf{y}$, respectively. Then we denote $n_{ij}$ as the number of members that belong to both the $x_i$ and $y_j$ clusters (overlap). The adjusted Rand index is defined as,

$$\text{ARI} = \frac{\sum_{i,j}\binom{n_{ij}}{2} - E}{\frac{1}{2}\left[\sum_i \binom{r_i}{2} + \sum_j \binom{c_j}{2}\right] - E}, \tag{2.8}$$

where $E = \left[\sum_i \binom{r_i}{2} \sum_j \binom{c_j}{2}\right] \Big/ \binom{N}{2}$ is the expected index.

The *precision* is defined as the clustering accuracy under the most favorable species label assignment for each cluster. That is, assuming that grouping $\mathbf{y}$ is the true species label, the precision can be expressed as

$$\text{Precision} = \frac{\sum_{i=1}^{r} \max(n_{i1}, \cdots, n_{is})}{N}. \tag{2.9}$$

On the other hand, the *recall* is defined as how well the best cluster for each species regroup all the cluster's contigs. That is, assuming that grouping $\mathbf{y}$ is the true species label, the recall is

$$\text{Recall} = \frac{\sum_{j=1}^{s} \max(n_{1j}, \cdots, n_{rj})}{N}. \tag{2.10}$$

## Reference


[1] Steven R Gill, Mihai Pop, Robert T DeBoy, Paul B Eckburg, Peter J Turnbaugh, Buck S Samuel, Jeffrey I Gordon, David A Relman, Claire M Fraser-Liggett, and Karen E Nelson. Metagenomic analysis of the human distal gut microbiome. *Science*, 312(5778): 1355–1359, 2006.

[2] Peter J Turnbaugh, Micah Hamady, Tanya Yatsunenko, Brandi L Cantarel, Alexis Duncan, Ruth E Ley, Mitchell L Sogin, William J Jones, Bruce A Roe, Jason P Affourtit, et al. A core gut microbiome in obese and lean twins. *Nature*, 457(7228):480–484, 2009.

[3] E Stackebrandt, W Liesack, and BM Goebel. Bacterial diversity in a soil sample from a subtropical australian environment as determined by 16s rdna analysis. *The FASEB Journal*, 7(1):232–236, 1993.

[4] Jill E Clarridge. Impact of 16s rrna gene sequence analysis for identification of bacteria on clinical microbiology and infectious diseases. *Clinical Microbiology Reviews*, 17(4): 840–862, 2004.

[5] Justin Kuczynski, Christian L Lauber, William A Walters, Laura Wegener Parfrey, José C Clemente, Dirk Gevers, and Rob Knight. Experimental and analytical tools for studying the human microbiome. *Nature Reviews Genetics*, 13(1):47–58, 2011.

[6] Hanno Teeling, Anke Meyerdierks, Margarete Bauer, Rudolf Amann, and Frank Oliver Glöckner. Application of tetranucleotide frequencies for the assignment of genomic fragments. *Environmental Microbiology*, 6(9):938–947, 2004.

[7] Takashi Abe, Hideaki Sugawara, Makoto Kinouchi, Shigehiko Kanaya, and Toshimichi Ikemura. Novel phylogenetic studies of genomic sequence fragments derived from uncultured microbe mixtures in environmental and clinical samples. *DNA Research*, 12(5): 281–290, 2006.

[8] Daniel H Huson, Alexander F Auch, Ji Qi, and Stephan C Schuster. Megan analysis of metagenomic data. *Genome research*, 17(3):377–386, 2007.

[9] Bo Liu, Theodore Gibbons, Mohammad Ghodsi, and Mihai Pop. Metaphyler: Taxonomic profiling for metagenomic sequences. In *Bioinformatics and Biomedicine (BIBM), 2010 IEEE International Conference on*, pages 95–100. IEEE, 2010.

[10] Derrick E Wood and Steven L Salzberg. Kraken: ultrafast metagenomic sequence classification using exact alignments. *Genome biology*, 15(3):R46, 2014.

[11] Rachid Ounit, Steve Wanamaker, Timothy J Close, and Stefano Lonardi. Clark: fast and accurate classification of metagenomic and genomic sequences using discriminative k-mers. *BMC genomics*, 16(1):236, 2015.

[12] Johannes Alneberg, Brynjar Smári Bjarnason, Ino de Bruijn, Melanie Schirmer, Joshua Quick, Umer Z Ijaz, Leo Lahti, Nicholas J Loman, Anders F Andersson, and Christopher Quince. Binning metagenomic contigs by coverage and composition. *Nature methods*, 11(11):1144–1146, 2014.

[13] Yu-Wei Wu, Blake A Simmons, and Steven W Singer. Maxbin 2.0: an automated binning algorithm to recover genomes from multiple metagenomic datasets. *Bioinformatics*, page btv638, 2015.

[14] Dongwan D Kang, Jeff Froula, Rob Egan, and Zhong Wang. Metabat, an efficient tool for accurately reconstructing single genomes from complex microbial communities. *PeerJ*, 3:e1165, 2015.

[15] Michael Imelfort, Donovan Parks, Ben J Woodcroft, Paul Dennis, Philip Hugenholtz, and Gene W Tyson. Groopm: an automated tool for the recovery of population genomes from related metagenomes. *PeerJ*, 2:e603, 2014.

[16] Cedric C Laczny, Tomasz Sternal, Valentin Plugaru, Piotr Gawron, Arash Atashpendar, Houry Hera Margossian, Sergio Coronado, Laurens Van der Maaten, Nikos Vlassis, and Paul Wilmes. Vizbin-an application for reference-independent visualization and human-augmented binning of metagenomic data. *Microbiome*, 3(1):1, 2015.

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

[18] Gideon Schwarz et al. Estimating the dimension of a model. *The annals of statistics*, 6 (2):461–464, 1978.

[19] Robert E Kass and Larry Wasserman. A reference bayesian test for nested hypotheses and its relationship to the schwarz criterion. *Journal of the american statistical association*, 90(431):928–934, 1995.

[20] Iwanka Kozarewa, Zemin Ning, Michael A Quail, Mandy J Sanders, Matthew Berriman, and Daniel J Turner. Amplification-free illumina sequencing-library preparation facilitates improved mapping and assembly of (g+ c)-biased genomes. *Nature methods*, 6(4): 291–295, 2009.

[21] Juliane C Dohm, Claudio Lottaz, Tatiana Borodina, and Heinz Himmelbauer. Substantial biases in ultra-short read data sets from high-throughput dna sequencing. *Nucleic acids research*, 36(16):e105–e105, 2008.

[22] Yen-Chun Chen, Tsunglin Liu, Chun-Hui Yu, Tzen-Yuh Chiang, and Chi-Chuan Hwang. Effects of gc bias in next-generation-sequencing data on de novo genome assembly. *PloS one*, 8(4):e62856, 2013.

[23] Chengwei Luo, Rob Knight, Heli Siljander, Mikael Knip, Ramnik J Xavier, and Dirk Gevers. Constrains identifies microbial strains in metagenomic datasets. *Nature biotechnology*, 33(10):1045–1052, 2015.

[24] Junjie Qin, Ruiqiang Li, Jeroen Raes, Manimozhiyan Arumugam, Kristoffer Solvsten Burgdorf, Chaysavanh Manichanh, Trine Nielsen, Nicolas Pons, Florence Levenez, Takuji Yamada, et al. A human gut microbial gene catalogue established by metagenomic sequencing. *nature*, 464(7285):59–65, 2010.

[25] Junjie Qin, Yingrui Li, Zhiming Cai, Shenghui Li, Jianfeng Zhu, Fan Zhang, Suisha Liang, Wenwei Zhang, Yuanlin Guan, Dongqian Shen, et al. A metagenome-wide association study of gut microbiota in type 2 diabetes. *Nature*, 490(7418):55–60, 2012.

[26] Nicola Segata, Levi Waldron, Annalisa Ballarini, Vagheesh Narasimhan, Olivier Jousson, and Curtis Huttenhower. Metagenomic microbial community profiling using unique clade-specific marker genes. *Nature methods*, 9(8):811–814, 2012.

[27] David Koslicki, Simon Foucart, and Gail Rosen. Wgsquikr: fast whole-genome shotgun metagenomic classification. *PloS one*, 9(3):e91784, 2014.

[28] Stinus Lindgreen, Karen L Adair, and Paul P Gardner. An evaluation of the accuracy and speed of metagenome analysis tools. *Scientific reports*, 6, 2016.

[29] Dinghua Li, Ruibang Luo, Chi-Man Liu, Chi-Ming Leung, Hing-Fung Ting, Kunihiko Sadakane, Hiroshi Yamashita, and Tak-Wah Lam. Megahit v1. 0: A fast and scalable metagenome assembler driven by advanced methodologies and community practices. *Methods*, 102:3–11, 2016.

[30] Sébastien Boisvert, Frédéric Raymond, Élénie Godzaridis, François Laviolette, Jacques Corbeil, et al. Ray meta: scalable de novo metagenome assembly and profiling. *Genome Biol*, 13(12):R122, 2012.

[31] Harry Sokol and Philippe Seksik. The intestinal microbiota in inflammatory bowel diseases: time to connect with the host. *Current opinion in gastroenterology*, 26(4): 327–331, 2010.

[32] Chaysavanh Manichanh, Lionel Rigottier-Gois, Elian Bonnaud, Karine Gloux, Eric Pelletier, Lionel Frangeul, Renaud Nalin, Cyrille Jarrin, Patrick Chardon, Phillipe Marteau, et al. Reduced diversity of faecal microbiota in crohns disease revealed by a metagenomic approach. *Gut*, 55(2):205–211, 2006.

[33] Yoav Benjamini and Yosef Hochberg. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 289–300, 1995.

[34] Edoardo Pasolli, Duy Tin Truong, Faizan Malik, Levi Waldron, and Nicola Segata. Machine learning meta-analysis of large metagenomic datasets: tools and biological insights. *PLoS Comput Biol*, 12(7):e1004977, 2016.

[35] Hao-Xun Chang, James S Haudenshield, Charles R Bowen, and Glen L Hartman. Metagenome-wide association study and machine learning prediction of bulk soil microbiome and crop productivity. *Frontiers in Microbiology*, 8, 2017.

[36] Bettina Glasl, Nicole S Webster, and David G Bourne. Microbial indicators as a diagnostic tool for assessing water quality and climate stress in coral reef ecosystems. *Marine Biology*, 164(4):91, 2017.

[37] Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 267–288, 1996.

[38] M Guindi and RH Riddell. Indeterminate colitis. *Journal of clinical pathology*, 57(12): 1233–1244, 2004.

[39] Giovanni Musso, Roberto Gambino, and Maurizio Cassader. Interactions between gut microbiota and host metabolism predisposing to obesity and diabetes. *Annual review of medicine*, 62:361–380, 2011.

[40] Herbert Tilg and Alexander R Moschen. Microbiota and diabetes: an evolving relationship. *Gut*, 63(9):1513–1521, 2014.

[41] Peter J Turnbaugh, Fredrik Bäckhed, Lucinda Fulton, and Jeffrey I Gordon. Diet-induced obesity is linked to marked but reversible alterations in the mouse distal gut microbiome. *Cell Host & Microbe*, 3(4):213–223, 2008.

[42] Alan Herbert, Norman P Gerry, Matthew B McQueen, Iris M Heid, Arne Pfeufer, Thomas Illig, H-Erich Wichmann, Thomas Meitinger, David Hunter, Frank B Hu, et al.

A common genetic variant is associated with adult and childhood obesity. *Science*, 312 (5771):279–283, 2006.

[43] Yuval Benjamini and Terence P Speed. Summarizing and correcting the gc content bias in high-throughput sequencing. *Nucleic acids research*, page gks001, 2012.

# Chapter 3

# Model-based Dictionary Learning: Sparse Coding beyond Gaussian Independent Model

## 3.1 Introduction

Dictionary learning aims at decomposing an $m$ dimensional random vector as a linear combination of $K$ interpretable vectors, a collection of which is also referred to as a dictionary. Each vector in a dictionary is referred to as an atom [1; 2]. Comparing to the wavelet or kernel estimation methods which use predefined basis functions [3], dictionary learning is more interpretable and flexible, which leads to the state-of-the-art discoveries in numerous scientific fields such as neuroscience, genomics, artificial intelligence and astronomics [4–7]. Particularly, dictionary learning has been shown to perform remarkably better than many existing approaches in brain image analysis [8]. Thus, effective and efficient dictionary learning algorithms are highly desirable.

More than often, not all atoms in a dictionary are important to approximate a random vector. The coefficients of a linear combination are usually sparse. Let $||\cdot||_1$ denote the $L_1$ norm of a vector. Observing random vectors $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n \in \mathbb{R}^m$, a sparse dictionary learning model assumes

$$\mathbf{x}_i = D\boldsymbol{\alpha}_i + \boldsymbol{\epsilon} \quad 1 \leq i \leq n \tag{3.1}$$

with the constraint that $||\boldsymbol{\alpha}_i||_1$ is less than a constant $\rho$, where $\boldsymbol{\alpha}_i \in \mathbb{R}^K$ is called the sparse coefficient for $\mathbf{x}_i$. $D \in \mathbb{R}^{m \times K}$ is the dictionary matrix with each column representing an atom, and $\boldsymbol{\epsilon}$ is the random noise. The dictionary and sparse coefficients are usually obtained

by minimizing some empirical loss function

$$L(\boldsymbol{\alpha}_1, \cdots, \boldsymbol{\alpha}_n, D) = \frac{1}{n}\sum_{i=1}^{n}(\frac{1}{2}||\mathbf{x}_i - D\boldsymbol{\alpha}_i||^2 + \lambda||\boldsymbol{\alpha}_i||_1),$$

where $\lambda$ is a trade-off between the sparsity of $\boldsymbol{\alpha}_i$ and goodness of fit.

The dictionary learning algorithms have been studied over decades. Although these algorithms are very intuitive and empirically effective, their theoretical performance is difficult to study and establish, simply due to the non-convexity of loss function and the iterative nature of the algorithm. Thus, despite the urgent need, the focus of the dictionary learning research is mainly on algorithm development. The majority of existing algorithms follow the alternative estimation approach where $\boldsymbol{\alpha}_i$ and $D$ are estimated iteratively. A few popular methods along this line of thinking are K-SVD [1], online dictionary learning [2] and recursive least squares [9], which all first estimate $\boldsymbol{\alpha}_i$ by either matching pursuit algorithm[10] or orthogonal matching pursuit [11] algorithm, then update $D$ by coordinate descent approach.

Despite the appealing empirical performance, the dictionary learning algorithms are extremely computationally intensive. The computational cost mainly spends on the search of the appropriate tuning parameter $\lambda$. Although there are several discussions on how to select $\lambda$ using CV in the early literature [12–14], these proposals have not been developed to their fruition. In this manuscript, we establish the model-based dictionary learning algorithm which is essentially a stylized version of the traditional sparse coding algorithm that is based on grid search of $\lambda$. MDL is a generalization of the model-based clustering method to the sparse coding problem. We show that the existing dictionary learning method is a special case of the model-based dictionary learning method. In the MDL algorithm, only finite steps are required to produce sparse dictionary learning estimates, which notably improves the computational efficiency. Besides the computational improvement, we show that MDL algorithm converges to a stationary point theoretically.

Beyond the algorithm, MDL approach provides a rich and flexible framework to overcome limitations contained in the existing dictionary learning algorithms. For example, when $\mathbf{x}_i$

is on a discrete domain, we use a mixture of Poisson distributions to model the counts data. When observations have spatial or temporal correlations such as observations from image data or time course data, we can easily incorporate these correlations into dictionary learning algorithms by using an data-adaptive covariance structure. We make theoretical and methodological efforts to generalize the existing dictionary learning methods and have a broad impact on applications in many areas.

The rest of the article is organized as follows. Section 2 will formally propose the probabilistic sparse coding model. An algorithm based on the model and discussion of the opportunities and challenges raised by the algorithm will be discussed in Section 3. Theoretical properties of the proposed algorithm will be discussed in Section 4. Simulations and applications will be collected in Section 5.

## 3.2 Model Set-up

Assume $\mathbf{x}_i$ follows a mixture distribution, of which each component is distributed from $f(\mathbf{x}_i|\theta_{ij})$ for $j = 1, \cdots, J$. The fractions of each component are $(\pi_1, \cdots, \pi_J)$. Notation-wise, we write

$$\mathbf{x}_i \sim \pi_1 f(\mathbf{x}_i|\theta_{i1}) + \cdots + \pi_J f(\mathbf{x}_i|\theta_{iJ}). \tag{3.2}$$

A binary membership labeling variable $z_{ij}$ for observation $i$ can be introduced such that $\mathbf{x}_i|z_{ij} = 1 \sim f(\mathbf{x}_i|\theta_{ij})$. In general, $\theta_{ij}$ is not estimable, as there are more parameters than observations. However, when $\theta_{ij}$ satisfies some sparsity constraints, the number of parameters will significantly reduce and model (3.2) is estimable.

Model (3.2) is very general where many popular models can be considered as its special cases. For example, if $J = 1$ and $f(\mathbf{x}_i|\theta_i)$ is the density function of a Gaussian distribution with mean $D\boldsymbol{\alpha}_i$ and covariance matrix $\sigma^2 I$, where $D \in \mathbb{R}^{m\times K}$ and $\boldsymbol{\alpha}_i \in \mathbb{R}^K$, we can show that model (3.2) is equivalent to the conventional dictionary learning model.

If we further assume $||\boldsymbol{\alpha}_i||_1 \leq \rho$ or equivalently $\boldsymbol{\alpha}_i$ is sparse, only a subset of atoms in a dictionary is involved in $\theta_i$. For example, if the $\ell$th entry of $\boldsymbol{\alpha}_i$ is zero, atom $\ell$ is excluded from estimating $\theta_i$. Notice that each element in $\boldsymbol{\alpha}_i$ can be classified into two categories: zero or nonzero. Thus, we have in total $2^K$ indicators denoted as $\{\boldsymbol{\gamma}_j\}_{j=1}^J$ in which $\boldsymbol{\gamma}_j$ is the index for a unique subset of parameters in the $j$th component. In another word, we can use a $2^K$ factorial design to quantify all the possible combinations of atoms and generate all possible sparse representations of a dictionary $D$. The only difference between a $2^K$ design matrix and $\{\boldsymbol{\gamma}_j\}_{j=1}^J$ is that we use $-1$ to represent that a factor is used as a control in a design matrix, while under dictionary learning, we use 0 instead of $-1$ when an atom is not included in a dictionary.

Heuristically, (3.2) can be casted as a probabilistic version of the dictionary learning model (3.1) for certain choice of $\rho$. To further illustrate how to use model (3.2) for sparse dictionary learning, we use a simple toy example with $K = 2$. Model (3.2) can be rewritten as

$$\mathbf{x}_i \sim \pi_1 N(\mathbf{x}_i|D\begin{pmatrix}0\\0\end{pmatrix}, \Sigma_i) + \pi_2 N(\mathbf{x}_i|D\begin{pmatrix}0\\[\boldsymbol{\alpha}_{i2}]_2\end{pmatrix}, \Sigma_i) + \pi_3 N(\mathbf{x}_i|D\begin{pmatrix}[\boldsymbol{\alpha}_{i3}]_1\\0\end{pmatrix}, \Sigma_i) + \pi_4 N(\mathbf{x}_i|D(\begin{pmatrix}[\boldsymbol{\alpha}_{i4}]_1\\[\boldsymbol{\alpha}_{i4}]_2\end{pmatrix}, \Sigma_i).$$

where $\boldsymbol{\alpha}_{ij} \in \mathbb{R}^2$ for $j = 1, \ldots, 4$, and $[\cdot]_k$ defines an operator that takes the $k$th entry of a vector. Correspondingly, for an optimal sparse model, we do not need to estimate $\rho$ but only need to estimate $\pi_1$ to $\pi_J$ to see which one is larger. The component with the larger fraction is the optimal sparse dictionary learning model.

In the next subsection, we will first discuss the sparse dictionary learning when $f(\mathbf{x}_i|\theta_{ij})$ is a Gaussian density with covariance matrix $\sigma_i^2 I$. Then we will generalize the Gaussian mixture model to incorporate spatial correlations. Finally, we will generalize the Gaussian mixture results to Poisson mixture to model counts data.

### 3.2.1 Simple $d$-sparse Gaussian MDL

In practice, the sparsity in dictionary learning can only be achieved for certain value of $\lambda$ because you cannot control the maximum number of nonzero coefficients using a $L_1$ penalty. In order to have real sparsity, we propose the following $d$-sparse model which requires that at most $d$ atoms have nonzero coefficients in a dictionary learning model. The $d$-sparse dictionary learning model is equivalent to conventional dictionary learning model but only for some fixed value of $\lambda$. Mathematically, we can formulate the $d$-sparse dictionary learning as

$$\mathbf{x}_i \quad \sim \quad \pi_1 \mathcal{N}(\mathbf{x}_i | D\boldsymbol{\alpha}_{i1} \circ \boldsymbol{\gamma}_1, \sigma_i^2 I) + \cdots + \pi_J \mathcal{N}(\mathbf{x}_i | D\boldsymbol{\alpha}_{iJ} \circ \boldsymbol{\gamma}_J, \sigma_i^2 I), \quad (3.3)$$

where $\boldsymbol{\alpha}_{ij} \in \mathbb{R}^K$, $\boldsymbol{\gamma}_j = (\gamma_{j1}, \cdots, \gamma_{jK})'$ is a $K$-dimensional binary vector that controls which atoms are selected to learn $\mathbf{x}_i$. Here, we use $\circ$ to denote the Hadamard product. To achieve $d$-sparse, i.e., $\sum_{l=1}^K \gamma_{jl} \leq d$, we only need to reduce the choices of $\boldsymbol{\gamma}_j$ from $2^K$ elements to $\sum_{\ell=1}^d \binom{K}{\ell}$ elements. It is easy to see that model (3.3) exhaustively include all possible $d$-sparse combinations of atoms, where each combination is referred to a specific component in (3.3).

### 3.2.2 Spatial $d$-sparse Gaussian MDL

Now let us turn our attention to some applications such as estimating the functional brain network using fMRI data and image denoising, both of which have significant spatial correlations between entries of $\mathbf{x}_i$s. For this type of application, we need to incorporate the spatial correlation into model (3.3). With a little abuse of notation, we let $\mathbf{x}_i = \{\mathbf{x}_i(s_\ell)\}$ denote observations measured at $s_\ell, \ell = 1, \cdots, m$ where $s_\ell \in \mathbb{R}^p$. We then assume that

$$\mathbf{x}_i \quad \sim \quad \pi_1 \mathcal{N}(\mathbf{x}_i | D\boldsymbol{\alpha}_{i1} \circ \boldsymbol{\gamma}_1, \Sigma_i) + \cdots + \pi_J \mathcal{N}(\mathbf{x}_i | D\boldsymbol{\alpha}_{iJ} \circ \boldsymbol{\gamma}_J, \Sigma_i), \quad (3.4)$$

where $\Sigma_i$ is the covariance matrix with the $\ell\ell'$th entry gives the covariance between $\mathbf{x}_i(s_\ell)$ and $\mathbf{x}_i(s_{\ell'})$. When $p = 1$, we use $\Sigma_i$ model the temporal covariance. When $p = 2$, we use

$\Sigma_i$ to model the spatial covariance. In general, we assume the covariance between random variables at two time points or two locations depends on the time lag or their inter-location distance $\Delta_{\ell\ell'}$. The most popular temporal covariance is the auto-correlated covariance which assumes that $\{\Sigma_i\}_{\ell\ell'} = \sigma_i^2 \omega_i^{-\Delta_{\ell\ell'}}$. For spatial covariance, a few popular models include the exponential model which assumes that $\{\Sigma_i\}_{\ell\ell'} = \sigma_i^2 \exp(-\omega_i \Delta_{\ell\ell'})$ and the Gaussian model which assumes $\{\Sigma_i\}_{\ell\ell'} = \sigma_i^2 \exp(-\omega_i \Delta_{\ell\ell'}^2)$.

### 3.2.3 Exponential family $d$-sparse MDL

In general, we assume that $f(\mathbf{x}_i|\theta_{ij})$ is a density in exponential family, i.e.,

$$\mathbf{x}_i|z_{ij} = 1 \sim h(\mathbf{x}_i, \phi_i) \exp\{\frac{\eta_{ij}' \mathbf{x}_i - A(\eta_{ij})}{c(\phi_i)}\}, \tag{3.5}$$

where $\eta_{ij}$ is a function of the mean, $c(\phi_i)$ is the dispersion parameter and $A(\eta_{ij})$ is the cumulant function. It is clear that both Gaussian distribution and Poisson distribution belong to the exponential family. When $\mathbf{x}_i$ has an independent normal distribution, $\eta_{ij}$ is the mean of $\mathbf{x}_i$ and $c(\phi_i) = \sigma_i^2$. For Poisson and binomial models without over-dispersion, we have $\phi_i = 1$. When $\mathbf{x}_i$ follows a Poisson distribution, we let $\eta_{ij}$ be the logarithm of its mean.

To achieve $d$-sparse dictionary learning, we further assume that $\eta_{ij}$ has the following decomposition, i.e.,

$$\eta_{ij} = D\boldsymbol{\alpha}_{ij} \circ \boldsymbol{\gamma}_j$$

for a given $\boldsymbol{\gamma}_j$. This family of distributions has broad applications in many scientific studies such as the RNA-seq analysis where $\mathbf{x}_i$ is on a discrete domain or network deconvolution where $\mathbf{x}_i$ is a binary vector. It can greatly broaden the application of sparse dictionary learning algorithm. For example, we use the exponential family $d$-sparse MDL to find cancer tissue related genetic signatures.

Let $\theta_{ij}$ denote the collection of all parameters for the $i$th observation and the $j$th component. In a simple $d$-sparse Gaussian MDL, $\theta_{ij} = \{D, \pi_j, \boldsymbol{\alpha}_{ij}, \sigma_i\}$ and in an exponential family $d$-sparse MDL, $\theta_{ij} = \{D, \pi_j, \boldsymbol{\alpha}_{ij}, \phi_i\}$. In spatial $d$-sparse Gaussian MDL, we usually assume $\Sigma_i = \sigma_i^2 R(\omega_i)$, where $R(\omega_i)$ is the correlation matrix related to spatial or temporal correlation structures. Correspondingly, $\theta_{ij} = \{D, \pi_j, \boldsymbol{\alpha}_{ij}, \sigma_i, \omega_i\}$. Observing $\mathbf{X} = (\mathbf{x}_1, \cdots, \mathbf{x}_n)$, the the likelihood function is

$$L(\theta_{11}, \cdots, \theta_{nJ}|\mathbf{X}) = \prod_{i=1}^{n} \sum_{j=1}^{J} \pi_j f(\mathbf{x}_i|\theta_{ij}).$$

Let $Z$ be a $n \times J$ model-labeling matrix with the $ij$th entry $z_{ij}$. Then, the complete likelihood function is

$$L(\theta_{11}, \cdots, \theta_{nJ}|\mathbf{X}, Z) = \prod_{i=1}^{n} \prod_{j=1}^{J} \big(\pi_j f(\mathbf{x}_i|\theta_{ij})\big)^{z_{ij}}, \tag{3.6}$$

and the log-likelihood of complete data is,

$$\ell(\theta|X, Z) = \sum_{i=1}^{n} \sum_{j=1}^{J} z_{ij}\big(\log \pi_j + \log f(\mathbf{x}_i|\theta_{ij})\big). \tag{3.7}$$

EM algorithm is one of the most common tools for the estimation in mixture models. It has been extensively discussed in the statistical literature. There are a large amount of EM variants that have been proposed to facilitate the computation. A few examples include the rejection-control EM [15], stochastic EM [16] and classification EM [17]. The classical EM algorithm has two steps: E step, which computes the expectation of the complete-data log-likelihood function (3.7) based on the parameters estimated in the $t$th iteration, i.e.,

$$Q(\theta|\theta^{(t)}) = \mathbb{E}_{Z|\mathbf{X},\theta^{(t)}} \ell(\theta|Z, \mathbf{X});$$

and M step, in which we found $\theta^{(t+1)}$ by maximizing the function $Q(\theta|\theta^{(t)})$. In terms of the sparse coding, the E step is the computation of

$$w_{ij} = \frac{\pi_j^{(t)} f(\mathbf{x}_i|\theta_{ij}^{(t)})}{\sum_{j=1}^{J} \pi_j^{(t)} f(\mathbf{x}_i|\theta_{ij}^{(t)})} \tag{3.8}$$

and the M-step involves maximization of

$$Q(\theta|\theta^{(t)}) = \sum_{i=1}^{n}\sum_{j=1}^{J} w_{ij}\log\pi_j \tag{3.9}$$

$$+\sum_{i=1}^{n}\sum_{j=1}^{J} w_{ij}\log f(\mathbf{x}_i|\theta_{ij}) \tag{3.10}$$

with respect to $\pi_j$ and $\theta_{ij}$, where $Q(\theta|\theta^{(t)})$ is generated by replacing $z_{ij}$ in (3.7) by the $w_{ij}$ obtained from E step. Clearly, maximizing $Q(\theta|\theta^{(t)})$ is equivalent to maximizing (3.9) with respect to $\pi_j$ and maximizing (3.10) with respect to $\theta_{ij}$ separately. Maximizing (3.9) leads to

$$\pi_j^{(t+1)} = \frac{1}{n}\sum_{i=1}^{n} w_{ij}.$$

### 3.3.1 Update dictionary for Gaussian distribution

Notice that for (simple or spatial ) $d$-sparse Gaussian MDL model, we rewrite (3.10) as

$$-\sum_{i=1}^{n}\sum_{j=1}^{J}\frac{w_{ij}}{2}\big((\mathbf{x}_i - D\boldsymbol{\alpha}_{ij}\circ\boldsymbol{\gamma}_j)'\Sigma_i^{-1}(\mathbf{x}_i - D\boldsymbol{\alpha}_{ij}\circ\boldsymbol{\gamma}_j) + \frac{m\log 2\pi + \log|\Sigma_i|}{2}\big). \tag{3.11}$$

**Update the variance:** For spatial $d$-sparse Gaussian MDL model we have $\Sigma_i = \sigma_i^2 R(\omega_i)$. Let $\boldsymbol{\eta}_{ij}^{(t)} = D^{(t)}\boldsymbol{\alpha}_{ij}^{(t)}\circ\boldsymbol{\gamma}_j$, where $D^{(t)}$ and $\boldsymbol{\alpha}_{ij}^{(t)}$ are the current estimate of $D$ and $\boldsymbol{\alpha}_{ij}$ respectively. Given $D^{(t)}$, $\boldsymbol{\alpha}_{ij}^{(t)}$ and $\omega_i^{(t)}$, maximizing (3.11) with respect to $\sigma_i^2$ leads to an updated estimate of $\sigma_i^2$, which is

$$\sigma_i^{(t+1)^2} = \frac{\sum_{j=1}^{J} w_{ij}(\mathbf{x}_i - \boldsymbol{\eta}_{ij}^{(t)})'R^{-1}(\omega_i^{(t)})(\mathbf{x}_i - \boldsymbol{\eta}_{ij}^{(t)})}{m},$$

and for simple $d$-sparse model where $R(\omega_i) = I$, we can update $\sigma_i^2$ by

$$\sigma_i^{2(t+1)} = \frac{1}{m}\sum_{j=1}^{J} w_{ij}(\mathbf{x}_i - \boldsymbol{\eta}_{ij}^{(t)})'(\mathbf{x}_i - \boldsymbol{\eta}_{ij}^{(t)}).$$

**Update dictionary and its coefficients:** Notice that $D\boldsymbol{\alpha}_{ij}\circ\boldsymbol{\gamma}_j = \boldsymbol{\gamma}_j \otimes D\boldsymbol{\alpha}_{ij}$. Let $D_j = [[\boldsymbol{\gamma}_j \otimes D]]$ be the sparse dictionary for the $j$th component and $D_j^{(t)} = [[\boldsymbol{\gamma}_j \otimes D^{(t)}]]$, where $[[\cdot]]$ defines an operator that only taking nonzero columns of a matrix. Given $\Sigma_i^{(t)}$,

maximizing (3.11) is equivalent to solving a weighted least square regression which leads to an estimate of $\boldsymbol{\alpha}_{ij}$, i.e.,

$$\boldsymbol{\alpha}_{ij}^{(t+1)} = \big(D_j^{(t)'}\Omega_i^{(t)}D_j^{(t)}\big)^{-1}D_j^{(t)'}\Omega_i^{(t)}\mathbf{x}_i,$$

where $\Omega_i^{(t)}$ is the inverse matrix of $\Sigma_i^{(t)}$ and is usually referred to as the precision matrix.

For simple sparse coding model, $\Omega_i^{(t)} = \frac{1}{\sigma_i^{2(t)}}I$ and $\boldsymbol{\alpha}_{ij}^{(t+1)}$ has a form of a conventional least square estimate, i.e.,

$$\boldsymbol{\alpha}_{ij}^{(t+1)} = \big(D_j^{(t)'}D_j^{(t)}\big)^{-1}D_j^{(t)'}\mathbf{x}_i.$$

Notice that $|\boldsymbol{\gamma}_j|_1 \leq d$, which implies that $\boldsymbol{\alpha}_{ij}^{(t+1)}$ is at most of dimension $d$ but not dimension $K$ as $\boldsymbol{\alpha}_{ij}$. Thus, we need to transform $\boldsymbol{\alpha}_{ij}^{(t+1)}$ to $K$ dimensional vector to generate final estimate of $\boldsymbol{\alpha}_{ij}$ denoted by $\boldsymbol{\alpha}_{ij}^{*(t+1)}$. We fill in entries of $\boldsymbol{\alpha}_{ij}^{*(t+1)}$ by zero if the corresponding entries of $\boldsymbol{\gamma}_j$ is zero and $\boldsymbol{\alpha}_{ij}^{(t+1)}$ otherwise. To ease the description, we still use $\boldsymbol{\alpha}_{ij}^{(t+1)}$ to denote the $\boldsymbol{\alpha}_{ij}^{*(t+1)}$ in the subsequent updates.

Next, we sequentially update each column of $D$ for given $\boldsymbol{\alpha}_{ij}^{(t+1)}$ and $\Sigma_i^{(t)}$ by using a block coordinate descent algorithm. Let $\mathbf{d}_k$ denote the $k$th column of the dictionary matrix $D$. Let $c_{ijk} = [\boldsymbol{\alpha}_{ij}^{(t+1)} \circ \boldsymbol{\gamma}_j]_k$. Now given $\mathbf{d}_k^{(t)}$, $\boldsymbol{\alpha}_{ij}^{(t+1)}$ and $\Sigma_i^{(t)}$, we can update $\mathbf{d}_k$ by

$$\mathbf{d}_k^{(t+1)} = M^{-1}\sum_{i=1}^{n}\sum_{j=1}^{J}M_{ijk}^{*}\big(x_i - c_{ijk}\mathbf{d}_{-k}\big),$$

where $M = \sum_{i=1}^{n}\sum_{j=1}^{J}w_{ij}c_{ijk}^2\Omega_i^{(t)}$, $M_{ijk}^{*} = w_{ij}c_{ijk}\Omega_i^{(t)}$ and $\mathbf{d}_{-k} = \sum_{l<k}\mathbf{d}_l^{(t+1)} + \sum_{l>k}\mathbf{d}_l^{(t)}$. Notice that $\Omega_i^{(t)} = R^{-1}(\omega_i^{(t)})/\sigma_i^{(t)}$, where $R(\omega_i^{(t)})$ quantifies the spatial correlations.

For the simple sparse coding model where $R(\omega_i^{(t)}) = I$, we can update $\mathbf{d}_k$ by the simple form

$$\mathbf{d}_k^{(t+1)} = \sum_{i=1}^{n}\sum_{j=1}^{J}\nu_{ijk}\big(x_i - c_{ijk}\mathbf{d}_{-k}\big),$$

where $\nu_{ijk} = \frac{w_{ij}c_{ijk}}{\sum_{i=1}^{n}\sum_{j=1}^{J}w_{ij}c_{ijk}^2}$.

**Update the spatial correlation parameter:** The final update in M-step is to maximize (3.11) with respect to the spatial hyper-parameter $\omega_i$ based on the updated $D$, $\boldsymbol{\alpha}_{ij}$

and $\sigma^2$. Newton-Raphson is the most popular algorithm for this type of minimization. Given $D^{(t+1)}$, $\alpha^{(t+1)}$, $\sigma_i^{(t+1)}$, we can recursively update $\omega_i$ by

$$\omega_i^{(t+1)} = \omega_i^{(t)} + b^{-1} f$$

where $b = \partial^2 Q(\theta|\theta^{(t)})/\partial^2 \omega_i$ and $f = \partial Q(\theta|\theta^{(t)})/\partial \omega_i$.

### 3.3.2 Update Dictionary for distribution in exponential family

**Update $\boldsymbol{\alpha}_{ij}$:** In general, if $\mathbf{x}_i|z_{ij} = 1$ follows a distribution in exponential family (3.5) with known dispersion parameter $c(\phi)$, the term (3.10) can be rewritten as

$$\sum_{i=1}^{n}\sum_{j=1}^{J} \frac{w_{ij}}{c(\phi_i)}\big(\boldsymbol{\eta}_{ij}'\mathbf{x}_i - A(\boldsymbol{\eta}_{ij})\big), \tag{3.12}$$

where $\boldsymbol{\eta}_{ij} = D\boldsymbol{\alpha}_{ij} \circ \boldsymbol{\gamma}_j$. Using chain rule, the maximizer of (3.12) with respect to $\boldsymbol{\alpha}_{ij}$ has the form

$$\boldsymbol{\alpha}_{ij}^{(t+1)} = (D_j^{(t)\prime} W^{(t)} D_j^{(t)})^{-1} D_j^{(t)\prime} W^{(t)} \mathbf{x}_i^*, \tag{3.13}$$

where $W^{(t)}$ is a $m \times m$ diagonal matrix with $l$th diagonal entry

$$W_{ll}^{(t)} = \frac{1}{c(\phi)\big(\partial^2 A(\boldsymbol{\eta}_{ij}^{(t)})/\partial^2 [\boldsymbol{\eta}_{ij}]_l\big)\big(\partial g([\boldsymbol{\eta}_{ij}^{(t)}]_l)/\partial [\boldsymbol{\eta}_{ij}]_l\big)^2}$$

where $g$ is the inverse link function, and $\mathbf{x}_i^* = \boldsymbol{\eta}_{ij}^{(t)} + B^{-1}(\mathbf{x}_i - g(\boldsymbol{\eta}_{ij}^{(t)}))$ of which $B$ is a $m \times m$ diagonal matrix with $l$th diagonal entry $\partial g([\boldsymbol{\eta}_{ij}^{(t)}]_l)/\partial [\boldsymbol{\eta}_{ij}]_l$. For Poisson distribution, the inverse link function is $g = \exp(\cdot)$. For binomial distribution, the inverse link function is $g(\cdot) = \exp(\cdot)/(1 + \exp(\cdot))$.

**Update D:** For general distributions of exponential family, the explicit form of $D$ is hard to obtain since the inverse link function is nonlinear. In practice, we use the gradient ascent algorithm to update $\mathbf{d}_k$ by

$$\mathbf{d}_k^{(t+1)} = \mathbf{d}_k^{(t)} + \tau U_{\mathbf{d}_k},$$

where $U_{\mathbf{d}_k}$ is the score function with respect to $\mathbf{d}_k$, and $\tau$ is the step size. We use Barzilai-Borwein method to choose a proper $\tau$ as

$$\tau = \frac{(\mathbf{d}_k^{(t)} - \mathbf{d}_k^{(t-1)})^T (U_{\mathbf{d}_k^{(t)}} - U_{\mathbf{d}_k^{(t-1)}})}{||U_{\mathbf{d}_k^{(t)}} - U_{\mathbf{d}_k^{(t-1)}}||^2}.$$

Since our algorithm uses the value of $\mathbf{d}_k^{(t)}$ for computing $\mathbf{d}_k^{(t+1)}$, a single iteration has empirically been found to be enough. As in Gaussian sparse coding, we do not need to evaluate all the possible $d$-sparse combination to estimate $\mathbf{d}_k^{(t)}$. We only need to focus on the combination containing the $k$th atom.

### 3.3.3 Rejection-control EM

With the exponentially increasing number of components, the E-step results in a huge number of $w_{ij}$s, many of which are extremely small. The small $w_{ij}$ can make the optimization in M-step very inefficient, unstable and sometimes even infeasible. To alleviate the computation and stabilize the algorithm, we use the rejection-control EM that is proposed in [15] instead of EM for our MDL algorithms. Let a small number $c$ (e.g., $c = 0.01$) be a rejection-control threshold. We approximate $z_{ij}$ by

$$w_{ij}^* = \begin{cases} w_{ij} & \textbf{\textit{if}}\ w_{ij} > c \quad \textbf{\textit{else}} \\ c & \text{with probability } w_{ij}/c \\ 0 & \text{with probability } 1 - w_{ij}/c \end{cases}$$

for $j = 1, \cdots, J$.

### 3.3.4 Information Criteria for selecting $d$

The sparsity of the traditional sparse coding method is controlled by the $||\boldsymbol{\alpha}_i||_1 \leq \rho$ where $\rho \in (0, \infty)$. A commonly used method is to set some grid points in some bounded interval $(0, c)$ and search the optimal estimation by cross validation. There are two difficulties for this

approach: the number of grid points is large; cross validation is computationally infeasible for the large data set. These problems can be solved by using the MDL method, which provides a framework for model comparison.

The Bayesian information criterion (BIC) [18] has been shown to be an effective method for model comparison in mixture models [19; 20]. The minimizer of BIC can well balance the model complexity and goodness-of-fit. In this article, we propose to use

$$BIC(d) = -2\ell(\hat{\theta}|X) + p(d)\log(n \times m) \tag{3.14}$$

where $\hat{\theta}$ is the final estimated parameters and $p(d)$ is the number of parameters to be estimated in the model. We set $p_{si}(d) = mK+2n+n\sum_{l=1}^{d} l\binom{K}{l}$, $p_{sp}(d) = mK+3n+n\sum_{l=1}^{d} l\binom{K}{l}$ and $p_{ex}(d) = mK + n + n\sum_{l=1}^{d} l\binom{K}{l}$ as the number of parameters for simple, spatial and exponential family $d$-sparse models respectively.

## 3.4 Convergence Analysis

The convergence of EM algorithm has been well studied in [21; 22]. The main difference of the $d$-sparse EM algorithm lies in the M-step where we update the $D$ and $\beta$ alternatively. In the following theorem, we show that MDL algorithm also converges as the traditional EM algorithm.

**Theorem 3.4.1.** *(Convergence to a stationary point). Let $\{\theta^{(t)}\}$ be an instance of a d-sparse MDL algorithm, then all the limiting points of $\theta^{(t)}$ are stationary points, denoted as $\theta^*$. Then we have $L(\theta^{(t)})$ converges monotonically to $L(\theta^*)$.*

*Sketch of the proof*, we are aimed to show the monotonicity of the likelihood function in the iterations, that is,

$$\log\big(L(\theta;X)\big) - \log(L(\theta^{(t)};X)\big) = \big(Q(\theta;\theta^{(t)}) - Q(\theta^{(t)};\theta^{(t)})\big) - \big(H(\theta;\theta^{(t)}) - H(\theta^{(t)};\theta^{(t)})\big)$$

where $H(\theta;\theta^{(t)}) = E_{(\theta^{(t)})}\big(\log k(x|z,\theta)|y\big)$, $k(x|z,\theta) = L(\theta;X)/L(\theta;X,Z)$. By the proof of theorem 1 in [22], $H(\theta;\theta^{(t)}) - H(\theta^{(t)};\theta^{(t)}) < 0$. In the M-step, we update the coefficients $\boldsymbol{\alpha}_{ij}$ and dictionary $D$ alternatively. First, we use the classical generalized linear model for computing the decomposition of $\mathbf{x}_i$ over the dictionary. The uniqueness of $\alpha_{ij}$ and increment of $Q$ function are guaranteed for the this step. Next, the new dictionary $D$ is computed by column-wise, which ensures the nondecreasing of the $Q$ function. Combining this two steps, we have the non-decreasing property for the $Q$ function. Thus, we have the likelihood function converge monotonically to some value $L^*$.

## 3.5 Empirical Studies

### 3.5.1 Gaussian $d$-sparse model

**SIMULATION:** Fifty mock data was simulated to compare the empirical performance of our MDL methods with the popular dictionary learning methods such as K-SVD and online dictionary learning. We generated each signal $\boldsymbol{x}_i \in \mathbb{R}^m$, where $m = 100$, from a mixture of normal distribution $\sum_{j=1}^{J} \pi_j \mathcal{N}(D\boldsymbol{\alpha}_{ij} \circ \boldsymbol{\gamma}_j, \sigma^2 R(\omega))$, $i = 1,\ldots,n$. Each element of the dictionary matrix $D \in \mathbb{R}^{m\times K}$, where $K = 30$, was fixed realization from Uniform$[0,1]$, and every column of $D$ was normalized. Each element of $\boldsymbol{\alpha}_{ij}$ was generated from Uniform$[1,10]$ and was kept fixed once generated. The weight $\pi_j$ was set to $1/J$, where $J = \sum_{l=1}^{2}\binom{30}{l}$, for $j = 1,\ldots,J$. The spatial locations of signals were randomly realized from the $[0,100]^2$ spatial domain. The exponential correlation function with $\omega = 1/25$ was employed to model the spatial correlation of signals. We set $\sigma^2 = ||D\boldsymbol{\alpha}_{ij} \circ \boldsymbol{\gamma}_j||_2/\text{SNR}$, where SNR $= 2,3,4,5$ respectively.

The spatial $d$-sparse Gaussian MDL (sp-MDL) algorithm, the simple $d$-sparse Gaussian MDL (si-MDL) algorithm, the online dictionary learning (Online) algorithm, and the K-SVD algorithm were implemented on the synthetic signals with sample size $n$ varying from 100 to 500. In each plot of Figure 3.1, we drew boxplots of the distance between the space spanned

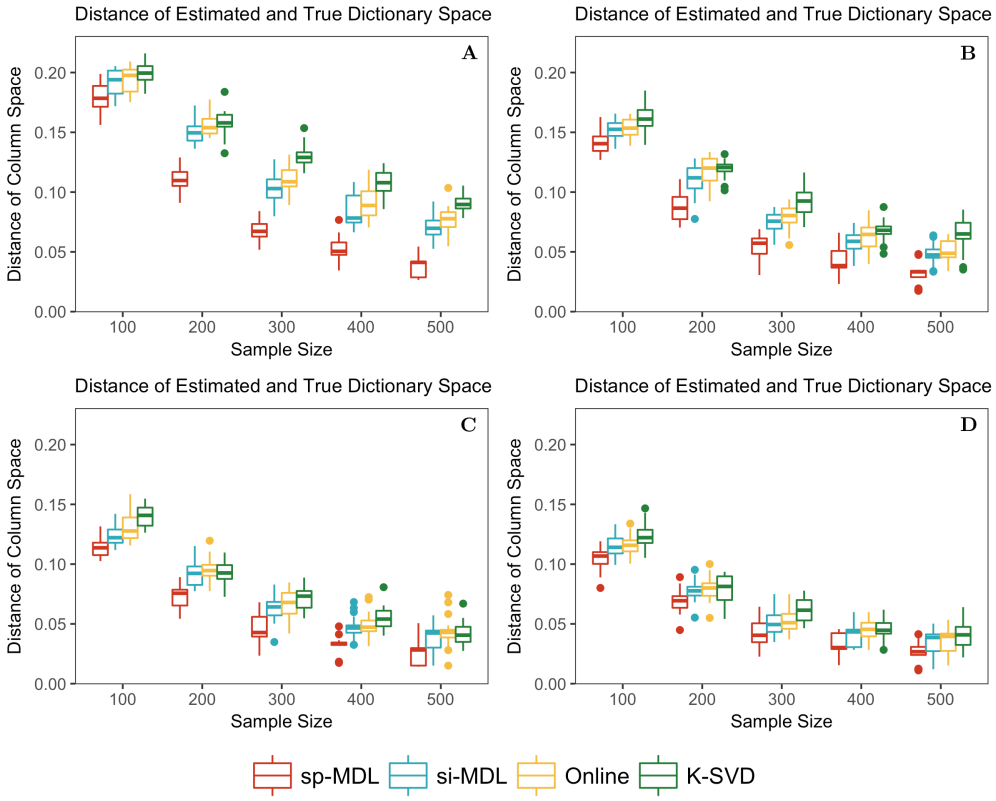



**Figure 3.1:** Simulation results for Gaussian $d$-sparse model with SNR $= 2, 3, 4, 5$ corresponding to Figure A, B, C, D respectively. The spatial $d$-sparse Gaussian MDL (sp-MDL) algorithm achieved the smallest distance between estimated and true dictionary space, compared with the simple $d$-sparse MDL (si-MDL) algorithm, the online dictionary learning (Online) algorithm, and the K-SVD algorithm.

by the estimated dictionary and the space spanned by the true dictionary for different sample sizes. Clearly, the sp-MDL algorithm outperformed other algorithms both in terms of average distance and in terms of standard deviation. Moreover, as we increased the noise level, the sp-MDL algorithm had a significant advantage over other comparable algorithms, which implies that the sp-MDL algorithm is especially useful for noisy data.

**APPLICATION (Image denoising):** In this example, Five 128pixel$\times$128pixel images were used for denoising. Noise from a Gaussian random field with covariance function


**Figure 3.2:** Plotted in the columns are generated images (first column), donoised images (2-5 columns) using K-SVD, online dictionary learning (Online), simple $d$-sparse Gaussian MDL (si-MDL) and spatial $d$-sparse Gaussian MDL (sp-MDL) respectively.

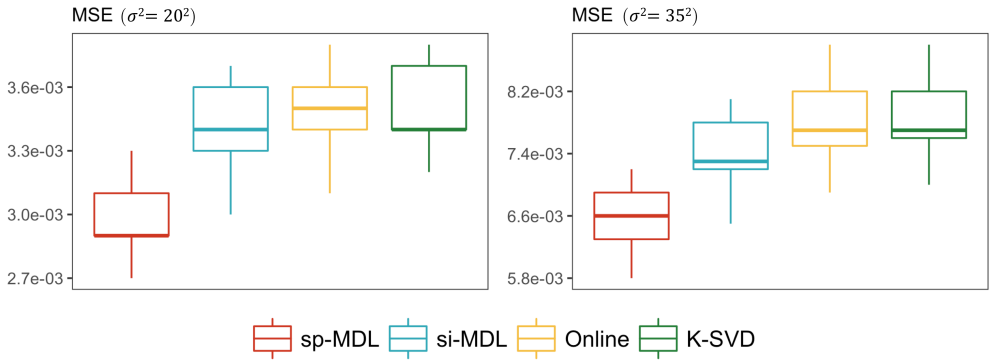



**Figure 3.3:** Plotted here are the MSE of denoised images using the spatial $d$-sparse Gaussian MDL (sp-MDL) algorithm, the simple $d$-sparse Gaussian MDL (si-MDL) algorithm, the online dictionary learning (Online) algorithm and the K-SVD algorithm.

$\{\Sigma_i\}_{\ell\ell'} = \sigma^2 \exp(-1/4\Delta_{\ell\ell'})$ was artificially added to the raw images. We compared the denoised images at $\sigma^2 = 20^2$ and $\sigma^2 = 35^2$ and plotted them in the first and second row of Figure 3.2 respectively. We have $n = 1600$ overlapping blocks with intensity of which were stretched as a $m = 144$(12pixel $\times$ 12pixel) dimensional vector. Clearly each entry in the vector were spatially correlated.

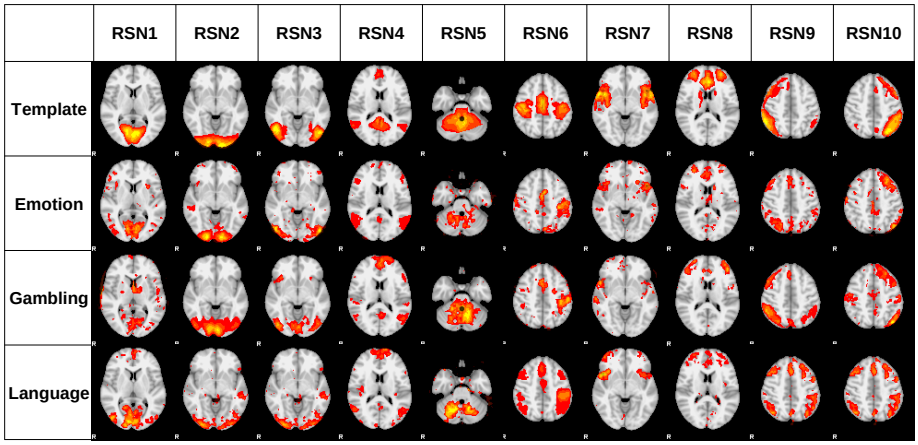



**Figure 3.4:** Ten Resting-state network (RSN 1-10) identified by sp-MDL algorithm.

Plotted in Figure 3.3 is the mean squared error (MSE) of the denoised images with $K = 64$ and $d = 2$ for spatial $d$-sparse Gaussian MDL (sp-MDL) and simple $d$-sparse Gaussian MDL (si-MDL), and with default setting for online dictionary learning as well as K-SVD. Clearly, sp-MDL significantly outperformed comparable algorithms in terms of estimation error. Both sp-MDL and si-MDL outperformed the existing algorithms, where the sp-MDL has a better estimation performance for spatially correlated data.

**APPLICATION (Brain connectivity study use fMRI data):** Understanding the organizational architecture of human brain function has been of intense interest since the inception of human neuroscience. After decades of active research using in-vivo functional neuroimaging techniques such as fMRI, there are accumulating evidence that human brain function emerges from and is realized by the interaction of multiple concurrent neural processes or networks, each of which is spatially distributed across the specific structural substrate of neuroanatomical areas. Although this discovery holds a lot of promise on constructing the concurrent functional networks and network-level interactions robustly and faithfully at the whole population level, the delivery of this promise, however, has not yet

been fully materialized, due to the lack of effective and efficient analytical tools for handling huge and complicated brain image data. Thus, it is largely unknown to what extent those multiple interacting functional networks spatially overlap with each other and jointly generate the total brain function.

To answer this question, we applied the proposed spatial $d$-sparse Gaussian MDL on the Human Connectome Project (HCP) Q1 release functional magnetic resonance imaging (fMRI) data. Three tasks ("Emotion", "Gambling", and "Language" ), each including fMRI images of 5 subjects, were selected to demonstrate how MDL can help understanding the human brain connectivity. The data was preprocessed using FSL [23]. Using $K = 100$ and Gaussian spatial correlation function $\{\Sigma_i\}_{\ell\ell'} = \sigma_i^2 \exp(-\omega_i \Delta_{\ell\ell'}^2)$, we found the optimal sparse level $d = 4$ based on the BIC criteria (3.14).

For each task, we mapped the learned atom $\{\mathbf{d}_k\}_{k=1}^{50}$ in the dictionary on the brain and compared them with the resting state networks (RSNs) [24; 25]. The intrinsic RSNs has also been observed in task-based fMRI data [24; 26]. As shown in Figure 3.4, we found the learned networks corresponding to the (RSNs 1-10) using task-based fMRI data. RSN 1-3 mainly include the visual cortex; RNS 4 is often referred to as the default mode network; RNS 5 covers the cerebellum; RSN6 dominantly features sensor-motor network; RSN7 covers the auditory network; RSN8 covers the executive control network; the symmetric RSN9 and RSN10 cover the left and right middle frontal, orbital and superior parietal areas, while for certain tasks (e.g. Language) it is observed that RSN9 and RSN10 will merge into the same network.

### 3.5.2 Poisson $d$-sparse Model

**SIMULATION:** Fifty mock data was simulated from a mixture of Poisson distribution $\sum_{j=1}^{J} \pi_j \text{Poisson}(\theta_{ij})$, where $\log \theta_{ij} = D\boldsymbol{\alpha}_{ij} \circ \boldsymbol{\gamma}_j$ with $D \in \mathbb{R}^{100\times 10}$. The dictionary $D$, coefficient $\boldsymbol{\alpha}_{ij}$ and weight $\pi_j$ was generated in the same way as the Gaussian $d$-sparse simulation.

The exponential family $d$-sparse MDL (ex-MDL) algorithm, the online dictionary learning (Online) algorithm, and the K-SVD algorithm were implemented on the synthetic signals with sample size $n$ varying from 100 to 500. The distance between the column space of an estimated dictionary and the column space of the true $D$ was plotted in Figure 3.5. Our exponential $d$-sparse MDL algorithm provided a significantly better estimate of the true dictionary.

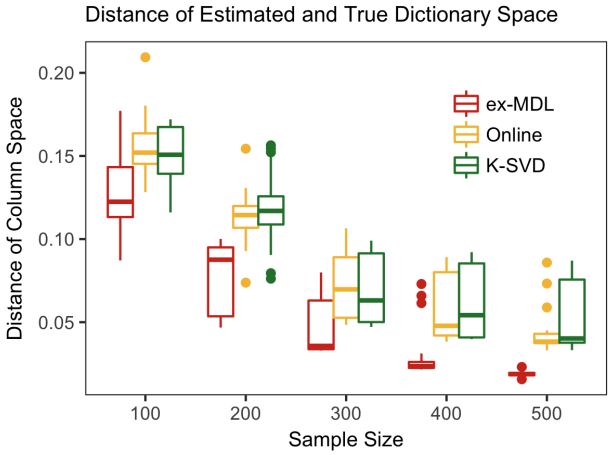



**Figure 3.5:** Estimation error of data with Poisson distribution.

**APPLICATION (Spatial Transcriptomic imaging for Breast Cancer Data):** In this real data analysis, we applied the exponential family $d$-sparse MDL (ex-MDL) algorithm to a breast cancer spatial transcriptomics dataset [27]. Spatial transcriptomics is a recent sequencing strategy that quantifies the gene expression within a tissue section with two-dimensional positional information. In the dataset, the sequenced reads are aligned to reference genome to count the number of reads mapped to a specific gene. We select 1573 genes with reads count larger than 100 at 254 locations in a histological section of a breast cancer biopsy including the invasive cancer areas, the cancer in situ areas, as well as the non-cancer areas. The input data vectors $\mathbf{x}_i \in \mathbb{R}^m$ are the mapped reads count at $m$ locations in a histological section for $i = 1, \ldots, n$ where $n$ is the number of selected genes.

We applied the ex-MDL algorithm to this dataset with dictionary size $K = 10$. The optimal $d$ was set as two based on the BIC criteria. The learned atoms were mapped back to the tissue image and plotted in Figure 3.6. There is a strong association between the atoms and the cell types. Three atoms were drawn on the same histological section of breast cancer biopsy, with the first one representing the gene expression of the invasive cancer areas (Figure 3.6 B), the second one representing the gene expression of the cancer in situ areas (Figure 3.6 C) and the last one representing the gene expression of the non-cancer areas (Figure 3.6 D). Comparing with traditional cell-type identification methods with human supervision, the exponential family $d$-sparse MDL provides a data-driven way for pathological analysis.

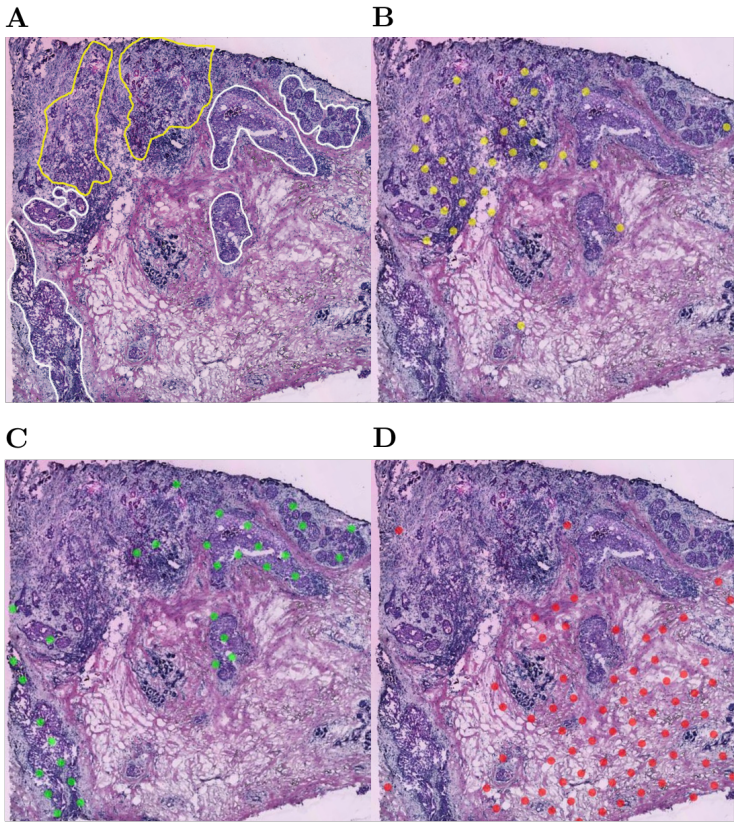



**Figure 3.6:** Plotted in A is the histological section of a breast cancer biopsy with invasive ductal cancer areas (yellow line), ductal cancer in situ areas (white line) and non-cancer areas (other areas). We plotted the predicted invasive ductal cancer areas in B, ductal cancer in situ areas in C and non cancer areas in D.

## 3.6 Discussion

The contribution of MDL procedure to the development of dictionary learning algorithm is two-fold. First, it can deal with any type of observations: continuous observations or discrete observations. The mixture model framework that the MDL procedure relies on includes classical dictionary learning algorithm as a special case. Therefore, MDL can be considered as a generalization of classical dictionary learning approach to the general data format. Second, as demonstrated by our simulation studies, MDL can effectively handle spatially correlated predictors, which can be challenging for existing dictionary learning methods. Another key of MDL is that it does not need to tune any model as required by classical dictionary learning method and its variants. Thus, the algorithm is computational efficient as we only need to estimate a finite number of sparse models instead of searching tuning parameter with infinite number of candidates. We believe that MDL should become an indispensable member of the repository of dictionary learning tools and recommend its broad use. We have implemented the MDL procedure in R, and the R package can be requested from the authors directly.

## Reference


[1] Michal Aharon, Michael Elad, and Alfred Bruckstein. K-svd: An algorithm for designing overcomplete dictionaries for sparse representation. *IEEE Transactions on signal processing*, 54(11):4311–4322, 2006.

[2] Julien Mairal, Francis Bach, Jean Ponce, and Guillermo Sapiro. Online dictionary learning for sparse coding. In *Proceedings of the 26th annual international conference on machine learning*, pages 689–696. ACM, 2009.

[3] Stephane G Mallat. A theory for multiresolution signal decomposition: the wavelet representation. *IEEE transactions on pattern analysis and machine intelligence*, 11(7): 674–693, 1989.

[4] Kangjoo Lee, Sungho Tak, and Jong Chul Ye. A data-driven sparse glm for fmri analysis using sparse dictionary learning with mdl criterion. *IEEE Transactions on Medical Imaging*, 30(5):1076–1089, 2011.

[5] Lawrence Carin, Alfred Hero, Joseph Lucas, David Dunson, Minhua Chen, Ricardo Heñao, Arnau Tibau-Piug, Aimee Zaas, Christopher W Woods, and Geoffrey S Ginsburg. High-dimensional longitudinal genomic data: an analysis used for monitoring viral infections. *IEEE signal processing magazine*, 29(1):108–123, 2012.

[6] Qiang Zhang and Baoxin Li. Discriminative k-svd for dictionary learning in face recognition. In *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*, pages 2691–2698. IEEE, 2010.

[7] Chujian Bi, Haoxiang Wang, and Rui Bao. Sar image change detection using regularized dictionary learning and fuzzy clustering. In *Cloud Computing and Intelligence Systems (CCIS), 2014 IEEE 3rd International Conference on*, pages 327–330. IEEE, 2014.

[8] Harini Eavani, Roman Filipovych, Christos Davatzikos, Theodore D Satterthwaite, Raquel E Gur, and Ruben C Gur. Sparse dictionary learning of resting state fmri networks. In *Pattern Recognition in NeuroImaging (PRNI), 2012 International Workshop on*, pages 73–76. IEEE, 2012.

[9] Karl Skretting and Kjersti Engan. Recursive least squares dictionary learning algorithm. *IEEE Transactions on Signal Processing*, 58(4):2121–2130, 2010.

[10] Shane F Cotter, R Adler, RD Rao, and K Kreutz-Delgado. Forward sequential algorithms for best basis selection. *IEE Proceedings-Vision, Image and Signal Processing*, 146(5):235–244, 1999.

[11] Yagyensh Chandra Pati, Ramin Rezaiifar, and PS Krishnaprasad. Orthogonal matching pursuit: Recursive function approximation with applications to wavelet decomposition.

In *Signals, Systems and Computers, 1993. 1993 Conference Record of The Twenty-Seventh Asilomar Conference on*, pages 40–44. IEEE, 1993.

[12] Ron Kohavi et al. A study of cross-validation and bootstrap for accuracy estimation and model selection. In *Ijcai*, volume 14, pages 1137–1145. Montreal, Canada, 1995.

[13] Gene H Golub, Michael Heath, and Grace Wahba. Generalized cross-validation as a method for choosing a good ridge parameter. *Technometrics*, 21(2):215–223, 1979.

[14] Sylvain Arlot, Alain Celisse, et al. A survey of cross-validation procedures for model selection. *Statistics surveys*, 4:40–79, 2010.

[15] Ping Ma and Wenxuan Zhong. Penalized clustering of large scale functional data with multiple covariates. *Journal of the American Statistical Association*, pages 625–636, 2008.

[16] Sren Nielsen. The stochastic em algorithm: Estimation and asymptotic results. 6, 11 1998.

[17] Gilles Celeux and Gérard Govaert. A classification em algorithm for clustering and two stochastic versions. *Comput. Stat. Data Anal.*, 14(3):315–332, October 1992. ISSN 0167-9473. doi: 10.1016/0167-9473(92)90042-E. URL `http://dx.doi.org/10.1016/0167-9473(92)90042-E`.

[18] Gideon Schwarz et al. Estimating the dimension of a model. *The annals of statistics*, 6 (2):461–464, 1978.

[19] Jonathan G Campbell, Chris Fraley, Fionn Murtagh, and Adrian E Raftery. Linear flaw detection in woven textiles using model-based clustering. *Pattern recognition letters*, 18 (14):1539–1548, 1997.

[20] Abhijit Dasgupta and Adrian E Raftery. Detecting features in spatial point processes with clutter via model-based clustering. *Journal of the American Statistical Association*, 93(441):294–302, 1998.

[21] CF Jeff Wu. On the convergence properties of the em algorithm. *The Annals of statistics*, pages 95–103, 1983.

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

[23] Mark Jenkinson, Christian F Beckmann, Timothy EJ Behrens, Mark W Woolrich, and Stephen M Smith. Fsl. *Neuroimage*, 62(2):782–790, 2012.

[24] Stephen M Smith, Peter T Fox, Karla L Miller, David C Glahn, P Mickle Fox, Clare E Mackay, Nicola Filippini, Kate E Watkins, Roberto Toro, Angela R Laird, et al. Correspondence of the brain's functional architecture during activation and rest. *Proceedings of the National Academy of Sciences*, 106(31):13040–13045, 2009.

[25] Stephen M Smith, Karla L Miller, Steen Moeller, Junqian Xu, Edward J Auerbach, Mark W Woolrich, Christian F Beckmann, Mark Jenkinson, Jesper Andersson, Matthew F Glasser, et al. Temporally-independent functional modes of spontaneous brain activity. *Proceedings of the National Academy of Sciences*, 109(8):3131–3136, 2012.

[26] Jinglei Lv, Binbin Lin, Wei Zhang, Xi Jiang, Xintao Hu, Junwei Han, Lei Guo, Jieping Ye, and Tianming Liu. Modeling task fmri data via supervised stochastic coordinate coding. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 239–246. Springer, 2015.

[27] P. L. Ståhl, F. Salmén, S. Vickovic, A. Lundmark, J. F. Navarro, J. Magnusson, S. Giacomello, M. Asp, J. O. Westholm, M. Huss, A. Mollbrink, S. Linnarsson, S. Codeluppi, Å. Borg, F. Pontén, P. I. Costea, P. Sahlén, J. Mulder, O. Bergmann, J. Lundeberg, and J. Frisén. Visualization and analysis of gene expression in tissue sections by spatial transcriptomics. *Science*, 353:78–82, July 2016. doi: 10.1126/science.aaf2403.