RANDOMIZED FUNCTIONAL DATA ANALYSIS AND ITS APPLICATION IN

ASTRONOMY

A Dissertation

by

XIAOMENG YAN

Submitted to the Graduate and Professional School of
Texas A&M University
in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

Chair of Committee,	Jianhua Z. Huang
Co-Chair of Committee,	Lan Zhou
Committee Members,	Nicholas B. Suntzeff
	Lifan Wang
	Raymond Ka Wai Wong
Head of Department,	Brani Vidakovic

May 2022

Major Subject: Statistics

ABSTRACT

Functional data analysis (FDA) methods have computational and theoretical appeals for some high dimensional data, but lack the scalability to modern large sample datasets. Covariance operators are fundamental concepts and modeling tools for many FDA methods, such as functional principal component analysis. However, the empirical (or estimated) covariance operator becomes too costly to compute when the functional dataset gets big. We study a randomized algorithm for covariance operator estimation. The algorithm works by sampling and rescaling observations from the large functional data collection to form a sketch of much smaller size, and performs computation on the sketch to obtain the subsampled empirical covariance operator. The proposed algorithm is theoretically justified via non-asymptotic bounds between the subsampled and the full-sample empirical covariance operator in terms of the Hilbert-Schmidt norm and operator norm. It is shown that the optimal sampling probability that minimizes the expected squared Hilbert-Schmidt norm of the subsampling error is determined by the norm of each function. Simulated and real data examples are used to illustrate the effectiveness of the proposed algorithm.

The idea of randomization is then used in a Type Ia supernova (SN Ia) spectrophotometric data modeling problem where we develop the Independent Component Estimation (ICE) method for sparse and irregularly spaced spectrophotometric data of Type Ia supernovae (SNe Ia) using functional principal component analysis (FPCA) and independent component analysis (ICA) to explore the separation of SN Ia intrinsic properties and interstellar dust reddening effect. This separation makes it possible to construct the intrinsic spectral energy distribution (SED) manifolds of SNe Ia, which facilitates supernova studies and their cosmological application.

DEDICATION

To my mother and my father.

ACKNOWLEDGMENTS

CONTRIBUTORS AND FUNDING SOURCES

Contributors

The data analyzed in this thesis are all publicly available. The second chapter contains the results of collaboration with Dr. Shiyuan He from the Institute of Statistics and Big Data, Renmin University of China and has been published by Stat [1]. The third chapter is based on a joint work with Drs. Lifan Wang, Shiyuan He and Jianhua Z. Huang.

Funding Sources

NOMENCLATURE

FDA	Functional Data Analysis
$\mathscr{H}$	Separable Hilbert Space
L_2	Space of Squared-Lebesgue-Integrable Functions
$\mathbb{R}$	Set of Real Numbers
UNI	Uniform Sampling Probability
IMPO	Importance Sampling Probability
FPCA	Functional Principal Component Analysis
ICA	Independent Component Analysis
SN Ia (plural SNe Ia)	Type Ia Supernova (plural Type Ia Supernovae)
SED	Spectral Energy Distribution

TABLE OF CONTENTS

Page

LIST OF FIGURES

1. OVERVIEW

1.1 Functional Data Analysis

The modern era brings us advanced technologies and instruments to record data continuously during a time interval or intermittently at several discrete observation points, such as the light curve, a smooth function that shows the brightness of an object over a period of time, spectrum which describes the electromagnetic radiation featured with localized emission and absorption lines, and the spectral energy distribution, a energy surface of an astronomical object observed at different time and wavelength. The nature of functional data is a stochastic process residing in a Hilbert space. Though digitization makes classical multivariate analysis possible, the smooth nature of the observations and the high dimensional fact can lead to ill conditioning of the problems and, eventually, profound consequences. Functional data analysis [3, FDA] is concerned with the development of both theoretical and applicable tools for statistical analysis of this commonly encountered type of data. FDA treats functional observations as realizations of the stochastic process with low intrinsic dimension in an infinite dimensional Hilbert space. This means functional variables can be effectively approximated by splines, wavelets or linear combinations of a few functional principal components. In contrast to classical multivariate methods, FDA takes these features into consideration to achieve optimal convergence rates.

1.2 Randomization

Advances in digital sensors, communications, computation, and storage have created huge collections of functional data, both sample size and dimension of which increase out of the capacity of classical statistical methods. Examples of the large scale datasets include: millions of spectra collected by astronomical surveys (e.g. LAMOST [4] and SDSS [5]); petabyte amounts of hyperspectral images recorded by airborne or satellite remote sensing [6]; brain medical images through the functional magnetic resonance imaging [7, 8, fMRI]. Despite FDA demonstrated value in theory and application, the current FDA methods lack the scalability for modern large scale datasets

[9].

Among the current approaches to making statistical methods scalable to large datasets, randomization has recently gained popularity. The randomized algorithms approach the problem by constructing a substantially smaller sketch of the large scale data at hand. Then, existing methods are applied to the sketch to reduce computation cost. When the sketch keeps the most relevant information, the result computed from the sketch should remain close to the result from the original dataset. There are a few ways to construct a sketch of a large data set. One approach is to draw subsamples with respect to some carefully designed probability, which will select informative samples with larger probability. These sampling probabilities include the importance sampling in matrix multiplication [10], the leverage sampling for least squares regression [11, 12], the subspace sampling for low rank matrix construction [13], etc. Effective sampling probabilities have also been proposed for logistic regression [14, 15] and generalized linear models [16] to minimize the asymptotic variance of the estimator. An alternative way towards sketching is to mix the original data with a random projection and draw samples from the projected data [17, 18].

Since the seminal works of [10, 19, 20] for matrix multiplication and approximation, the idea of randomized algorithm has been successfully applied to optimization [21, 22], low rank matrix estimation [23, 13], least squares regression [11], nonparametric kernel regression [24], etc. These algorithms are able to yield comparatively accurate results at reduced computational and storage costs. See the references [25, 26] for an overview. Most theoretical analysis of these randomized algorithms is conducted from the algorithmic perspective, where the analysis is carried out conditionally on an arbitrarily fixed dataset. Some recent works [12, 27, 18] also draw analysis from the statistical perspective, where statistical properties such as bias and average prediction error are considered.

1.3 Main Contributions

The current literature of randomized algorithms focuses on multivariate statistical methods which are not directly applicable to the functional data setting. Their theoretical results also does not naturally extend to functional data in an infinite dimensional space. Existing work on random-

2

ized algorithms have focused on algorithmic and numerical issues arising in the finite dimensional Euclidean space, but none of them addresses these issues in infinite dimensional Hilbert space and the generalization of the results is non-trivial.

In Chapter 2, we explore randomized estimation of covariance operator for functional data. We propose to estimate the covariance operator via a sketched dataset instead. The sketched data is constructed by sampling and rescaling a small number of functions from the original large dataset. The sampling probability should be chosen to minimize the loss of information. The common strategies are selecting informative samples with larger probability. In determining empirical covariance operator, we argue that informative samples are those with large norm values. To theoretically justify the randomized algorithm, we develop non-asymptotic bounds between the subsampled and full-sample empirical covariance operator. The bounds are developed in terms of Hilber-Schmidt norm and operator norm. For an eigenfunction with a positive eigengap, the operator norm bound directly implies a bound on the eigenfunctions. The bound suggests the potential application of randomized algorithms to functional principal component analysis.

The idea of randomization is also applied to the Type Ia supernova (SN Ia) spectrophotometric data modeling in Chapter 3. We develop the Independent Component Estimation (ICE) method for sparse and irregularly spaced spectrophotometric data of SNe Ia mainly based on functional principal component analysis (FPCA) and assisted by independent component analysis (ICA) using a collection of SN Ia spectra from the Nearby Supernova Factory collaboration (SNFactory). The application of FPCA facilitates a unique low-rank representation of SN Ia data by decomposing the data into a linear combination of multiple components which are efficiently estimated using a stochastic gradient descent algorithm [28]. From astronomical perspective, the ICA allows for the separation of independent physical effects such as the dust extinction and intrinsic color differences of SNe Ia. This separation makes it possible to construct the intrinsic SED manifolds of SNe Ia. A mean extinction curve is derived based on ICA which is in excellent agreement with the dust extinction laws found in the Milky Way but with a significantly steeper wavelength dependence. The intrinsic SED manifolds also allows for the calculation of the dust extinction properties of each

individual SN. Typical values of the total-to-selective extinction ratio R_V are found to be around 1-2 for the intrinsic SED constructed with ICA, which is in agreement with independently derived R_V values of several well observed nearby SNe with NIR data. A numerical toolbox is built to calculate A_V and R_V for any SNe with well calibrated spectral sequence. The toolbox simultaneously builds an intrinsic spectral time series for the SN. With this tool, the color correction in SN Ia distance standardization can be performed separately for the intrinsic color and the host galaxy reddening.

The rest of the work is organized as follows. In Chapter 2, we develop our randomized algorithm for empirical covariance operator estimation. In Chapter 3, we develop the Independent Component Estimation (ICE) method for sparse and irregularly spaced spectrophotometric data of SNe Ia. We end the thesis with our major results and conclusions in Chapter 4.

2. RANDOMIZED COVARIANCE OPERATOR ESTIMATION[*]

The covariance operator generalizes covariance matrix in classical multivariate statistics to measure variable interdependence. It is a fundamental modeling concept for many FDA methods, such as functional principal components, functional canonical analysis, functional time series, etc. Interested readers may refer to the book of [29] for a thorough review. There are several works, e.g., [30, 31] considering the estimation of covariance operator but not the computation for large datasets. We focus on the regime with fully observed functional data. In this case, each observation is regarded as a function whose value is known at any point of its domain. In practical applications, our setup is relevant when functions are recorded with a fixed high frequency over a time interval or recorded on a dense grid in a spatial domain. This type of data has been widely collected by the astronomical surveys, remote sensing, etc, as discussed at the beginning of this section. Given a collection of fully observed functions, the empirical covariance operator, as an estimator of the population covariance operator, is well defined and readily computable. However, the computation becomes expensive as the sample size grows large, since it involves a tensor product for each sample as well as the summations of these tensor products.

This work explores randomized estimation of covariance operator for functional data. The optimal sampling probability that minimizes the expected squared Hilbert-Schimidt norm of the subsampling error has been shown to be proportional to the norm of each function (pairs). This sampling probability is indeed a importance sampling probability which regards the data with larger L_2 norm as more informative one. We also developed concentration bounds between the subsampled and the full-sample empirical covariance operator in terms of the Hilbert-Schmidt norm and the operator norm. The bounds suggest that, under some regularity conditions, the target precision can be guaranteed with high probability if the subsample size is large enough, roughly speaking, inversely proportional to squared precision.

The rest of this chapter is organized as follows. Section 2.1 develops the randomized algorithms for covariance operator estimation. The theoretical guarantees of the algorithms are developed in Section 2.2. A simulation study to assess the proposed algorithms is provided in 2.3. Finally, our randomized algorithm is illustrated on two real world datasets in Section 2.4. The technical proofs are provided in Section 2.5

2.1 Introduction

We start by defining some basic notions. Let x (or y) be a zero-mean random function in a separable Hilbert space $\mathscr{H}_X$ (or $\mathscr{H}_Y$ resp.). The Hilbert space $\mathscr{H}_X$ (or $\mathscr{H}_Y$) is equipped with an inner product $\langle \cdot, \cdot \rangle_X$ (or $\langle \cdot, \cdot \rangle_Y$ resp.) and the induced norm $\| \cdot \|_X$ (or $\| \cdot \|_Y$ resp.). For example, when x is a continuous function over a compact interval Ω, the space $\mathscr{H}_X = L_2(\Omega)$ can be made of all square integrable functions with domain Ω. For $x, x' \in \mathscr{H}_X$, their inner product is $\langle x, x' \rangle_X = \int_\Omega x(\omega) x'(\omega) \, d\omega$ and the induced norm is $\|x\|_X = \left(\int_\Omega x^2(\omega) \, d\omega \right)^{1/2}$. In the following, when the notation is clear from the context, the subscript X (or Y) will be dropped for the norm $\| \cdot \|_X$ (or $\| \cdot \|_Y$ resp.) and the inner product $\langle \cdot, \cdot \rangle_X$ (or $\langle \cdot, \cdot \rangle_Y$ resp.) for simplicity.

Let x (or y) be a zero-mean random function in a separable Hilbert space $\mathscr{H}_X$ (or $\mathscr{H}_Y$ resp.). The Hilbert space $\mathscr{H}_X$ (or $\mathscr{H}_Y$) is equipped with an inner product $\langle \cdot, \cdot \rangle_X$ (or $\langle \cdot, \cdot \rangle_Y$ resp.) and the induced norm $\| \cdot \|_X$ (or $\| \cdot \|_Y$ resp.). For example, when x is a continuous function over a compact interval Ω, the space $\mathscr{H}_X = L_2(\Omega)$ can be made of all square integrable functions with domain Ω. For $x, x' \in \mathscr{H}_X$, their inner product is $\langle x, x' \rangle_X = \int_\Omega x(\omega) x'(\omega) \, d\omega$ and the induced norm is $\|x\|_X = \left(\int_\Omega x^2(\omega) \, d\omega \right)^{1/2}$. In the following, when the notation is clear from the context, the subscript X (or Y) will be dropped for the norm $\| \cdot \|_X$ (or $\| \cdot \|_Y$ resp.) and the inner product $\langle \cdot, \cdot \rangle_X$ (or $\langle \cdot, \cdot \rangle_Y$ resp.) for simplicity.

The population *covariance operator* $\mathscr{C}_{XY}$ generalizes the covariance matrix in multivariate statistics to measure the covariance between x and y. It is a mapping from $\mathscr{H}_X$ to $\mathscr{H}_Y$, such that $u \in \mathscr{H}_X$ is mapped to $\mathscr{C}_{XY} u \in \mathscr{H}_Y$. The covariance operator is determined by the relation

$$\langle \mathscr{C}_{XY} u, v \rangle = \mathbb{E}(\langle x, u \rangle \langle y, v \rangle), \text{ for any } u \in \mathscr{H}_X, v \in \mathscr{H}_Y.$$

Note $\langle \mathscr{C}_{XY} u, v \rangle$ quantifies the correlation between the random scalars $\langle x, u \rangle$ and $\langle y, v \rangle$. In the case where $\mathscr{H}_X = \mathscr{H}_Y = L_2(\Omega)$, the covariance operator can be expressed as an integral operator. Let $k(\omega, \omega') = \mathbb{E}(x(\omega) \cdot y(\omega'))$, then $\mathscr{C}_{XY}$ maps any $u \in L_2(\Omega)$ to $\mathscr{C}_{XY} u \in L_2(\Omega)$ defined by $(\mathscr{C}_{XY} u)(\omega') = \int_{\Omega} k(\omega, \omega') u(\omega) \, d\omega$.

Suppose we have observed N independent pairs of functions $\{(x_n, y_n)\}_{n=1}^{N}$ with identical distribution as the random pair (x, y). The covariance operator can be estimated by the *empirical covariance operator*

$$\widehat{\mathscr{C}}_{XY} = (1/N) \sum_{n=1}^{N} x_n \otimes y_n,$$

where $\otimes$ denotes the tensor product determined by the relation $(x_n \otimes y_n) u = \langle x_n, u \rangle \times y_n$ for any $u \in \mathscr{H}_X$.

The cost of computing the empirical covariance operator $\widehat{\mathscr{C}}_{XY}$ will grow overwhelming when the sample size N becomes large, as it involves the computation of N tensor products and their summations. To reduce the computing cost, we can draw a subset of samples from the large dataset $\{(x_n, y_n)\}_{n=1}^{N}$. The sampling is taken with replacement and according to some appropriate probability distribution $\{p_n\}_{n=1}^{N}$, which satisfies $p_n \geq 0$ and $\sum_{n=1}^{N} p_n = 1$. By this way, we obtain $C \ (\ll N)$ subsamples $\{(\tilde{x}_c, \tilde{y}_c)\}_{c=1}^{C}$. In particular, suppose the c-th subsample is indexed by i_c in the original whole dataset, then, we actually have set $\tilde{x}_c = x_{i_c}$, $\tilde{y}_c = y_{i_c}$. The corresponding sampling probability will be denoted as $\tilde{p}_c = p_{i_c}$ for the c-th subsample. The smaller dataset $\{(\tilde{x}_c, \tilde{y}_c)\}_{c=1}^{C}$ then forms a sketch of the original large dataset $\{(x_n, y_n)\}_{n=1}^{N}$. Given the sketch, the covariance operator can be estimated from this smaller dataset as

$$\widetilde{\mathscr{C}}_{XY} = \frac{1}{CN} \sum_{c=1}^{C} (\tilde{x}_c \otimes \tilde{y}_c)/\tilde{p}_c.$$

This procedure is listed in Algorithm 1. Obviously, conditional on the full dataset $\{(x_n, y_n)\}_{n=1}^{N}$, $\widetilde{\mathscr{C}}_{XY}$ is an unbiased estimator of $\widehat{\mathscr{C}}_{XY}$ under any strictly positive $\{p_n\}_{n=1}^{N}$. Specific choices of the sampling probability $\{p_n\}_{n=1}^{N}$ affect algorithmic performance. The probability distribution of the form $p_n \propto \|x_n\| \cdot \|y_n\|$ will be proven to be optimal in that it minimizes $\mathbb{E}\|\widetilde{\mathscr{C}}_{XY} - \widehat{\mathscr{C}}_{XY}\|_{HS}^2$, see

7

Algorithm 1: Randomized Algorithm for Sample Covariance Operator

---

Input: Dataset $\{(x_n, y_n)\}_{n=1}^N$; sampling probability $\{p_n\}_{n=1}^N$; subsample size C.

Output: $\widetilde{\mathscr{C}}_{XY}$.

  1: **for** $c = 1, \cdots, C$ **do**

  2:     Obtain $(\tilde{x}_c, \tilde{y}_c)$ and $\tilde{p}_c$ with probability $\mathbb{P}\big((\tilde{x}_c, \tilde{y}_c, \tilde{p}_c) = (x_n, y_n, p_n)\big) = p_n$.

  3: **end for**

  4: Compute $\widetilde{\mathscr{C}}_{XY} = \frac{1}{C \cdot N} \sum_{c=1}^C \frac{1}{\tilde{p}_c} \tilde{x}_c \otimes \tilde{y}_c$.

Algorithm 2: Randomized Algorithm for Sample Covariance Operator

---

Input: Dataset $\{x_n\}_{n=1}^N$; sampling probability $\{p_n\}_{n=1}^N$; subsample size C.

Output: $\widetilde{\mathscr{C}}_{XX}$.

  1: **for** $c = 1, \cdots, C$ **do**

  2:     Sample $\tilde{x}_c$ from $\{x_n\}_{n=1}^N$ and get $\tilde{p}_c$ with probability $\mathbb{P}\big((\tilde{x}_c, \tilde{p}_c) = (x_n, p_n)\big) = p_n$.

  3: **end for**

  4: Compute $\widetilde{\mathscr{C}}_{XX} = \frac{1}{C \cdot N} \sum_{c=1}^C \frac{1}{\tilde{p}_c} \tilde{x}_c \otimes \tilde{x}_c$.

Theorem 1.

Similarly, the covariance operator $\mathscr{C}_{XX}$ can be defined involving x with itself. In particular, it is a mapping from $\mathscr{H}_X$ to $\mathscr{H}_X$, such that $u \in \mathscr{H}_X$ is mapped to $\mathscr{C}_{XX}u \in \mathscr{H}_X$. The covariance operator is determined by the relation $\langle \mathscr{C}_{XX}u, v \rangle = \mathbb{E}(\langle x, u \rangle \langle x, v \rangle)$ for any $u, v \in \mathscr{H}_X$. After getting N observations $\{x_n\}_{n=1}^N$ with identical distribution as x, the finite sample empirical estimator $\widehat{\mathscr{C}}_{XX} = (1/N) \sum_{n=1}^N x_n \otimes x_n$ can be calculated correspondingly. We also note Algorithm 1 can be easily adapted for randomized estimation of $\mathscr{C}_{XX}$, as in Algorithm 2.

In some applications, each observation x_n (or y_n) is recorded digitally as a high dimensional vector of length T. For example, when $\mathscr{H}_X = L_2(\Omega)$ with $\Omega = [0, 1]$, then x_n could be recorded as $(x_n(1/T), x_n(2/T), \cdots, x_n(1))^T \in \mathbb{R}^T$. For the spectral dataset of LAMOST survey [4], N is in the order of millions and $T \sim 3500$ for each spectrum. Evaluating the full sample empirical covariance operator $\widehat{\mathscr{C}}_{XX}$ requires a complexity of $\mathscr{O}(NT^2)$. On the other hand, to compute the subsampled empirical covariance operator $\widetilde{\mathscr{C}}_{XX}$, we need to evaluate the sampling probability p_n in the first place. Suppose the sampling probability is chosen as $p_n \propto \|x_n\|^2$. Then, computing

$\{p_n\}_{n=1}^N$ requires one scan over the whole dataset with complexity $\mathscr{O}(NT)$. After obtaining C subsamples, evaluating $\widehat{\mathscr{C}}_{XX}$ only costs $\mathscr{O}(CT^2)$. Therefore, the total complexity of the algorithm is $\mathscr{O}(NT + CT^2)$. When $C \ll N$, the algorithm approximately costs $\mathscr{O}(NT)$, which reduces the cost of the full sample algorithm by a factor of T.

Closely related to the self-adjoint operator $\widehat{\mathscr{C}}_{XX}$ is the concept of functional principal component. Traditionally, functional principal components can be obtained via the *eigenvalue-eigenfunction decomposition* of the full sample empirical covariance operator $\widehat{\mathscr{C}}_{XX}$. The decomposition admits the form

$$\widehat{\mathscr{C}}_{XX} = \sum_{r=1}^\infty \hat{\sigma}_r^2 (\hat{\theta}_r \otimes \hat{\theta}_r),$$

where $\hat{\sigma}_1^2 \geq \hat{\sigma}_2^2 \geq \cdots$ is a decreasing sequence of eignvalues, and $\hat{\theta}_r$'s are the corresponding orthonormal eigenfunctions. The leading R eigenfunctions are usually referred to as functional principal components, which explains most of the variability of the dataset $\{x_n\}_{n=1}^N$. As we have noted, computing the full sample empirical covariance operator $\widehat{\mathscr{C}}_{XX}$ is expensive when the sample size N is large. The proposed algorithm provides a faster way to estimate the covariance operator and obtain the estimate of functional principal components. That is, with the eigenvalue-eigenfunction decomposition of the subsampled empirical covariance operator,

$$\widetilde{\mathscr{C}}_{XX} = \sum_{r=1}^\infty \tilde{\sigma}_r^2 (\tilde{\theta}_r \otimes \tilde{\theta}_r),$$

the eigenfunctions $\tilde{\theta}_1, \cdots, \tilde{\theta}_R$ can be used as surrogate of $\hat{\theta}_1, \cdots, \hat{\theta}_R$ for functional principal component analysis.

2.2 The Non-asymptotic Bound

In this section, we quantify the uncertainty associated with the subsampling estimation error $\widetilde{\mathscr{C}}_{XY} - \widehat{\mathscr{C}}_{XY}$, where $\widetilde{\mathscr{C}}_{XY}$ is obtained from Algorithm 1. When the Hilbert spaces are finite-dimensional, the theoretical results in this section degenerate to that of [10] and that of Section 6.4 from [32] in Euclidean space. The results presented here is more general and is framed with

9

infinite-dimensional Hilbert space.

The magnitude of the subsampling estimation error can be quantified by the Hilbert–Schmidt norm. Suppose $\{\theta_{Xj}\}_{j=1}^{\infty}$ and $\{\theta_{Yj}\}_{j=1}^{\infty}$ are two groups of complete orthonormal basis for $\mathscr{H}_X$ and $\mathscr{H}_Y$, respectively. The *Hilbert–Schmidt norm* for an operator $\mathscr{C}_{XY}$ is computed as $\|\mathscr{C}_{XY}\|_{HS} := \left(\sum_{i,j}\langle\mathscr{C}_{XY}\theta_{Xi},\theta_{Yj}\rangle^2\right)^{1/2}$. The next theorem analytically evaluates $\mathbb{E}\|\widetilde{\mathscr{C}}_{XY} - \widehat{\mathscr{C}}_{XY}\|_{HS}^2$ and minimize it with optimal sampling probability.

Theorem 1. *Conditional on the full observation* $\{(x_n, y_n)\}_{n=1}^{N}$, *we have*

$$\mathbb{E}\|\widetilde{\mathscr{C}}_{XY} - \widehat{\mathscr{C}}_{XY}\|_{HS}^2 = (1/C)\left[(1/N^2)\sum_{n=1}^{N}\|x_n\|^2\cdot\|y_n\|^2/p_n - \|\widehat{\mathscr{C}}_{XY}\|_{HS}^2\right]. \qquad (2.1)$$

Furthermore, the choice of

$$p_n = \|x_n\|\cdot\|y_n\|/\left[\sum_{m=1}^{N}\|x_m\|\cdot\|y_m\|\right] \qquad (2.2)$$

minimizes the expected error, which yields the optimal $\mathbb{E}\|\widetilde{\mathscr{C}}_{XY} - \widehat{\mathscr{C}}_{XY}\|_{HS}^2 = (1/C)\left[(1/N^2)\left(\sum_{i=1}^{N}\|x_i\|\cdot\|y_i\|\right)^2 - \|\widehat{\mathscr{C}}_{XY}\|_{HS}^2\right]$.

Theorem 1 extends Lemma 4 of [10] in Euclidean space. Note the analysis is carried out conditional on the whole dataset $\{(x_n, y_n)\}_{n=1}^{N}$, and the expectation is taken with respect to the subsampling process. As expected, the squared Hilbert-Schmidt error is inversely proportional to the subsample size C in (2.1). The optimal sampling probability $p_n \propto \|x_n\|\cdot\|y_n\|$ minimizes the constant in the bracket of (2.1) and is called *importance sampling probability* (IMPO). Using Jensen's inequality and the fact $\|\widetilde{\mathscr{C}}_{XY} - \widehat{\mathscr{C}}_{XY}\| \le \|\widetilde{\mathscr{C}}_{XY} - \widehat{\mathscr{C}}_{XY}\|_{HS}$, we can directly obtain bounds for $\mathbb{E}\|\widetilde{\mathscr{C}}_{XY} - \widehat{\mathscr{C}}_{XY}\|_{HS}$ and $\mathbb{E}\|\widetilde{\mathscr{C}}_{XY} - \widehat{\mathscr{C}}_{XY}\|$ from the above. Corollary 2 below is a direct consequence of Theorem 1 for $\widetilde{\mathscr{C}}_{XX}$ by setting $x_n = y_n$.

Corollary 2. *Conditional on the full observation* $\{x_n\}_{n=1}^{N}$, *the choice of sampling probability*

$$p_n = \|x_n\|^2/\left[\sum_{m=1}^{N}\|x_m\|^2\right] \qquad (2.3)$$

yields minimal $\mathbb{E}\|\widetilde{\mathscr{C}}_{XX} - \widehat{\mathscr{C}}_{XX}\|^2_{HS} = (1/C)\big[(1/N^2)\big(\sum_{i=1}^N \|x_i\|^2\big)^2 - \|\widehat{\mathscr{C}}_{XX}\|^2_{HS}\big].$

Non-asymptotic concentration bounds in terms of operator norm can also be obtained. The *operator norm* is defined by $\|\widehat{\mathscr{C}}_{XY}\| := \sup_{u \in \mathscr{H}_X : \|u\| \leq 1} \|\widehat{\mathscr{C}}_{XY}u\|.$ The subsampling error in operator norm $\|\widetilde{\mathscr{C}}_{XY} - \widehat{\mathscr{C}}_{XY}\|$ will be measured relative to the size of $\|\widehat{\mathscr{C}}_{XX}\|$ and $\|\widehat{\mathscr{C}}_{YY}\|.$ As $\widehat{\mathscr{C}}_{XX}$ and $\widehat{\mathscr{C}}_{YY}$ are self-adjoint and positive semi-definite, their operator norms equal their largest eigenvalues. In an infinite-dimensional Hilbert space, operator concentration inequalities are developed with the notion of intrinsic dimension. For the operator $\widehat{\mathscr{C}}_{XX}$, its *intrinsic dimension* is defined by $\mathrm{intdim}(\widehat{\mathscr{C}}_{XX}) := \mathrm{tr}(\widehat{\mathscr{C}}_{XX})/\|\widehat{\mathscr{C}}_{XX}\|.$ The *trace* for $\mathscr{C}_{XY}$ is defined as $\mathrm{tr}(\widehat{\mathscr{C}}_{XY}) := \sum_{i,j}\langle \widehat{\mathscr{C}}_{XX}\theta_{Xi}, \theta_{Yj}\rangle.$ Because $\widehat{\mathscr{C}}_{XX}$ is computed from a finite amount of N samples, it obviously holds that $\mathrm{intdim}(\widehat{\mathscr{C}}_{XX}) \leq \mathrm{rank}(\widehat{\mathscr{C}}_{XX}) \leq N.$

Theorem 3. *Denote by* $\hat{\delta}_x = \mathrm{intdim}(\widehat{\mathscr{C}}_{XX})$ *and* $\hat{\delta}_y = \mathrm{intdim}(\widehat{\mathscr{C}}_{YY})$ *the intrinsic dimensions of* $\widehat{\mathscr{C}}_{XX}$ *and* $\widehat{\mathscr{C}}_{YY}$, *respectively. In addition, define* $\hat{\sigma}^2_{x,1} = \|\widehat{\mathscr{C}}_{XX}\|,$ $\hat{\sigma}^2_{y,1} = \|\widehat{\mathscr{C}}_{YY}\|,$ *and* $\hat{\eta} = \|\widehat{\mathscr{C}}_{XY}\|/(\hat{\sigma}_{x,1}\hat{\sigma}_{y,1}).$ *Suppose for some* $\beta \geq 1$, *the sampling probability of Algorithm 1 satisfies*

$$p_n \geq \frac{1}{\beta}\frac{\|x_n\|^2/\hat{\sigma}^2_{x,1} + \|y_n\|^2/\hat{\sigma}^2_{y,1}}{\sum_{m=1}^N \big(\|x_m\|^2/\hat{\sigma}^2_{x,1} + \|y_m\|^2/\hat{\sigma}^2_{y,1}\big)}. \tag{2.4}$$

Then for ϵ *satisfying* $\epsilon \cdot C \geq \sqrt{C\beta(\hat{\delta}_x + \hat{\delta}_y)} + \beta(\hat{\delta}_x + \hat{\delta}_y)/6 + \hat{\eta}/3$, *it holds that*

$$\mathbb{P}\big(\|\widetilde{\mathscr{C}}_{XY} - \widehat{\mathscr{C}}_{XY}\| \geq \epsilon \cdot \hat{\sigma}_{x,1}\hat{\sigma}_{y,1}\big) \leq 4(\hat{\delta}_x + \hat{\delta}_y)\exp\Big(-\frac{C\epsilon^2}{2\beta(\hat{\delta}_x + \hat{\delta}_y)(1 + \epsilon/6) + 2\hat{\eta}\epsilon/3}\Big).$$

The above theorem provides performance guarantee for Algorithm 1 in terms of the relative error $\frac{1}{\hat{\sigma}_{x,1}\hat{\sigma}_{y,1}}\|\widetilde{\mathscr{C}}_{XY} - \widehat{\mathscr{C}}_{XY}\|.$ Note (2.4) is a lower bound requirement for the sampling probability. Smaller value of $\beta(\geq 1)$ will deliver smaller bound for the error probability. The error bound from this theorem allows us to give a conservative estimate of the subsample size C to ensure that the approximation error is within certain upper bound. For example, suppose we want to control the relative error within $\epsilon(\leq 1)$ and with probability 0.9. The theorem states that the subsample size

11

C should exceed

$$C \geq \frac{4}{\epsilon^2} \left(\beta(\hat{\delta}_x + \hat{\delta}_y) + \hat{\eta} \right) \log \left(40(\hat{\delta}_x + \hat{\delta}_y) + e \right). \tag{2.5}$$

This required subsample size scales inverse proportionally to the squared precision ϵ^2, and scales superlinearly with the intrinsic dimension $\hat{\delta}_x + \hat{\delta}_y$.

Following similar argument as in Theorem 3, we establish concentration bound for $\|\widetilde{\mathscr{C}}_{XX} - \widehat{\mathscr{C}}_{XX}\|$ in the next theorem.

Theorem 4. *Let* $\hat{\delta}_x = \mathrm{intdim}(\widehat{\mathscr{C}}_{XX})$ *and* $\hat{\sigma}^2_{x,1} = \|\widehat{\mathscr{C}}_{XX}\|$. *Suppose the sampling probability satisfies* $p_n \geq \frac{1}{\beta} \frac{\|x_n\|^2}{\sum_{m=1}^{N} \|x_m\|^2}$ *for some* $\beta \geq 1$. *Then, we have that*

$$\mathbb{P}\left(\|\widetilde{\mathscr{C}}_{XX} - \widehat{\mathscr{C}}_{XX}\| \geq \epsilon \cdot \hat{\sigma}^2_{x,1} \right) \leq 4\hat{\delta}_x \exp\left(-\frac{C\epsilon^2/2}{\beta\hat{\delta}_x + (\beta\hat{\delta}_x + 1)\epsilon/3} \right),$$

for ϵ satisfying $C \cdot \epsilon > \sqrt{C\beta \cdot \hat{\delta}_x} + (\beta\hat{\delta}_x + 1)/3$.

The interpretation of Theorem 4 is similar to that of Theorem 3. We can compute exactly $p_n = \|x_n\|^2 / \left(\sum_{m=1}^{N} \|x_m\|^2 \right)$ with $\beta = 1$ through one scan over the whole dataset. Suppose our target is to control the relative error $\|\widetilde{\mathscr{C}}_{XX} - \widehat{\mathscr{C}}_{XX}\| / \hat{\sigma}^2_{x,1}$ within $\epsilon(\leq 1)$ and with probability 0.9, Theorem 4 indicates the subsample size should satisfy the lower bound

$$C \geq \frac{8}{3\epsilon^2} (\hat{\delta}_x + 1) \log(40\hat{\delta}_x + e). \tag{2.6}$$

In practical applications, the exact value of this lower bound is unknown unless we have $\hat{\delta}_x$ from the full sample computation, which contradicts our goal of reducing computation cost via subsampling. To determine C in practice, we need to estimate the intrinsic dimension $\hat{\delta}_x$. Recall the definition $\hat{\delta}_x = \mathrm{intdim}(\widehat{\mathscr{C}}_{XX}) = \mathrm{tr}(\widehat{\mathscr{C}}_{XX}) / \|\widehat{\mathscr{C}}_{XX}\|$. The trace value in the numerator, $\mathrm{tr}(\widehat{\mathscr{C}}_{XX}) = (1/N) \sum_{n=1}^{N} \|x_i\|^2$, can be easily obtained at the same time when we compute the sampling probability $p_n = \|x_n\|^2 / \left(\sum_{m=1}^{N} \|x_m\|^2 \right)$. It only remains to estimate $\|\widehat{\mathscr{C}}_{XX}\|$ in the de-

nominator of $\hat{\delta}_x$. For this, we can apply Algorithm 2 to get a pilot covariance operator $\widetilde{\mathscr{C}}'_{XX}$ with a relatively small C' (e.g $C' = 1,000$ or $C' = 5,000$). The largest eigenvalue $(\hat{\sigma}'_{x,1})^2$ of $\widetilde{\mathscr{C}}'_{XX}$ serves as an estimate for $\hat{\sigma}^2_{x,1}$. Combining this $(\hat{\sigma}'_{x,1})^2$ with the computed $\mathrm{tr}(\widehat{\mathscr{C}}_{XX})$, we can obtain an estimate for $\hat{\delta}_x$, and thereby be able to determine an appropriate value for C according to (2.6). With this specified subsample size C, Algorithm 2 will be executed again to get the final randomized estimate $\widehat{\mathscr{C}}_{XX}$.

To determine C for the estimation of $\widehat{\mathscr{C}}_{XY}$ via (2.5), we can carry out a similar two-step procedure in practice. Given a relatively small subsample size C', we firstly apply Algorithm 1 and Algorithm 2 with sampling probability (2.2) and (2.3), respectively. They will produce pilot operator estimates $\widetilde{\mathscr{C}}'_{XX}$, $\widetilde{\mathscr{C}}'_{YY}$ and $\widetilde{\mathscr{C}}'_{XY}$. Their largest eigenvalues and trace values help us to estimate $\hat{\delta}_x$, $\hat{\delta}_y$, $\hat{\sigma}^2_{x,1}$, $\hat{\sigma}^2_{y,1}$ and $\hat{\eta}$. These estimates can be plugged in (2.5). Now, the only unspecified quantity in bound (2.5) is the value of β. To get an appropriate value of β, we denote $(\hat{\sigma}'_{x,1})^2 = \|\widetilde{\mathscr{C}}'_{XX}\|$ and $(\hat{\sigma}'_{y,1})^2 = \|\widetilde{\mathscr{C}}'_{YY}\|$. Suppose the pilot estimates satisfy the event

$$\|\widetilde{\mathscr{C}}'_{XX} - \widehat{\mathscr{C}}_{XX}\| \leq \hat{\sigma}^2_{x,1}/2 \text{ and } \|\widetilde{\mathscr{C}}'_{YY} - \widehat{\mathscr{C}}_{YY}\| \leq \hat{\sigma}^2_{y,1}/2,$$

then the perturbation bounds imply

$$|(\hat{\sigma}'_{x,1})^2 - \hat{\sigma}^2_{x,1}| \leq \|\widetilde{\mathscr{C}}'_{XX} - \widehat{\mathscr{C}}_{XX}\| \leq \hat{\sigma}^2_{x,1}/2 \text{ and } |(\hat{\sigma}'_{y,1})^2 - \hat{\sigma}^2_{y,1}| \leq \|\widetilde{\mathscr{C}}'_{YY} - \widehat{\mathscr{C}}_{YY}\| \leq \hat{\sigma}^2_{y,1}/2.$$

When this event happens, it follows that

$$p_n = \frac{\|x_n\|^2/(\hat{\sigma}'_{x,1})^2 + \|y_n\|^2/(\hat{\sigma}'_{y,1})^2}{\sum_{m=1}^N \left(\|x_m\|^2/(\hat{\sigma}'_{x,1})^2 + \|y_m\|^2/(\hat{\sigma}'_{y,1})^2\right)} \geq \frac{1}{\beta} \frac{\|x_n\|^2/\hat{\sigma}^2_{x,1} + \|y_n\|^2/\hat{\sigma}^2_{y,1}}{\sum_{m=1}^N \left(\|x_m\|^2/\hat{\sigma}^2_{x,1} + \|y_m\|^2/\hat{\sigma}^2_{y,1}\right)}$$

with $\beta = 3$. In practical applications, we can be more conservative by specifying a slightly larger value of $\beta > 3$ in (2.5).

Recall we have discussed that functional principal components can be obtained from the subsampled covariance oeperator $\widetilde{\mathscr{C}}_{XX}$ to reduce the cost of computation. By a direct result of per-

turbation theory Lemma 1 of [33], the eigenfunctions $\tilde{\theta}_r$'s of the subsample estimator $\widetilde{\mathscr{C}}_{XX}$ should remain close to the eigenfunctions $\hat{\theta}_r$'s of the full sample estimator $\widehat{\mathscr{C}}_{XX}$, as long as the subsample size C and the eigengap are large enough. Assume the eigenvalue of the r-th eigenfunction $\hat{\theta}_r$ has multiplicity one. Let $g_r = \hat{\sigma}_r^2 - \hat{\sigma}_{r+1}^2$ be the r-th *eigengap* and set $G_r = \min\{g_r, g_{r-1}\}$. When $G_r > 0$, it follows immediately that $\|\hat{\theta}_r - s_r\tilde{\theta}_r\| \leq (2\sqrt{2}/G_r)\|\widetilde{\mathscr{C}}_{XX} - \widehat{\mathscr{C}}_{XX}\|$, with $s_r = \text{sign}(\langle\hat{\theta}_r, \tilde{\theta}_r\rangle)$, see Lemma 2.3 of [29]. As a result, Theorem 4 also directly implies a bound for the leading functional principal components.

2.3 Simulation Study
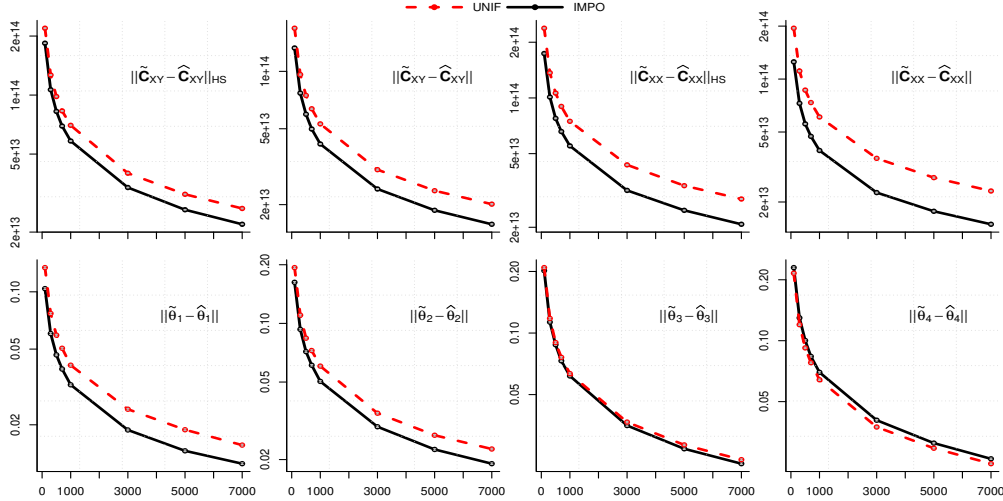
Figure 2.1: Randomized covariance operator estimation and FPCA for the nearly uniform (NU) setting. The vertical axes represent comparison metrics in log scale. The red dashed and black solid lines correspond to the results of the UNIF and IMPO sampling probability, respectively.

To evaluate the randomized algorithm on estimating covariance operator and functional principal components, in each replication of simulations, we generate one synthetic dataset $\{(x_n, y_n)\}_{n=1}^N$. Each x_n (or y_n) is a continuous function defined over the interval $\Omega = [0, 1]$. In particular, each pair is generated according to $x_n(\omega) = \sum_{r=1}^{50} \sigma_r \xi_{nr}^x \theta_r(\omega)$ and $y_n(\omega) = \sum_{r=1}^{50} \sigma_r \xi_{nr}^y \theta_r(\omega)$. We set θ_r as the Fourier function which is $\theta_r(\omega) = \sqrt{2}\sin(2\pi r\omega)$ when r is odd, and $\theta_r(\omega) = \sqrt{2}\cos(2\pi r\omega)$
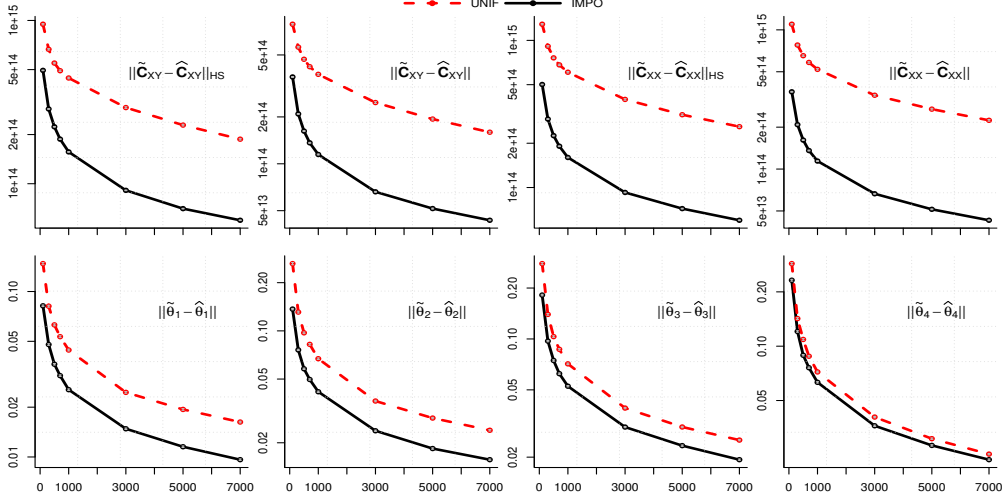
Figure 2.2: Randomized covariance operator estimation and FPCA for the moderately nonuniform (MN) setting.

when r is even. The eigenvalues $\sigma_1^2, \cdots, \sigma_{50}^2$ are specified by a exponentially decaying sequence of values, i.e. $\sigma_r^2 = c \cdot \kappa^r$ with $c = 2^{51}$ and $\kappa = 0.5$.

Each random score pair, (ξ_{nr}^x, ξ_{nr}^y)'s, independently follows a bi-variate distribution with zero mean and a fixed covariance matrix. The covariance matrix has ones on its diagonal and 0.5's on its off-diagonal entries. In particular, the distribution of (ξ_{nr}^x, ξ_{nr}^y) is set to one of the following three settings. The first is called nearly uniform (NU) distribution, for which (ξ_{nr}^x, ξ_{nr}^y)'s are drawn from bi-variate Gaussian distribution. The second is called moderately nonuniform (MN) distribution, where (ξ_{nr}^x, ξ_{nr}^y)'s are generated by bi-variate t-distribution with 3 degree of freedom. The third is denoted as very nonuniform (VN), where (ξ_{nr}^x, ξ_{nr}^y)'s follows bi-variate t-distribution with 1 degree of freedom. Notice in these three cases, the distribution of the score pairs for NU has the lightest tail, while VN has the heaviest tail. As a result, the function pairs in a synthetic dataset from NU will behave homogeneously. On the other hand, the synthetic datasets from MN or VN are likely to contain some function pairs with extreme magnitude, which could affect the data-driven sampling distribution in an adverse way. From the above specification, we can deduce that the population covariance operator $\mathscr{C}_{XY}$ for (x_n, y_n) can be represented as $(\mathscr{C}_{XY}u)(\omega') = \int_0^1 k(\omega, \omega')u(\omega)\,\mathrm{d}\omega$ for any $u \in L_2([0,1])$, with $k(\omega, \omega') =$

15
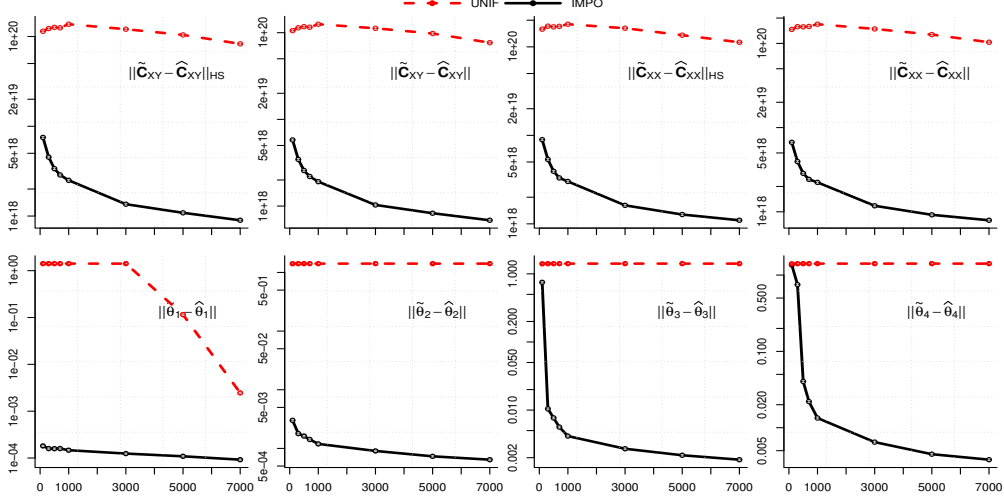
Figure 2.3: Randomized covariance operator estimation and FPCA for the very non-uniform (VN) setting.

$\sum_{r=1}^{50} 0.5\sigma_r^2 \theta_r(\omega)\theta_r(\omega')$. Similarly, the true covariance operator $\mathscr{C}_{XX}$ for x_n is determined by the integral $(\mathscr{C}_{XX}u)(\omega') = \int_0^1 k(\omega, \omega')u(\omega)\,\mathrm{d}\omega$ with $k(\omega, \omega') = \sum_{r=1}^{50} \sigma_r^2 \theta_r(\omega)\theta_r(\omega')$. The r-th eigenvalue and eigenfunction of $\mathscr{C}_{XX}$ are σ_r^2 and θ_r, respectively.

By three different choices (NU, MN, or VN) of score pair distributions, we have three distinct ways to generate a full sample dataset $\{(x_n, y_n)\}_{n=1}^N$ of size $N = 10,000$. The full sample empirical covariance operator $\widehat{\mathscr{C}}_{XY}$ and $\widehat{\mathscr{C}}_{XX}$ will be computed. We then apply the randomized algorithms to estimate the subsampled empirical covariance operator $\widetilde{\mathscr{C}}_{XY}$ and $\widetilde{\mathscr{C}}_{XX}$. Based on the eigenvalue-eigenfunction decomposition of the subsampled $\widetilde{\mathscr{C}}_{XX}$, the first four eigenfunctions $\tilde{\theta}_1, \cdots, \tilde{\theta}_4$ can also be obtained as estimates of the full sample counterparts $\hat{\theta}_1, \cdots, \hat{\theta}_4$ (eigenfunctions of $\widehat{\mathscr{C}}_{XX}$).

Two sampling probabilities are plugged into Algorithm 1 for comparision. These include: (i) the naive uniform sampling probability (UNIF) $p_n^{\text{unif}} = 1/N$; and (ii) the optimal sampling probability (IMPO, suggested by Theorem 1 and Corollary 2) $p_n^{\text{impo}} = \|x_n\|\|y_n\|/\left(\sum_{m=1}^N \|x_m\|\|y_m\|\right)$ for estimating $\widehat{\mathscr{C}}_{XY}$ or $p_n^{\text{impo}} = \|x_n\|^2/\left(\sum_{m=1}^N \|x_m\|^2\right)$ for $\widehat{\mathscr{C}}_{XX}$.

The algorithm performance will be evaluated by three comparison metrics, including the Hilbert-Schmidt norm ($\|\widehat{\mathscr{C}}_{XY} - \widetilde{\mathscr{C}}_{XY}\|_{HS}$ and $\|\widehat{\mathscr{C}}_{XX} - \widetilde{\mathscr{C}}_{XX}\|_{HS}$) and the operator norm ($\|\widehat{\mathscr{C}}_{XY} - \widetilde{\mathscr{C}}_{XY}\|$ and $\|\widehat{\mathscr{C}}_{XX} - \widetilde{\mathscr{C}}_{XX}\|$). Notice these are differences between the full sample estimates and the subsample

estiamtes, as bounded by our theoretical results. In addition, the estimation error for eigenfunctions will be also computed by $\|\hat{\theta}_r - \text{sign}(\langle \hat{\theta}_r, \tilde{\theta}_r \rangle)\tilde{\theta}_r\|$.

The procedure is repeated $5,000$ times to compute the average errors. The results are presented in Figures 2.1–2.3. The three figures correspond to the cases with NU, MN and VN distributions, respectively. In each figure, the four panels in the first row show the estimation error of $\|\widehat{\mathscr{C}}_{XY} - \widetilde{\mathscr{C}}_{XY}\|_{HS}$, $\|\widehat{\mathscr{C}}_{XY} - \widetilde{\mathscr{C}}_{XY}\|$, $\|\widehat{\mathscr{C}}_{XX} - \widetilde{\mathscr{C}}_{XX}\|_{HS}$ and $\|\widehat{\mathscr{C}}_{Xx} - \widetilde{\mathscr{C}}_{XX}\|$, respectively. The panels in the second row draw the estimation error of the individual $\tilde{\theta}_1$, $\tilde{\theta}_2$, $\tilde{\theta}_3$ and $\tilde{\theta}_4$, respectively. In addition, for each panel, the horizontal axes represent the subsample size C, while the vertical axes are the comparison metrics in the logarithm scale with base 10. The red dashed and black solid lines correspond to the results of the UNIF and IMPO sampling probability, respectively.

It is seen from the figures that, as the subsample size increases, the average errors tend to decrease monotonically for both random sampling methods. The advantage of IMPO over UNIF increases substantially when the data distribution changes from nearly uniform to moderately nonuniform and then to very nonuniform. The optimal sampling algorithm IMPO always outperforms the naive uniform sampling UNIF in estimating the covariance operators, which corroborates our theoretical results. In particular, UNIF almost fails completely for the VN datasets in Figure 2.3, where the observations have extreme values, whereas IMPO can substantially preserve the information contained in covariance operator by assigning extreme observations with a larger sampling probability. For estimation of the eigenfunctions, the accuracy of IMPO is higher than UNIF in all cases except the estimation of $\hat{\theta}_4$ in the NU setting, where UNIF is expected to work well since the data have nearly uniform distribution.

2.4 Real Data Analysis

Our randomized algorithm is illustrated on two real datasets: one is the hourly air pollution dataset from the Environmental Protection Agency (EPA, `https://aqs.epa.gov/aqsweb/airdata/download_files.html`), and the other is the stellar spectrum dataset from the survey of the Large sky Area Multi-Object fiber Spectroscopic Telescope [4, LAMOST].

The EPA dataset has hourly-level measurements of O_3, SO_2 and CO from the state, local and

17
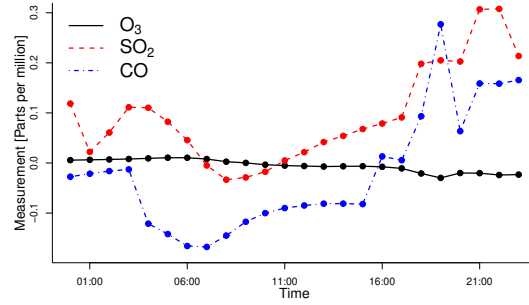
Figure 2.4: The level of O_3, SO_2 and CO collected from Site Number 2, Contra Costa County, California on 2019-01-01.
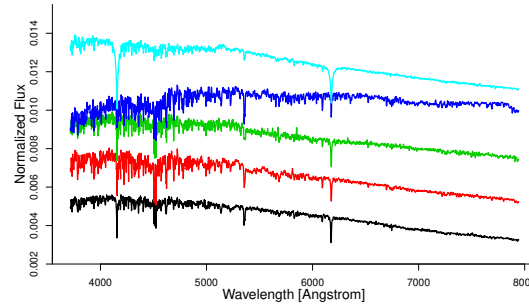
Figure 2.5: Five stellar spectra from LAMOST with 0.0002 offset in the vertical direction.

tribal monitoring agencies across the United States. The level of the criteria gases O_3, SO_2 and CO is part of for the Air Quality Index. At an exemplary site, the hourly measurement in one day is plotted in Figure 2.4. The horizontal axis is hour of the day, and the vertical axis is the criteria gases level. Each curve in the plot can be viewed as functional data, where the gas level varies as a function of time. We will treat the O_3, SO_2 and CO measurements from one site in a day as a realization of a random function triple (x, y, z). From the released 2019 hourly dataset, we select 20,770 observation triples by ensuring each curve having more than 20 measurement points. The missing measurements are imputed by linear interpolation. The randomized algorithm is applied to estimate the covariance operator between the three criteria gases.

18

The LAMOST survey began carrying out its scientific spectroscopic recording of millions of stars and galaxies in 2012. Its current data release includes more than nine million spectra. From the current LAMOST data release, we randomly select 110,000 spectra whose r-band signal-to-noise ratio is greater than 35 as our full dataset. Each spectrum is normalized with unit norm. Five randomly selected spectra are plotted in Figure 2.5. The horizontal axis is wavelength, and the vertical axis is normalized flux. The flux measures the brightness of an object. Each spectrum observation can also be viewed as functional data, where the flux value varies as a function of wavelength. For this dataset, we will apply the randomized algorithm to estimate the covariance operator and its leading five eigenfunctions.

Two sampling strategies (UNIF and IMPO) in Section 2.3 are utilized for the above two datasets. The procedure is repeated 300 times. Each time the subsampled estimates are compared with the full sample results by the metrics introduced in Section 2.3, and the full sample estimators are computed with brute force. Table 2.1 and Table 2.2 present the average estimation accuracy and related standard errors for the EPA and LAMOST dataset, respectively. The metrics are reported in the scale of $\log_{10}$. The comparison results are similar to that of Section 2.3. The IMPO sampling strategy delivers smaller error in estimating the covariance operators in all cases. In Table 2.2, it also has smaller estimation error for all the eigenfunctions. The running time for constructing sampling probabilities and performing Algorithm 2 for spectra dataset is reported in the last row in Table 2.2. The time to compute full sample empirical covariance operator is 257.82 secs, whereas, the running time for the proposed randomized algorithms with sampling probability being IMPO and subsample size C being $1,000$, $5,000$ and $10,000$ is 7.69 secs, 15.78 secs and 27.38 secs, respectively. This suggests that the randomized algorithms are more computationally efficient.

2.5 Technical Proofs

Theorem 7.3.1 of [32] developed intrinsic matrix Bernstein concentration inequality. Together with the techniques in Section 3.2 of [34], it can be easily extended to compact operators. Similar results can also be found in Lemma 5 of [35]. In the below, $\mathscr{H}_1$ and $\mathscr{H}_2$ are two Hilbert spaces.

Table 2.1: Randomized covariance operator estimation for the Criteria Gases dataset.

	$C = 100$		$C = 1,000$		$C = 10,000$	
	UNIF	IMPO	UNIF	IMPO	UNIF	IMPO
$\|\widehat{\mathscr{C}}_{XY} - \widetilde{\mathscr{C}}_{XY}\|_{HS}$	-2.932	-3.246	-3.268	-3.743	-3.738	-4.243
	(0.008)	(0.002)	(0.004)	(0.002)	(0.003)	(0.002)
$\|\widehat{\mathscr{C}}_{XY} - \widetilde{\mathscr{C}}_{XY}\|$	-2.965	-3.356	-3.277	-3.845	-3.760	-4.349
	(0.010)	(0.003)	(0.004)	(0.003)	(0.004)	(0.003)
$\|\widehat{\mathscr{C}}_{XZ} - \widetilde{\mathscr{C}}_{XZ}\|_{HS}$	-3.645	-3.865	-4.134	-4.364	-4.631	-4.865
	(0.003)	(0.002)	(0.002)	(0.001)	(0.002)	(0.001)
$\|\widehat{\mathscr{C}}_{XZ} - \widetilde{\mathscr{C}}_{XZ}\|$	-3.718	-3.984	-4.218	-4.478	-4.712	-4.983
	(0.003)	(0.003)	(0.003)	(0.002)	(0.003)	(0.002)
$\|\widehat{\mathscr{C}}_{YZ} - \widetilde{\mathscr{C}}_{YZ}\|_{HS}$	-1.661	-2.126	-1.859	-2.625	-2.059	-3.126
	(0.031)	(0.001)	(0.013)	(0.002)	(0.003)	(0.001)
$\|\widehat{\mathscr{C}}_{YZ} - \widetilde{\mathscr{C}}_{YZ}\|$	-1.764	-2.267	-2.015	-2.768	-2.260	-3.268
	(0.030)	(0.002)	(0.015)	(0.002)	(0.005)	(0.002)

Lemma 5. *For a finite sequence of random operators $\mathscr{Z}_i$ mapping from $\mathscr{H}_1$ to $\mathscr{H}_2$, they satisfy $\mathbb{E}\mathscr{Z}_i = 0$ and $\|\mathscr{Z}_i\| \leq L$. Suppose their summation is $\mathscr{S} = \sum_i \mathscr{Z}_i$ and $\mathscr{S}^*$ is its adjoint. Let $\mathscr{V}_1$ and $\mathscr{V}_2$ be semidefinite upper bounds for $\mathrm{Var}_1(\mathscr{S})$ and $\mathrm{Var}_2(\mathscr{S})$, respectively. That is, $\mathscr{V}_1 \succeq \mathrm{Var}_1(\mathscr{S}) = \mathbb{E}(\mathscr{S}\mathscr{S}^*) = \sum_i \mathbb{E}(\mathscr{Z}_i \mathscr{Z}_i^*)$, and $\mathscr{V}_2 \succeq \mathrm{Var}_2(\mathscr{S}) = \mathbb{E}(\mathscr{S}^*\mathscr{S}) = \sum_i \mathbb{E}(\mathscr{Z}_i^* \mathscr{Z}_i)$. Define an intrinsic bound $d = \mathrm{intdim}(\mathscr{V}_1) + \mathrm{intdim}(\mathscr{V}_2)$ and a variance bound $v = \max\{\|\mathscr{V}_1\|, \|\mathscr{V}_2\|\}$. Then, for $t \geq \sqrt{v} + L/3$, $\mathbb{P}(\|\mathscr{S}\| \geq t) \leq 4d \exp\left(-\frac{t^2/2}{v + Lt/3}\right)$.*

Lemma 6. *For a sequence of random self-adjoint operators $\mathscr{Z}_i$ mapping from $\mathscr{H}_1$ to $\mathscr{H}_1$, they satisfy $\mathbb{E}\mathscr{Z}_i = 0$ and $\|\mathscr{Z}_i\| \leq L$. Suppose their summation is $\mathscr{S} = \sum_i \mathscr{Z}_i$, and $\mathscr{V}$ is the semidefinite upper bound for $\sum_i \mathbb{E}(\mathscr{Z}_i^2)$. Define an intrinsic bound $d = \mathrm{intdim}(\mathscr{V})$ and a variance bound $v = \|\mathscr{V}\|$. Then, for $t \geq \sqrt{v} + L/3$, $\mathbb{P}(\|\mathscr{S}\| \geq t) \leq 4d \exp\left(-\frac{t^2/2}{v + Lt/3}\right)$.*

2.5.1 Proof of Theorem 1

For any $u \in \mathscr{H}_X$ and $v \in \mathscr{H}_Y$, $\langle \widetilde{\mathscr{C}}_{XY} u, v \rangle = \frac{1}{CN} \sum_{c=1}^{C} \frac{1}{\tilde{p}_c} \langle \tilde{x}_c, u \rangle \langle \tilde{y}_c, v \rangle$. For each sampled summand, it holds that

$$\mathbb{E}\left[\frac{1}{\tilde{p}_c} \langle \tilde{x}_c, u \rangle \langle \tilde{y}_c, v \rangle\right] = \sum_{n=1}^{N} \langle x_n, u \rangle \langle y_n, v \rangle = N \langle \widehat{\mathscr{C}}_{XY} u, v \rangle, \quad \mathbb{E}\left[\frac{1}{\tilde{p}_c} \langle \tilde{x}_c, u \rangle \langle \tilde{y}_c, v \rangle\right]^2 = \sum_{n=1}^{N} \frac{1}{p_n} \langle x_n, u \rangle^2 \langle y_n, v \rangle^2.$$

20

Table 2.2: Randomized covariance operator estimation for the spectra dataset.

	$C = 1,000$		$C = 5,000$		$C = 10,000$	
	UNIF	IMPO	UNIF	IMPO	UNIF	IMPO
$\|\widehat{\mathscr{C}}_{XX} - \widetilde{\mathscr{C}}_{XX}\|_{HS}$	-7.64	-8.23	-7.96	-8.58	-8.12	-8.73
	(0.01)	(0.00)	(0.01)	(0.00)	(0.01)	(0.00)
$\|\widehat{\mathscr{C}}_{XX} - \widetilde{\mathscr{C}}_{XX}\|$	-7.67	-8.35	-8.00	-8.70	-8.15	-8.86
	(0.01)	(0.00)	(0.01)	(0.00)	(0.01)	(0.00)
$\|\hat{\theta}_1 - \text{sign}(\langle\hat{\theta}_1, \tilde{\theta}_1\rangle)\tilde{\theta}_1\|$	-1.71	-2.04	-2.03	-2.39	-2.17	-2.54
	(0.01)	(0.00)	(0.01)	(0.00)	(0.01)	(0.00)
$\|\hat{\theta}_2 - \text{sign}(\langle\hat{\theta}_2, \tilde{\theta}_2\rangle)\tilde{\theta}_2\|$	-0.53	-1.01	-0.99	-1.36	-1.17	-1.51
	(0.01)	(0.00)	(0.01)	(0.00)	(0.01)	(0.00)
$\|\hat{\theta}_3 - \text{sign}(\langle\hat{\theta}_3, \tilde{\theta}_3\rangle)\tilde{\theta}_3\|$	-0.35	-0.89	-0.73	-1.27	-1.03	-1.41
	(0.01)	(0.00)	(0.02)	(0.00)	(0.02)	(0.00)
$\|\hat{\theta}_4 - \text{sign}(\langle\hat{\theta}_4, \tilde{\theta}_4\rangle)\tilde{\theta}_4\|$	-0.03	-0.36	-0.21	-0.63	-0.30	-0.76
	(0.00)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)
$\|\hat{\theta}_5 - \text{sign}(\langle\hat{\theta}_5, \tilde{\theta}_5\rangle)\tilde{\theta}_5\|$	-0.02	-0.21	-0.14	-0.57	-0.21	-0.71
	(0.00)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)
CPU Time (secs)	2.60	7.69	11.46	15.78	23.42	27.38

The subsampled estimator $\widetilde{\mathscr{C}}_{XY}$ has C summands, and its related expectation is $\mathbb{E}\langle\widetilde{\mathscr{C}}_{XY}u, v\rangle = \frac{1}{CN} \times C \times \mathbb{E}\left[\frac{1}{\tilde{p}_c}\langle\tilde{x}_c, u\rangle\langle\tilde{y}_c, v\rangle\right] = \langle\widehat{\mathscr{C}}_{XY}u, v\rangle$, and the variance is

$$
\begin{aligned}
\text{Var}\langle\widetilde{\mathscr{C}}_{XY}u, v\rangle =& \frac{1}{CN^2} \times \text{Var}\left[\frac{1}{\tilde{p}_c}\langle\tilde{x}_c, u\rangle\langle\tilde{y}_c, v\rangle\right] \\
=& \frac{1}{CN^2} \times \left[\mathbb{E}\left(\frac{1}{\tilde{p}_c}\langle\tilde{x}_c, u\rangle\langle\tilde{y}_c, v\rangle\right)^2 - \left(\mathbb{E}\frac{1}{\tilde{p}_c}\langle\tilde{x}_c, u\rangle\langle\tilde{y}_c, v\rangle\right)^2\right] \\
=& \frac{1}{CN^2} \times \left[\sum_{n=1}^{N}\frac{1}{p_n}\langle x_n, u\rangle^2\langle y_n, v\rangle^2 - N^2\langle\widehat{\mathscr{C}}_{XY}u, v\rangle^2\right]. \quad (2.7)
\end{aligned}
$$

Suppose $\{\theta_{Xj}\}_{j=1}^{\infty}$ and $\{\theta_{Yj}\}_{j=1}^{\infty}$ are two groups of complete orthonormal basis for $\mathscr{H}_X$ and $\mathscr{H}_Y$, respectively. By definition of Hilbert–Schmidt norm,

$$
\begin{aligned}
\mathbb{E}\|\widetilde{\mathscr{C}}_{XY} - \widehat{\mathscr{C}}_{XY}\|_{HS}^2 =& \mathbb{E}\sum_{i,j=1}^{\infty}\langle(\widetilde{\mathscr{C}}_{XY} - \widehat{\mathscr{C}}_{XY})\theta_{Xi}, \theta_{Yj}\rangle^2 \\
=& \frac{1}{CN^2}\sum_{i,j=1}^{\infty}\left[\sum_{n=1}^{N}\frac{1}{p_n}\langle x_n, \theta_{Xi}\rangle^2\langle y_n, \theta_{Yj}\rangle^2 - N^2\langle\widehat{\mathscr{C}}_{XY}\theta_{Xi}, \theta_{Yj}\rangle^2\right]
\end{aligned}
$$

$$= \frac{1}{C} \left[\frac{1}{N^2} \sum_{n=1}^{N} \frac{\|x_n\|^2 \cdot \|y_n\|^2}{p_n} - \|\widehat{\mathscr{C}}_{XY}\|_{HS}^2 \right].$$

In the above, the second equation uses (2.7). The third equation depends on that $\|x_n\|^2 = \sum_{i=1}^{\infty} \langle x_n, \theta_{Xi} \rangle^2$ and $\|y_n\|^2 = \sum_{i=1}^{\infty} \langle y_n, \theta_{Yi} \rangle^2$. When $p_n = \frac{\|x_n\| \cdot \|y_n\|}{\sum_{m=1}^{N} \|x_m\| \cdot \|y_m\|}$ is used in this expression, we will get

$$\mathbb{E}\|\widetilde{\mathscr{C}}_{XY} - \widehat{\mathscr{C}}_{XY}\|_{HS}^2 = \frac{1}{C} \left[\frac{1}{N^2} \left(\sum_{n=1}^{N} \|x_n\| \cdot \|y_n\| \right)^2 - \|\widehat{\mathscr{C}}_{XY}\|_{HS}^2 \right].$$

To prove that this choice of p_n's minimizes $\mathbb{E}\|\widetilde{\mathscr{C}}_{XY} - \widehat{\mathscr{C}}_{XY}\|_{HS}^2$, we can find that

$$\sum_{n=1}^{N} \frac{\|x_n\|^2 \cdot \|y_n\|^2}{p_n} \overset{(i)}{=} \sum_{n=1}^{N} \left(\frac{\|x_n\| \cdot \|y_n\|}{\sqrt{p_n}} \right)^2 \cdot \sum_{n=1}^{N} (\sqrt{p_n})^2 \overset{(ii)}{\geq} \left(\sum_{n=1}^{N} \frac{\|x_n\| \cdot \|y_n\|}{\sqrt{p_n}} \cdot \sqrt{p_n} \right)^2 = \left(\sum_{n=1}^{N} \|x_n\| \cdot \|y_n\| \right)^2.$$

The equation (i) uses $\sum_{n=1}^{N} p_n = 1$. The inequality (ii) is due to Cauchy–Schwarz inequality, and the equality holds if and only if $p_n = \gamma \|x_n\| \cdot \|y_n\|$ for some $\gamma > 0$. Therefore, the optimal probability is $p_n = \frac{\|x_n\| \cdot \|y_n\|}{\sum_{m=1}^{N} \|x_m\| \cdot \|y_m\|}$.

2.5.2 Proof of Theorem 3

First of all, notice $\frac{1}{N} \sum_{n=1}^{N} \left(\|x_n\|^2 / \hat{\sigma}_{x,1}^2 + \|y_n\|^2 / \hat{\sigma}_{y,1}^2 \right) = \mathrm{tr}(\widehat{\mathscr{C}}_{XX}) / \hat{\sigma}_{x,1}^2 + \mathrm{tr}(\widehat{\mathscr{C}}_{YY}) / \hat{\sigma}_{y,1}^2 = \hat{\delta}_x + \hat{\delta}_y$. As a result, the sampling probability (2.4) satisfies $p_n \geq (\|x_n\|_2^2 / \hat{\sigma}_{x,1}^2 + \|y_n\|_2^2 / \hat{\sigma}_{y,1}^2) / (\beta(\hat{\delta}_x + \hat{\delta}_y) N)$.

Let $\mathscr{Z}_c = \frac{1}{\tilde{p}_c N} \frac{\tilde{x}_c}{\hat{\sigma}_{x,1}} \otimes \frac{\tilde{y}_c}{\hat{\sigma}_{y,1}} - \frac{\widehat{\mathscr{C}}_{XY}}{\hat{\sigma}_{x,1}\hat{\sigma}_{y,1}}$, with its adjoint operator $\mathscr{Z}_c^* = \frac{1}{\tilde{p}_c N} \frac{\tilde{y}_c}{\hat{\sigma}_{y,1}} \otimes \frac{\tilde{x}_c}{\hat{\sigma}_{x,1}} - \frac{\widehat{\mathscr{C}}_{XY}^*}{\hat{\sigma}_{x,1}\hat{\sigma}_{y,1}}$, our goal is to bound

$$\frac{1}{\hat{\sigma}_{x,1}\hat{\sigma}_{y,1}} \left[\widetilde{\mathscr{C}}_{XY} - \widehat{\mathscr{C}}_{XY} \right] = \frac{1}{\hat{\sigma}_{x,1}\hat{\sigma}_{y,1}} \times \frac{1}{C} \sum_{c=1}^{C} \left[\frac{1}{\tilde{p}_c N} \tilde{x}_c \otimes \tilde{y}_c - \widehat{\mathscr{C}}_{XY} \right] = \frac{1}{C} \sum_{c=1}^{C} \mathscr{Z}_c = \frac{1}{C} \mathscr{S},$$

with Lemma 5. It is easy to see that $\mathbb{E}\mathscr{Z}_c = 0$ and

$$\|\mathscr{Z}_c\| \leq \frac{1}{\tilde{p}_c N} \left\| \frac{\tilde{x}_c}{\hat{\sigma}_{x,1}} \otimes \frac{\tilde{y}_c}{\hat{\sigma}_{y,1}} \right\| + \left\| \frac{\widehat{\mathscr{C}}_{XY}}{\hat{\sigma}_{x,1}\hat{\sigma}_{y,1}} \right\| \leq (\hat{\delta}_x + \hat{\delta}_y)\beta \times \frac{\|\tilde{x}_c \otimes \tilde{y}_c\| / (\hat{\sigma}_{x,1}\hat{\sigma}_{y,1})}{\|\tilde{x}_c\|^2 / \hat{\sigma}_{x,1}^2 + \|\tilde{y}_c\|^2 / \hat{\sigma}_{y,1}^2} + \|\widehat{\mathscr{C}}_{XY}\| / (\hat{\sigma}_{x,1}\hat{\sigma}_{y,1})$$

22

$$\overset{(i)}{\leq} \frac{(\hat{\delta}_x + \hat{\delta}_y)\beta}{2} \times \left\| \frac{\tilde{x}_c}{\|\tilde{x}_c\|} \otimes \frac{\tilde{y}_c}{\|\tilde{y}_c\|} \right\| + \|\widehat{\mathscr{C}}_{XY}\|/(\hat{\sigma}_{x,1}\hat{\sigma}_{y,1}) \overset{(ii)}{=} \frac{\beta(\hat{\delta}_x + \hat{\delta}_y)}{2} + \hat{\eta} = L \,, \qquad (2.8)$$

In the above, (i) uses the inequality $a^2 + b^2 \geq 2ab$ for any $a, b \in \mathbb{R}$, and (ii) uses the fact the operator norm of $\frac{\tilde{x}_c}{\|\tilde{x}_c\|} \otimes \frac{\tilde{y}_c}{\|\tilde{y}_C\|}$ is equal to one.

Next, we need to control over $\mathbb{E}\mathscr{S}\mathscr{S}^* = \sum_{c=1}^{C} \mathbb{E}\mathscr{Z}_c\mathscr{Z}_c^*$ and $\mathbb{E}\mathscr{S}^*\mathscr{S} = \sum_{c=1}^{C} \mathbb{E}\mathscr{Z}_c^*\mathscr{Z}_c$. For any $u \in \mathcal{H}_1$, we have that $\mathscr{Z}_c u = \frac{1}{\hat{\sigma}_{x,1}\hat{\sigma}_{y,1}}\left[\frac{1}{\tilde{p}_c N}\langle \tilde{x}_c, u\rangle \tilde{y}_c - \widehat{\mathscr{C}}_{XY} u\right]$, and moreover,

$$\begin{aligned}
(\hat{\sigma}_{x,1}^2\hat{\sigma}_{y,1}^2)\mathscr{Z}_c^*(\mathscr{Z}_c u) &= \frac{1}{\tilde{p}_c^2 N^2}\langle \tilde{x}_c, u\rangle\langle \tilde{y}_c, \tilde{y}_c\rangle \tilde{x}_c - \frac{1}{\tilde{p}_c N}\langle \tilde{x}_c, u\rangle\widehat{\mathscr{C}}_{XY}^*\tilde{y}_c - \frac{1}{\tilde{p}_c N}\langle \tilde{y}_c, \widehat{\mathscr{C}}_{XY}u\rangle\tilde{x}_c + \widehat{\mathscr{C}}_{XY}^*\widehat{\mathscr{C}}_{XY}u \\
&= \frac{1}{\tilde{p}_c^2 N^2}\langle \tilde{x}_c, u\rangle\langle \tilde{y}_c, \tilde{y}_c\rangle \tilde{x}_c - \frac{1}{\tilde{p}_c N}\widehat{\mathscr{C}}_{XY}^*(\tilde{x}_c \otimes \tilde{y}_c)u - \frac{1}{\tilde{p}_c N}(\tilde{x}_c \otimes \tilde{y}_c)^*\widehat{\mathscr{C}}_{XY}u + \widehat{\mathscr{C}}_{XY}^*\widehat{\mathscr{C}}_{XY}u \,.
\end{aligned}$$

Notice $\mathbb{E}\left[\frac{1}{\tilde{p}_c N}\widehat{\mathscr{C}}_{XY}^*(\tilde{x}_c \otimes \tilde{y}_c)u\right] = \mathbb{E}\left[\frac{1}{\tilde{p}_c N}(\tilde{x}_c \otimes \tilde{y}_c)^*\widehat{\mathscr{C}}_{XY}u\right] = \widehat{\mathscr{C}}_{XY}^*\widehat{\mathscr{C}}_{XY}u$. As a result,

$$(\hat{\sigma}_{x,1}^2\hat{\sigma}_{y,1}^2)\mathbb{E}[\mathscr{Z}_c^*(\mathscr{Z}_c u)] = \sum_{n=1}^{N} \frac{1}{p_n N^2}\langle x_n, u\rangle\langle y_n, y_n\rangle x_n - \widehat{\mathscr{C}}_{XY}^*\widehat{\mathscr{C}}_{XY}u.$$

It follows that

$$\begin{aligned}
\mathbb{E}[\mathscr{Z}_c^*\mathscr{Z}_c] &= \frac{1}{\hat{\sigma}_{x,1}^2\hat{\sigma}_{y,1}^2}\left[\sum_{n=1}^{N} \frac{1}{p_n N^2}\langle y_n, y_n\rangle x_n \otimes x_n - \widehat{\mathscr{C}}_{XY}^*\widehat{\mathscr{C}}_{XY}\right] \preceq \frac{1}{\hat{\sigma}_{x,1}^2\hat{\sigma}_{y,1}^2}\sum_{n=1}^{N} \frac{1}{p_n N^2}\langle y_n, y_n\rangle x_n \otimes x_n \\
&\preceq \frac{\beta(\hat{\delta}_x + \hat{\delta}_y)}{\hat{\sigma}_{x,1}^2 N}\sum_{n=1}^{N} \frac{\langle y_i, y_i\rangle/\hat{\sigma}_{y,1}^2}{\|x_n\|^2/\hat{\sigma}_{x,1}^2 + \|y_n\|^2/\hat{\sigma}_{y,1}^2}x_n \otimes x_n \preceq \frac{\beta(\hat{\delta}_x + \hat{\delta}_y)}{\hat{\sigma}_{x,1}^2}\widehat{\mathscr{C}}_{XX} \,. \qquad (2.9)
\end{aligned}$$

Now we get that $\mathbb{E}\mathscr{S}\mathscr{S}^* \preceq \mathscr{V}_1 := \frac{C\beta(\hat{\delta}_x + \hat{\delta}_y)}{\hat{\sigma}_{x,1}^2}\widehat{\mathscr{C}}_{XX}$. It has operator norm $\|\mathscr{V}_1\| = C\beta(\hat{\delta}_x + \hat{\delta}_y)$ and intrinsic dimension $\text{intdim}(\mathscr{V}_1) = \text{intdim}(\widehat{\mathscr{C}}_{XX}) = \hat{\delta}_x$. With a similar argument, we can deduce that $\mathbb{E}\mathscr{S}^*\mathscr{S} \preceq \mathscr{V}_2 := \frac{c\beta(\hat{\delta}_x + \hat{\delta}_y)}{\hat{\sigma}_{y,1}^2}\widehat{\mathscr{C}}_{YY}$. It has operator norm $\|\mathscr{V}_2\| = C\beta(\hat{\delta}_x + \hat{\delta}_y)$ and intrinsic dimension $\text{intdim}(\mathscr{V}_2) = \text{intdim}(\widehat{\mathscr{C}}_{YY}) = \hat{\delta}_y$.

Now denote $v = \max\{\|\mathscr{V}_1\|, \|\mathscr{V}_2\|\} = C\beta(\hat{\delta}_x + \hat{\delta}_y)$, $d = \text{intdim}(\mathscr{V}_1) + \text{intdim}(\mathscr{V}_2) = \hat{\delta}_x + \hat{\delta}_y$ and $L = \frac{\beta(\hat{\delta}_x + \hat{\delta}_y)}{2} + \hat{\eta}$ from (2.8). Applying Lemma 5, for ϵ satisfying $\epsilon \cdot C \geq \sqrt{C\beta(\hat{\delta}_x + \hat{\delta}_y)} +$

$\beta(\hat{\delta}_x + \hat{\delta}_y)/6 + \hat{\eta}/3$, we have that

$$\mathbb{P}\Big(\|\widetilde{\mathscr{C}}_{XY} - \widehat{\mathscr{C}}_{XY}\| \geq \epsilon \cdot \hat{\sigma}_{x,1}\hat{\sigma}_{y,1}\Big)$$

$$=\mathbb{P}\left(\|\mathscr{S}\| \geq C\epsilon\right) \leq 4(\hat{\delta}_x + \hat{\delta}_y) \exp\Big(-\frac{C^2\epsilon^2/2}{C\beta(\hat{\delta}_x + \hat{\delta}_y) + [\beta(\hat{\delta}_x + \hat{\delta}_y)/2 + \hat{\eta}] \times C\epsilon/3}\Big)$$

$$=4(\hat{\delta}_x + \hat{\delta}_y) \exp\Big(-\frac{C\epsilon^2}{2\beta(\hat{\delta}_x + \hat{\delta}_y)(1 + \epsilon/6) + 2\hat{\eta}\epsilon/3}\Big).$$

2.5.3   Proof of Theorem 4

Our goal is to apply Lemma 6 to $\frac{1}{\hat{\sigma}_{x,1}^2}\Big[\widetilde{\mathscr{C}}_{XX} - \widehat{\mathscr{C}}_{XX}\Big] = \frac{1}{\hat{\sigma}_{x,1}^2} \times \frac{1}{C}\sum_{c=1}^{C}\Big[\frac{1}{\tilde{p}_c N}\tilde{x}_c \otimes \tilde{x}_c - \widehat{\mathscr{C}}_{XX}\Big] = \frac{1}{C}\sum_{c=1}^{C}\mathscr{Z}_c = \frac{1}{C}\mathscr{S}$.

In the above, we have set $\mathscr{Z}_c := \frac{1}{\hat{\sigma}_{x,1}^2}\Big[\frac{1}{\tilde{p}_c N}\tilde{x}_c \otimes \tilde{x}_c - \widehat{\mathscr{C}}_{XX}\Big]$. It has the operator norm upper bound $\|\mathscr{Z}_c\| \leq \frac{1}{N\tilde{p}_c} \cdot \frac{\|\tilde{x}_c\|^2}{\hat{\sigma}_{x,1}^2} + 1$. At the same time, for the probability $p_n \geq \frac{1}{\beta}\frac{\|x_n\|^2}{\sum_{m=1}^{N}\|x_m\|^2}$ in the theorem, we have

$$\tilde{p}_c \geq \|\tilde{x}_c\|^2/(\beta\sum_{n=1}^{N}\|x\|^2) = \|\tilde{x}_c\|^2/(\beta N \cdot \text{tr}(\widehat{\mathscr{C}}_{XX})). \tag{2.10}$$

This implies

$$\|\mathscr{Z}_c\| \leq \beta\text{tr}(\widehat{\mathscr{C}}_{XX})/\hat{\sigma}_{x,1}^2 + 1 = \beta\hat{\delta}_x + 1. \tag{2.11}$$

Similar to the derivation of (2.9), it holds that

$$\mathbb{E}[\mathscr{Z}_c^*\mathscr{Z}_c] = \frac{1}{\hat{\sigma}_{x,1}^4}\Big[\sum_{n=1}^{N}\frac{1}{p_n N^2}\langle x_n, x_n\rangle x_n \otimes x_n - \widehat{\mathscr{C}}_{XX}^*\widehat{\mathscr{C}}_{XX}\Big] \preceq \frac{1}{\hat{\sigma}_{x,1}^4}\sum_{n=1}^{N}\frac{1}{p_n N^2}\langle x_n, x_n\rangle x_n \otimes x_n. \tag{2.12}$$

Together with (2.10), it holds that

$$\mathbb{E}[\mathscr{Z}_c^*\mathscr{Z}_c] \preceq \frac{1}{N}\sum_{n=1}^{N}\beta\hat{\delta}_x \times \frac{x_n \otimes x_n}{\hat{\sigma}_{x,1}^2} = \mathscr{V} := (\beta\hat{\delta}_x/\hat{\sigma}_{x,1}^2)\widehat{\mathscr{C}}_{XX}. \tag{2.13}$$

Therefore, $\|\mathscr{V}\| = C(\beta\hat{\hat{\delta}}_x/\hat{\sigma}_{x,1}^2)\|\mathscr{C}_{XX}\| = C\beta\hat{\hat{\delta}}_x$, and $\mathrm{intdim}(\mathscr{V}) = \hat{\hat{\delta}}_x$. Combining (2.11) and (2.13) with Lemma 6, we conclude that for ϵ satisfying $C \cdot \epsilon > \sqrt{C\beta \cdot \hat{\hat{\delta}}_x} + (\beta\hat{\hat{\delta}}_x + 1)/3$, it holds that

$$\mathbb{P}\left(\left\|\widetilde{\mathscr{C}}_{XX} - \widehat{\mathscr{C}}_{XX}\right\| \geq \epsilon\hat{\sigma}_{x,1}^2\right) \leq 4\hat{\hat{\delta}}_x \exp\left(-\frac{\epsilon^2 C^2/2}{C\beta\hat{\hat{\delta}}_x + (\beta\hat{\hat{\delta}}_x + 1)C\epsilon/3}\right) = 4\hat{\hat{\delta}}_x \exp\left(-\frac{\epsilon^2 C/2}{\beta\hat{\hat{\delta}}_x + (\beta\hat{\hat{\delta}}_x + 1)\epsilon/3}\right).$$

3. ICE: INDEPENDENT COMPONENT ESTIMATION OF SPECTROSCOPIC DATA OF TYPE IA SUPERNOVAE

With the advent of large dedicated Type Ia supernova (SN Ia) surveys, it is possible to statistically improve the SN Ia spectral energy distribution (SED) and investigate the SN Ia intrinsic properties. A purely data-driven statistical approach conducting such tasks can enable a direct utilization of the data product and contribute to a better understanding of SN Ia intrinsic properties. Based on a collection of SN Ia spectra from the Nearby Supernova Factory collaboration (SNFactory), we develop the Independent Component Estimation (ICE) method for sparse and irregularly spaced spectrophotometric data of SNe Ia using functional principal component analysis (FPCA) and independent component analysis (ICA). The application of FPCA facilitates a unique low-rank representation of SN Ia data by decomposing the data into a linear combination of multiple components. The ICA allows for the separation of independent physical effects such as the dust extinction and intrinsic color differences of SNe Ia. This separation makes it possible to construct the intrinsic SED manifolds of SNe Ia. A mean extinction curve is derived based on ICA which is in excellent agreement with the dust extinction laws found in the Milky Way but with a significantly steeper wavelength dependence. The intrinsic SED manifolds also allows for the calculation of the dust extinction properties of each individual SN. Typical values of the total-to-selective extinction ratio R_V are found to be around 1-2 for the intrinsic SED constructed with ICA, which is in agreement with independently derived R_V values of several well observed nearby SNe with NIR data. A numerical toolbox is built to calculate A_V and R_V for any SNe with well calibrated spectral sequence. The toolbox simultaneously builds an intrinsic spectral time series for the SN. With this tool, the color correction in SN Ia distance standardization can be performed separately for the intrinsic color and the host galaxy reddening.

3.1 Introduction

Type Ia supernovae (SNe Ia) have played an important role in cosmological studies. The luminosity distance versus redshift relation derived from SNe Ia provides powerful evidences for the acceleration of the expansion of the Universe and SNe Ia continue to be one of the most critical tools in cosmology [36, 37, 38]. The luminosity dispersion of SNe Ia is naturally low and can be further reduced by well-addressed procedures [39, 40]. They are visible out to a high redshift, making them precise distance indicators for cosmological studies. With the advent of current and next-generation dedicated SN Ia surveys such as the LSST [41] and the WFRIST/Roman Telescope [42], the number of well observed SNe Ia will increase dramatically, and the SNe will be complimentary to various other astronomical probes in examining and distinguishing candidate dark energy models.

There is a pressing need to improve the cosmological application of SNe Ia to understand the systematic uncertainties encoded in the measurements. A number of studies of the heterogeneity of SNe Ia have been carried out in the literature based largely on light curve shape and intrinsic luminosity correlations. The light curve shape is measured mainly by parameters describing the decline rate after optical maximum and the color around optical maximum. The light curves of intrinsically brighter SN Ia tend to be broader and bluer [43, 44, 39]. Several metrics are developed to describe these correlations, such as the magnitude decline from peak to at 15 days after B-band maximum, Δm_{15} [40] and the light curve stretch factor s [45, 46]. The treatment of the luminosity-color dependence is ambivalent as the intrinsic color variations among different SNe Ia are difficult to be decoupled from the reddening due to host galaxy extinction. To derive the host galaxy reddening, a subset of SNe Ia from hosts of late type galaxies is needed to define a extinction free sub-sample SNe Ia [36, 47, 48]. Alternatively, color corrections can also be applied without distinguishing the difference between the intrinsic color and dust reddening [49, 50, 51]. However, the confounding of the intrinsic color and dust reddening presents a systematic limitation to the application of SN Ia in cosmological study [52]. Generally speaking, the reddening effect follows the extinction law [2, hereafter CCM] and [53], which can be expressed as an exponential

multiplier to the flux depending on the ratio of total to selective dust extinction parameter R_B. The average value of R_B is ~ 4.1 for the interstellar dust in the Milky Way but appears to have smaller values for extragalactic dust [54, 55], whereas the equivalent color correction coefficients that minimizes residuals on the Hubble diagrams is found to be around 2.0 [56]. This discrepancy remains poorly understood. The dust reddening may further subject to complications introduced by the presence of dust located in the vicinity of the SNe [57, 58, 59, 60, 61].

Though a handful of well-established standardization procedures have been generally applied to correct the light curve broadness-color heterogeneity, the dissimilarity among spectral features does not follow a simple relation. There have been several empirical attempts to address this problem. [62] constructed a mean spectral time series template by interpolating the observations in two dimensional time and wavelength grids, re-matching the photometric and spectroscopic measurements, and taking the average over available spectra. A "mangling" function is used to rescale the observed spectral data. SALT2 and SALT3 [50, 51] model the spectral energy distribution (SED) by the sum of a two principle components, including a mean surface and a surface that captures the dominant spectral variation. A time-independent color term c is defined to capture the color offset between the flux surface of a fiducial SN with first order correction constructed by the principle component analysis (PCA). A nonlinear transformation function is found for the relation between the dominant component x_1 of SALT2 to Δm_{15} or the stretch parameter s. In a recent work, the SNEMO [63] uses Gaussian Process to interpolate the non-uniformly collected observations to generate a equally gridded spectral time series, and examines the time series further with multivariate factor analysis. The color difference is removed by aligning the SN spectra to an average fiducial supernova assuming an extinction law with $R_V = 3.1$ [63]. Three SNEMO models have been finally proposed based on different criteria, including two-component model SNEMO2 for comparison purpose, seven-component model SNEMO7 determined by standardizing supernova magnitudes, and fifteen-component model SNEMO15 chosen according to Akaike Information Criterion [64], which maximizes the information encoded in the spectral time series. However, in either SALT3 or SNEMO, the fiducial models themselves do not distinguish intrinsic

color and dust reddening. In SALT3, the requirement of c to be independent of x_1 helps to isolate the effect of dust extinction from that of the intrinsic color, but the effectiveness of this method is compromised as the assumption of the color correction to be following certain functional forms of wavelength dependency will lead to the eigenvectors of the PCA to be color dependent. For the same reason, the use of a single average fiducial supernova in SNEMO to account for the color corrections results in parameters and SN flux surfaces that are corrected by a mixture of dust and intrinsic color effect. The methods do not lead to estimates of the intrinsic flux surfaces of SNeIa.

Deep neural network approaches have also been developed recently [65, 66, 67]. These models show great promises in capturing the essential information contained in a multi-epoch data sequence of SNe Ia, and eventually link the observed data with certain theoretical models such as TARDIS [68, 65].

In this work, we develop a data-driven tool to decouple the SN Ia spectral-feature-dependent intrinsic properties from dust reddening mainly based on functional principal component analysis [69, FPCA] and assisted by independent component analysis [70, 71, ICA]. Functional principal component analysis treats irregular observations of SN Ia SED as realizations of some random function over the domain generated by the tensor product of wavelength and phase dimension and extracts the dominant modes residing within a function space. No additional SED interpolation and evaluation are needed to fulfill the requirement of common-grid measurements in multivariate principal component analysis, which mitigates the error propagating throughout the training process. Using FPCA, the SN Ia SED is represented as a linear combination of a mean surface and a few principal component surfaces and the corresponding coefficients are called principal component scores, providing a unique low-rank representation of SNe Ia. FPCA was adopted by SALT2 where third-order B splines were used to represent the spectral features and their evolution over time. Alternatively, to increase the model capability of describing the localized and multi-resolutional spectral feature strengths in a more stable, robust and compact way, we design a wavelet-spline tensor product basis during the model training process, consisting of a over-complete fourth-order wavelet basis in the wavelength dimension and a B-spline basis in the phase dimension.

Two models are proposed, including a broadly considered flux-based FPCA model, referred as fb-FPCA model, which models the spectrophotometric data with original flux unit and a magnitude-based FPCA model, referred as mb-FPCA model, which considers the SED decomposition in the logarithmic scale (magnitude). The motivation of mb-FPCA model is to adapt to the additive nature of the dust extinction law and extract it from the rest of the signals. To enable an effective construction of the physical surface that is independent of time and least correlated to the surfaces governing the intrinsic time and spectral variations and in turn a better separation of dust reddening effect and supernova intrinsic properties, ICA is applied to the imputed spectrophotometric data for a desired SED representation that minimizes the statistical dependence between the latent components.

As one of the most important contributions of this work, a novel Independent Component Estimation (ICE) procedure of SNe Ia is proposed to decouple the effects due to dust extinction and intrinsic SN property. Several tasks are conducted to achieve this goal:

I. Construct a dust extinction free SN Ia flux ratio surface to extract SN Ia intrinsic property parameters. It opens the possibility of analyzing SN Ia intrinsic property without considering the confounding of dust along the line of sight.

II. Extract the heuristic color correction curves (CCRs) from the spectrophotometric data. The construction of the CCRs is based on FPCA and ICA, with the working hypothesis that the effects of dust extinction and the SN intrinsic properties originate from distinct independent physical processes, and that the extinction curves are smooth functions of wavelength and time-independent. ICA is used as it enables a better signal separation of signals in data with mixed physical origins.

III. Propose and statistically test two simplified hypotheses about the contribution of the dust extinction to the CCRs. This enables us to quantify the intrinsic color and interstellar dust reddening separately.

IV. Derive the mean extinction curves with FPCA and ICA after removing the SN Ia

intrinsic color effect. The mean extinction curves encode the averaged effect of dust reddening of the training dataset.

V. Construct the intrinsic spectrophotometric surface of SNe Ia by correcting the effect due to dust extinction. A comprehensive low-rank representation of the intrinsic surface is also provided.

VI. Calculate the total extinction A_V in V-band and the ratio of absolute to selective extinction R_V under the parametrization of CCM [2] and characterize their distributions.

The chapter is structured as below. In Section 3.2, we introduce the data analyzed in this work. The mathematical model formulation, model training implementation and SED fitting of the FPCA are described in Section 3.3. We proceed in Section 3.4 to report the model selection result and compare the parameters estimated from different models for the original data. Our main contribution is presented in Section 3.5 where we construct a data-driven pipeline to study SN Ia intrinsic property and separate the interstellar dust effect from SN Ia intrinsic features.

3.2 The Type Ia Supernova Dataset

The supernovae data used in this work are spectrophotometric time series observed by The Nearby Supernova Factory [72] between 2004 and 2013 and published in [73]. The data product have been analyzed by SNEMO [63] and SUGAR [74] companion projects. Our work is based on the SNEMO training data. The supernova observations are flux-calibrated, corrected for Milky Way dust extinction and shifted to a common rest-frame at $z = 0$, as described in the [63].

The original SNEMO training dataset consists of 1759 spectra for 122 SNe Ia, with redshift ranging from $z = 0.01$ to 0.08. Peculiar SNe, such as SN 1991T and SN 1991bg-like objects are also included in the training dataset. We restrict our study to the spectral range from 3500Å to 8000Å and the phase coverage is chosen to be $[-10, 20]$ with respect to the phase of B band maximum. To ensure an adequate statistical model construction, the SNe Ia in the training dataset are required to have more than 5 valid spectra. A total number of 1087 spectra of 122 SNe Ia meet

the selection criteria and are used for further analysis.

Essential data processing procedures, such as light curve decline rate alignment and the corresponding variance spectra generation, have been carried out before applying the FPCA. These are described below:

Decline rate alignment. We scaled each spectrum to the mean B-band light curve [75] to reduce the spectral diversity introduced by the light curve decline rate. This mean light curve function was originally developed by identifying a function which describes the the averaged behaviour of the SN Ia light curves and has a peak B-band magnitude of zero. The scaling procedure was proceeded by multiplying a wavelength independent factor to the original spectrum to match the mean B-band light curve at the corresponding light curve phase. With this simple photometric scaling, the spectral features are preserved and the time axis is stretched or squeezed to match the mean light curve of [75].

Variance spectra generation. With a flux-calibrated training set, we generated the variance spectra based on the scrolling window approach with a fixed window size h such that the local variance of one pixel is determined by its $h - 1$ nearest neighborhoods. This self-generated error spectra is similar to what has been used previously in the wavelet decomposition of SN spectra [76]. We did not use the error spectra of the original data [73] as the self-generated errors, though not as accurate, are applicable to a broad range of data without error spectra. As in wavelet decomposition in [76], a preliminary true signal estimation is required so that pixel to pixel fluctuations can be quantified. We fitted the individual spectrum using least absolute shrinkage and selection operator, known as Lasso [77], with the design matrix constructed from an over-complete fourth-order wavelet basis, which will be discussed in Section 3.3.2. Lasso encourages a sparse estimation of the regression coefficients through penalizing the least squares criterion with an additional l_1 norm of the coefficient vector and it's commonly used for basis selection. Once the true signal was estimated and got removed from the raw observations, the variance for a pixel was calculated by averaging the squared measurements of its closest $h - 1$ neighborhoods.

3.3 The Statistical Model

We aim at modeling the averaged behaviour of SN Ia spectrophotometric time series and its dominant variation with a few components in a purely data-driven fashion. In this section, we will present the mathematical formulation of the FPCA model, the details of the model training and the new SN Ia SED surface fitting procedures.

3.3.1 Model Construction

Functional data analysis provides a reliable toolbox to represent and analyze spectrophotometric time series data in the form of functions. Our model construction is based on FPCA, which extracts the dominant modes of the data variation. Let $f^{(s)}(\lambda, t)$ be the calibrated SED surface of the s-th supernova, which is a function of rest-frame wavelength λ and rest-frame time t with respect to its B-band maximum flux (i.e. phase). We represent the surface using a truncated version of the functional principal component expansion [69]

$$f^{(s)}(\lambda, t) = \phi_0(\lambda, t) + \sum_{r=1}^{R} \beta_r^{(s)} \phi_r(\lambda, t), \qquad (3.1)$$

where $\phi_0(\lambda, t)$ is the mean spectral surface and $\phi_1(\lambda, t), \ldots, \phi_R(\lambda, t)$ are R principal component surfaces, which depict the dominant variations shared by the supernova population under study. The elements of coefficient vector $\boldsymbol{\beta} = \left(\beta_1^{(s)}, \ldots, \beta_R^{(s)}\right)$ are called the principal component scores (or scores for short). Additional assumptions are imposed on the principal component surfaces to ensure model identifiability and promote interpretability. In particular, we require $\phi_r(\lambda, t)$'s to be orthonormal to each other, i.e. $\int \phi_r(\lambda, t)\phi_{r'}(\lambda, t)d\lambda dt = \delta_{rr'}$ with $\delta_{rr'} = 1$ if $r = r'$ and $\delta_{rr'} = 0$ otherwise. With these properties, each object is uniquely represented by a few number of scores, which will facilitate the study of supernova properties.

In the model (3.1), functional basis expansion is applied directly to spectra in flux scale, leading to a flux-based FPCA model (denoted as fb-FPCA). Since this model is a linear combination of different components, the effect of dust extinction is not expected to be explicitly isolated as one particular component. This may pose a problem in model interpretability that the model confounds

the color variation due to dust extinction with the variation due to intrinsic SN heterogeneity.

As the dust extinction is usually expressed as an exponential multiplier to the flux, we build an alternative model in the multiplicative form of components,

$$f^{(s)}(\lambda, t) = \phi_0(\lambda, t) \prod_{r=1}^{R} \exp\left(\beta_r^{(s)} \phi_r(\lambda, t)\right). \tag{3.2}$$

This is equivalent to applying the model (3.1) to the spectrophotometric data in the magnitude scale, i.e., we take logarithm of (3.2) to get

$$\log\left(f^{(s)}(\lambda, t)\right) = \phi_0(\lambda, t) + \sum_{r=1}^{R} \beta_r^{(s)} \phi_r(\lambda, t), \tag{3.3}$$

where the mean spectral surface, principal component surfaces and supernova-specific scores are defined similarly as those of model (3.1). This model is referred to as the magnitude-based FPCA model (abbreviated as mb-FPCA).

In summary, we build two FPCA models: the fb-FPCA model (3.1) and mb-FPCA model (3.2). For both models, the mean spectral surface, principal component surfaces, and the scores of each supernova are unknown parameters to be estimated.

3.3.2 Model Training

The procedure to estimate the unknown model parameters using the flux-calibrated spectra observations is referred to as model training and will be presented in this section. The methodology and terminologies developed here are for model (3.1) but can be extended to model (3.3) in a straightforward way.

The training spectrophotometric observations are irregularly observed in the two dimensional wavelength-phase space. In particular, the spectral time series of the s-th supernova is recorded at a wavelength-phase pair $(\lambda_i^{(s)}, t_j^{(s)})$ with flux $y_{i,j}^{(s)}$ and uncertainty $\sigma_{i,j}^{(s)}$ for $i \in \{1, \cdots I_j^{(s)}\}$ and $j \in \{1, \cdots J^{(s)}\}$. We decompose the observed value $y_{i,j}^{(s)}$ as the summation of the signal surface

and a stochastic noise term, i.e.,

$$y_{i,j}^{(s)} = f^{(s)}(\lambda_i^{(s)}, t^{(s)}) + \sigma_{i,j}^{(s)} \epsilon_{i,j}^{(s)}, \tag{3.4}$$

where the signal surface $f^{(s)}(\lambda_i^{(s)}, t_j^{(s)})$ has the FPCA expansion (3.1), $\sigma_{i,j}^{(s)}$ is the corresponding measurement uncertainty, and $\epsilon_{i,j}^{(s)}$ is a random term with zero mean and unit variance.

The spectral surface $f^{(s)}(\lambda, t)$ of a supernova has two salient features: 1) high frequency and localized variability along the dimension of wavelength λ; 2) smooth evolution along the dimension of phase t. The estimated mean surface and component surface should capture these characteristics effectively.

For this purpose, we use a wavelet-spline tensor product basis to expand the mean and component surfaces. For the wavelength dimension, a fourth order wavelet basis $\boldsymbol{W}(\lambda) = \{w_1(\lambda), w_2(\lambda), \ldots, w_p(\lambda)\}^\top$ of [78] is utilized, which is over-complete and highly redundant. The over-completeness is usually coupled with sparse wavelet coefficient to only capture the significant but localized spectral features (i.e. emission and absorption lines) and to remove small scale noise. On the other hand, we use an orthonormal cubic spline basis $\boldsymbol{B}(t) = \{b_1(t), b_2(t), \ldots, b_q(t)\}^\top$ for the phase dimension. By utilising the tensor product of the two sets of bases, we represent the model (3.1) as

$$f^{(s)}(\lambda, t) = \underbrace{\boldsymbol{W}(\lambda)^\top \Theta_0 \boldsymbol{B}(t)}_{\phi_0(\lambda, t)} + \sum_{r=1}^{R} \beta_r^{(s)} \times \underbrace{\boldsymbol{W}(\lambda)^\top \Theta_r \boldsymbol{B}(t)}_{\phi_r(\lambda, t)}, \tag{3.5}$$

where the mean surface and the component surfaces are parameterized by coefficient matrices Θ_r of size $p \times q$, for $r = 0, \ldots, K$. A more comprehensible expression of Equation (3.5) is based on the wavelet-spline tensor product basis presented as following,

$$f^{(s)}(\lambda, t) = \boldsymbol{W}(\lambda)^\top \otimes \boldsymbol{B}(t)^\top \text{vec}(\Theta_0) + \boldsymbol{W}(\lambda)^\top \otimes \boldsymbol{B}(t)^\top \boldsymbol{\Theta} \boldsymbol{\beta}^{(s)}, \tag{3.6}$$

where the operator $\text{vec}(\cdot)$ converts a matrix into a column vector by stacking the transpose of the rows on top of one another. In addition, $\boldsymbol{\Theta}$ is a $pq \times R$ matrix, storing all the parameters in

the coefficient matrix of R principal component surfaces, i.e., $\boldsymbol{\Theta} = \big(\text{vec}(\Theta_1), \ldots, \text{vec}(\Theta_R) \big)$. The coefficient vector $\boldsymbol{\beta}^{(s)} = \big(\beta_1^{(s)}, \ldots, \beta_R^{(s)} \big)^\top$ stores principal component scores and is unique for each supernova. We require the coefficient matrix $\boldsymbol{\Theta}$ to be orthonormal, i.e., $\boldsymbol{\Theta}^T \boldsymbol{\Theta} = \mathbf{I}_R$, to guarantee the orthonormality of the principal component surfaces and identifiability of the statistical model.

All the parameters are estimated by minimizing weighted least squares. Two additional regularization penalties are added to the weighted least square criterion: a group Lasso penalty [79] on the rows of Θ_r to encourage the selection of important spectral feature; and a roughness penalty [69] on the columns of Θ_r to encourage the smooth evolution over phase. We use the fast iterative shrinkage-thresholding algorithm [80, FISTA] to estimate the mean surface. After that, we combine a stochastic gradient descent algorithm [28, ADAM] with a manifold optimization technique to estimate the principal component surfaces efficiently. See Supplemental Materials 3.6.1 for details.

3.3.3 Spectrophotometric Data Fitting

After the model training described in Section 3.3.2, we obtain the estimation of the mean and principal component surfaces $\{\phi_r(\lambda, t)\}$, for $r = 0, \ldots, R$. They can be applied to fit new spectrophotometric data. The empirical estimations are denoted by $\{\widehat{\phi}_r(\lambda, t)\}_{r=0}^R$ and are considered as fixed during new data fitting. The construction of the full SED surface reduces to the evaluation of the principal component score $\boldsymbol{\beta} = (\beta_1, \cdots, \beta_R)^\top$ assuming that the data have been processed through the procedures discussed in Section 3.2.

We use maximum likelihood estimate (MLE) method to derive the estimated parameter $\boldsymbol{\beta}$ and provide uncertainty quantification for the estimated parameter using parametric bootstrap method [81]. To facilitate the estimation of unknown parameters $\boldsymbol{\beta}$ for the new supernova, we assume that the random term $\epsilon_{i,j}$ in Equation (3.4) follows the standard normal distribution. The score of the new supernova $\widehat{\boldsymbol{\beta}} = \big(\widehat{\beta}_1, \ldots, \widehat{\beta}_R \big)^\top$ is determined by MLE method, which is equivalent to generalizing the least squares by combining Equation (3.4) and Equation (3.5). The parametric bootstrap procedure with B bootstrap samples is described as follows. For $b = 1, \ldots, B$, we generate a new set of spectral data at its own observational wavelength-phase grid according to the

model,

$$y_{i,j}^{(*b)} = \widehat{\phi}_0(\lambda_i, t_j) + \sum_{r=1}^{R} \widehat{\beta}_r \widehat{\phi}_r(\lambda_i, t_j) + \sigma_{i,j} \epsilon_{i,j}^{(*b)}$$

where $\epsilon_{i,j}^{(*b)}$ is drawn independently from the standard normal distribution $\mathcal{N}(0,1)$. The bth boot-strap sample is the collection of sampled flux $\{y_{i,j}^{(*b)}\}$ for this supernova, its observed wavelength-phase pairs $\{\lambda_i, t_j\}$ and observation uncertainty $\{\sigma_{i,j}\}$ for $i \in \{1, \ldots, I_j\}$ and $j \in \{1, \ldots, J\}$. The estimation procedure described above is applied to each of bootstrap sample, leading to the B bootstrapped parameter estimates $\widehat{\boldsymbol{\beta}}^{(*1)}, \ldots, \widehat{\boldsymbol{\beta}}^{(*B)}$. The element-wise standard deviation of $\widehat{\boldsymbol{\beta}}^{(*1)}, \ldots, \widehat{\boldsymbol{\beta}}^{(*B)}$ serves as an estimate of the uncertainty of $\widehat{\boldsymbol{\beta}}$ for the new supernova data.

3.4 The FPCA Surfaces of the Original Data

In this section, we present the model training results of the fb-FPCA and mb-FPCA models for the original data with no color correction but the data are processed as described in Section 3.2. The model selection problem, i.e., the determination of the number of components R and the interpretation of the estimated principal component surfaces is discussed. We also compare the proposed two FPCA models, SALT2 and two-component SNEMO model (SNEMO2) in terms of parameter interpretation. For simplicity, we denote the estimated mean SED surface by $\phi_0(\lambda, t)$ and the rth principal component SED surface by $\phi_r(\lambda, t)$ for $r = 1, 2, \ldots, R$. The estimated rth principal component scores for fb-FPCA and mb-FPCA models are represented by β_r^{fb} and β_r^{mb} respectively.

3.4.1 Model Selection

We use cross-validation [82, 83, 84] to select the number of components (R). It is a commonly used method to select the optimal hyper-parameters and to evaluate model prediction performance. The procedure of K-fold cross-validation is: randomly partitioning dataset into K disjoint subsets, training the model using $K-1$ subsets with prefixed hyper-parameter, evaluating the model on the remaining subset, repeating the procedure K times using different training and test partitions, and ultimately averaging the K evaluation metrics. The average evaluation metric (referred to as cross-validation score) is calculated for every candidate hyper-parameter, and the optimal hyper-

parameter is selected as the one minimizing the cross-validation score.

We use $\boldsymbol{\xi}$ to denote the penalty parameters η, ρ in Supplemental Materials 3.6.1. For our models, both $\boldsymbol{\xi}$ and the number of components R need to be tuned during the model training. For any candidate hyper-parameter pair $(\boldsymbol{\xi}, R)$, we conduct K-fold cross-validation with the model training (in Section 3.3.2) and data fitting (in Section 3.3.3) procedures. In particular, the full SED surface $\widehat{f}^{(s)}(\lambda, t)$ is fitted for each supernova in the test set, and the fitted surface is employed to predict its actual observations. The overall cross-validation score (CV) is calculated as,

$$\mathrm{CV}(\boldsymbol{\xi}, R) = \frac{1}{K} \sum_{k=1}^{K} \frac{1}{|\mathscr{S}_k|} \sum_{s \in \mathscr{S}_k} \frac{1}{I_j^{(s)} \times J^{(s)}} \sum_{i,j} \left[y_{i,j}^{(s)} - \widehat{f}^{(s)}(\lambda_i^{(s)}, t_j^{(s)}) \right]^2, \qquad (3.7)$$

where the outermost summation corresponds to the K cross-validation repetitions, $\mathscr{S}_k$ is the index set (with cardinality $|\mathscr{S}_k|$) of test supernovae in the k-th fold, and the innermost summation computes the prediction error for the s-th supernova in the test data of the k-th fold.

To determine the optimal R using equation (3.7), we define a $\boldsymbol{\xi}$-free metric called *optimal CV score*, $\mathrm{opCV}(R) := \min_{\boldsymbol{\xi}} \mathrm{CV}(\boldsymbol{\xi}, R)$, which minimizes the cross-validation score with respect to the hyper-parameter $\boldsymbol{\xi}$ for a given R. It accesses the model performance with a fixed R under the best-case scenarios across different values of $\boldsymbol{\xi}$. The optimal R is then selected as the one with the minimal $\mathrm{opCV}(R)$ value. We note that, unlike the classical PCA method where the selection of R is based on the proportion of in-sample variance explained by the model, this new metric focuses on the decrease of the prediction error by adding R components into the model and selects the model with better out-of-sample generalization power.

The model selection procedure is applied to both fb-FPCA and mb-FPCA model to select the optimal R from a sequence of candidates ranging from 0 to 25. Due to the scale difference of fb-FPCA and mb-FPCA models, their $\mathrm{opCV}(R)$'s are not directly comparable. To overcome this problem, we take the ratio of $\mathrm{opCV}(R)$ and $\mathrm{opCV}(0)$ to define a quantity called *optimal CV score ratio* as $\mathscr{R}_{opCV} := \mathrm{opCV}(R)/\mathrm{opCV}(0)$. This adjust the scales of $\mathrm{opCV}(R)$ to the same level for both models.
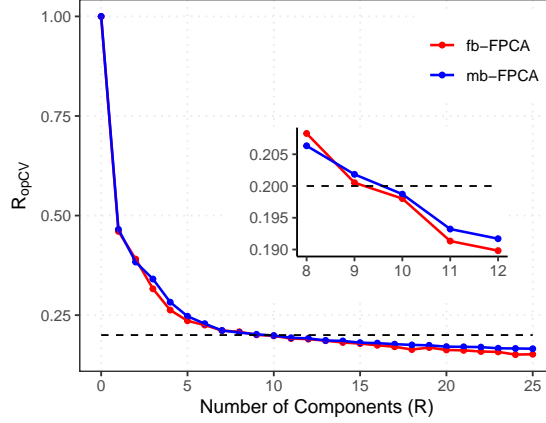
Figure 3.1: The optimal CV score ratio of fb-FPCA model and mb-FPCA model. The red and blue points correspond to fb-FPCA and mb-FPCA model, respectively. The black dashed horizontal line corresponds to $y = 0.20$. The inset panel zooms in on part of x-axis with R ranging from 8 to 12.

Figure 3.1 shows the change of $\mathscr{R}_{opCV}$ with the number of components R. The red and blue lines represent the results of fb-FPCA and mb-FPCA model, respectively. In general, the average prediction error decreases as more components are included in the model. The $\mathscr{R}_{opCV}$ is 46.03% for fb-FPCA model and 46.54% for mb-FPCA model when $R = 1$, which indicates that the inclusion of a single component dramatically reduces the prediction error by more than 50% for both models. However, the difference between the fb-FPCA and mb-FPCA is very small which suggests that the dominate source of the data diversities can be well captured by both models, such as can be expected if the effect is due mostly to interstellar dust extinction but with the large majority of the SNe showing very small amount of reddening. As the component number increases, the decreasing trend of the ratio becomes gradually flat and the ratio eventually stabilizes at around 15%. The black dashed lines in Figure 3.1 correspond to $y = 0.20$ which intersects with two $\mathscr{R}_{opCV}$ curves between $R = 9$ and $R = 10$. In this study, we will present three models: **one-component model** which explores only the dominant variation within the spectral data; **two-component model** to compare with SNEMO2 and SALT2 SED models; and a comprehensive **ten-component model** which captures more than 80% of the information of the spectral time series in the training dataset.

3.4.2 Model Comparison

This section presents the model training results of the fb-FPCA and mb-FPCA models, including the estimated mean, principal component surfaces and SN Ia spectral fitting results. We compare the derived principal component scores with parameters extracted from other SN Ia SED models at the end of this section.

Our fb-FPCA and mb-FPCA models are constructed in the wavelength range $[3500\text{Å}, 8000\text{Å}]$ and phase range $[-10, 20]$. Figure 3.2 and Figure 3.3 show the estimated mean surfaces and principal component surfaces for the two-component models of fb-FPCA and mb-FPCA. For illustrative purpose, we only show several slices in the phase and wavelength dimension. The three rows correspond to mean surface $\phi_0(\lambda, t)$, the first principal component surface $\phi_1(\lambda, t)$ and the second principal component surface $\phi_2(\lambda, t)$, respectively. From the left column to the right, the eigenspectrum slices at four demonstrative phases denoted with $\{\phi_r(\lambda, \cdot)\}_{r=0,1,2}$, the monochromatic light curves at four wavelengths denoted with $\{\phi_r(\cdot, t)\}_{r=0,1,2}$ and the broad-band light curves in the rest-frame Kron-Cousins B, V and R filters denoted with $\{\widetilde{\phi}_r(t)\}_{r=0,1,2}$ are shown. In the second panel of the left column of Figures 3.2 and 3.3, the re-scaled and vertically re-located CCM extinction law with $R_V = 3.1$ is overplotted using grey solid lines. By comparing the first eigenspectrum slices $\phi_1(\lambda, \cdot)$ at different phases with the extinction law, we find that the $\phi_1(\lambda, t)$ presents a homogeneous trend regardless of the phases of the data, especially in the mb-FPCA model. This behavior suggests that the surface of the first principal component effectively captures the effect of interstellar dust extinction. Note however, the same component may also capture partially the time-independent color variations intrinsic to different sub-types of SNe Ia. The second eigenspectra are distributed around a mean value of 0 and describe the spectral features and their time evolution. We notice that the estimated mean spectral surface and the eigenspectra time series of the mb-FPCA model are unstable at the two wavelength boundaries, in particular, the second eigenspectra depict a fake spectral feature at around 8000Å. This issue is most likely due to a combination of the artifacts introduced by taking logarithm of the flux close to 0 and the increased spectral diversities of SNe Ia at this wavelength range caused by the high velocity component of

40

the Ca II IR triplet [85, 86].
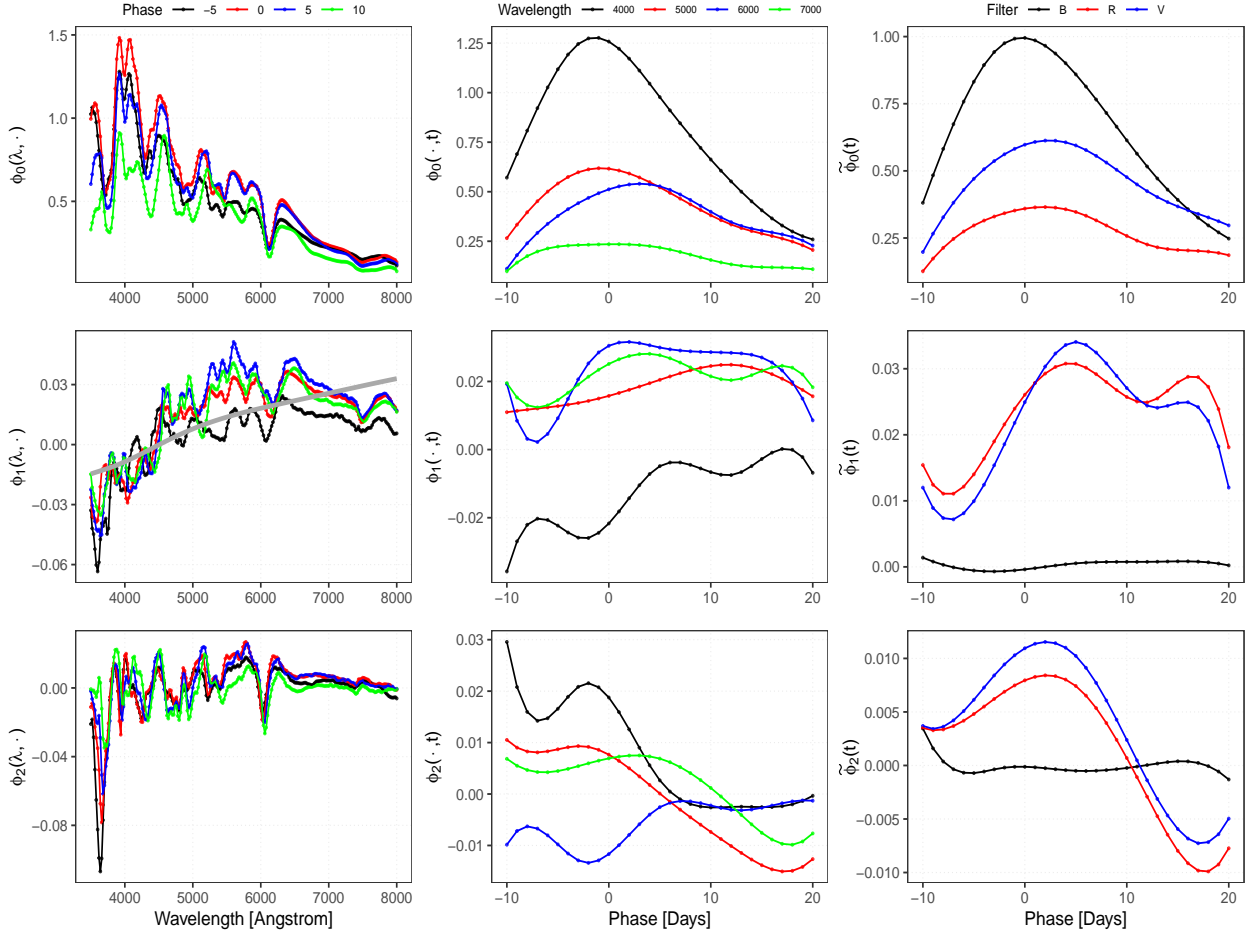
Figure 3.2: Demonstrative spectra, monochromatic and broad band light curves of the two-component fb-FPCA model. From top to bottom, the mean eigenspectra $\phi_0(\lambda, t)$, the first eigen-spectra $\phi_1(\lambda, t)$, and the second eigenspectra $\phi_2(\lambda, t)$ are shown. The left column presents eigen-spectrum slices at four demonstrative phases of day -5 (black), 0 (red), 5 (blue), and 10 (green) past B band maximum, the middle panel shows monochromatic light curves at four different wave-lengths: 4000Å (black), 5000Å (red), 6000Å (blue), and 7000Å (green), and the right column shows the broad-band light curves in the rest-frame Kron-Cousins B (black), V (red) and R (blue) filters. The grey solid line in the second panel of the left column represents the re-scaled and re-located CCM extinction law with $R_V = 3.1$.
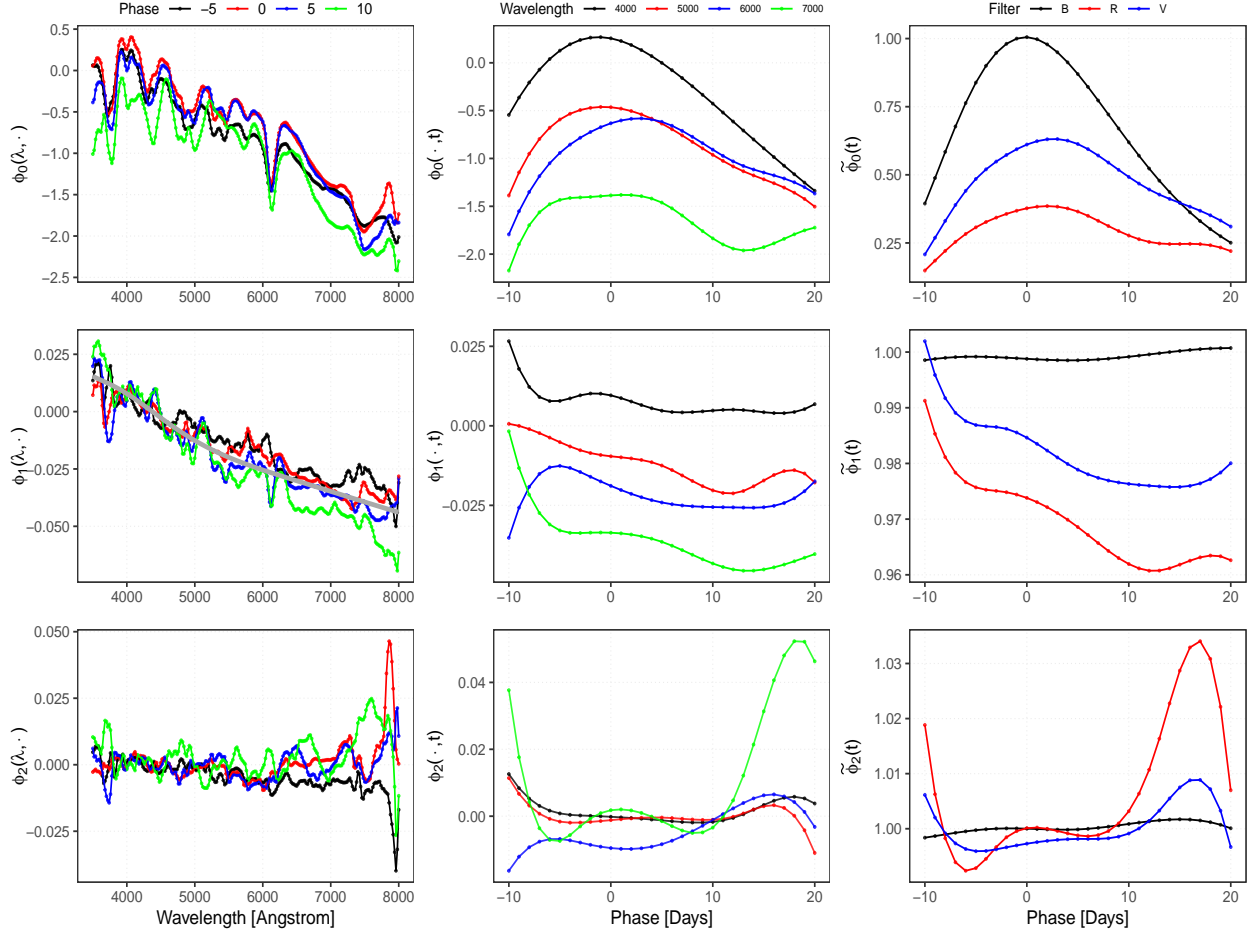
Figure 3.3: Same as Figure 3.2 but for the two-component mb-FPCA model.

Examples of the fb-FPCA model and the mb-FPCA model fitting results are shown in Figure 3.4 and 3.5 for three SNe: SN33, SN67, and SN106. Each column shows the fitting results of the spectra of one object observed at different phases with varying component numbers. In general, one component model (plus mean) is able to capture the overall spectral trend and the time evolution of the observed spectra. However, higher order principal component functions are required to precisely depict the details of localized spectral features. Of these three SNe, SN67 is SN 1991T-like with shallow Si II lines, the other two are normal SNe Ia but at different phases. The one and two component models perform well for data around optical maximum, and deteriorates at later phases, whereas the ten component model provide good fits to data of all epochs.
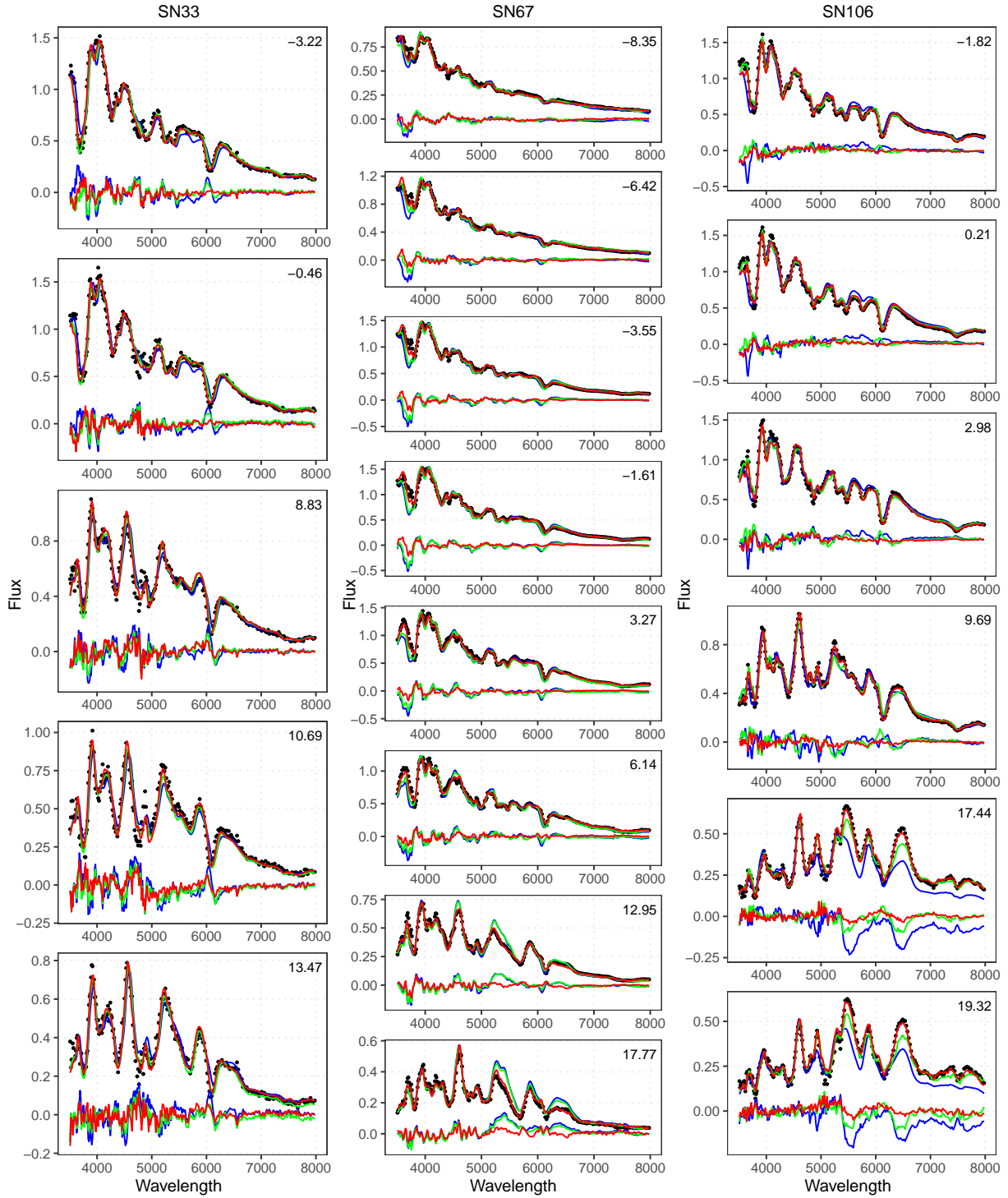
Figure 3.4: SED fitting examples of the fb-FPCA model with different component numbers. The blue, green and red solid lines correpond to $R = 1, 2$, and 10, respectively. From left to right, the columns are for one SN33, SN67, and SN106 with phases of the spectra marked at the upper right corner of each panel. The residual spectra are plotted at the bottom in each panel.
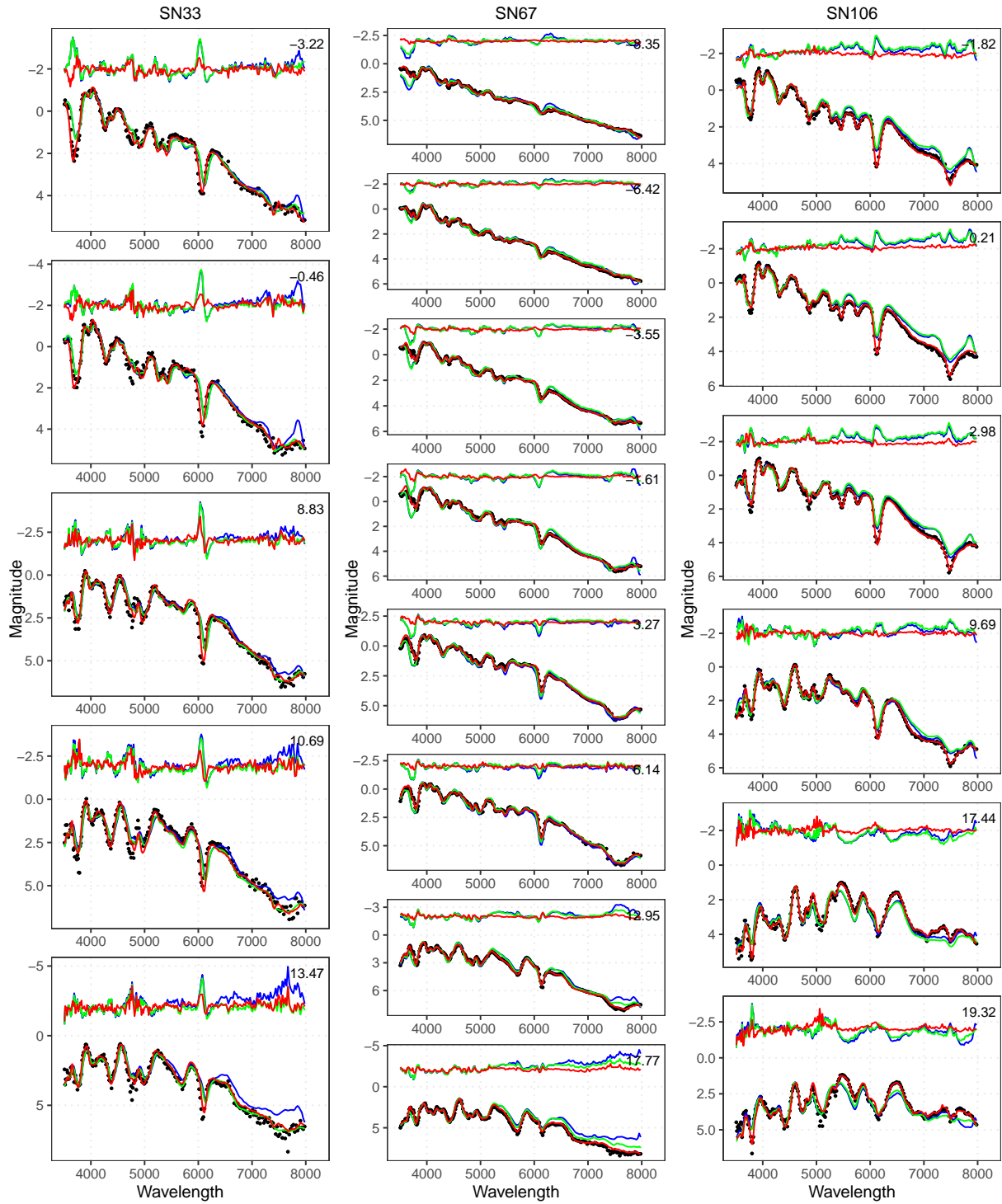
Figure 3.5: Same as in Figure 3.4 but for the mb-FPCA models. An offset of 2 is added to the residual spectra to separate them from the supernova spectra, and are shown above the supernova spectra.

Figure 3.6 shows the estimated scores of the two-component fb-FPCA and mb-FPCA model versus the parameters extracted from two other standard models SALT2 and SNEMO2. SALT2 and SNEMO2 are both two-component plus dust extinction models in which the supernova spectra are decomposed into a linear combination of two components with an extra color term multiplied to the linear term to account for interstellar extinction – in SNEMO2, the color difference is modeled by widely used extinction laws with a prefixed parameter R_V, whereas SALT2 constructs a color correction function using a third order polynomial with additional constraints, whose effect has been proven to be similar to the CCM extinction law [2] [1]. In accordance with SNEMO2, a constant value of $R_V = 3.1$ has been assumed in this comparison analysis. However, we note that the assumed R_V and extinction curve are needed only for SNEMO2, our algorithm does not need these assumptions to derive the extinction corrections. We adopt the same notations from the original SNEMO2 and SALT2 model such that (c_1, A_s) and (x_1, c) represent the coefficients of the first component and color parameter, of SNEMO2 and SALT2, respectively. The corresponding parameters for our models are the FPCA scores denoted as $\left(\beta_1^{fb}, \beta_2^{fb}\right)$ and $\left(\beta_1^{mb}, \beta_2^{mb}\right)$ for the fb-FPCA and mb-FPCA model, respectively.

The first row of Figure 3.6 shows the relation between the first score estimated in our models and the color parameters computed according to SNEMO2 and SALT2. The strong correlations between β_1^{fb} and β_1^{mb} with A_s and c confirm our previous observation that the first principal component ϕ_1 of our models are closely related to interstellar dust reddening. For example, the coefficient of determination is calculated to be 0.945 and 0.951 for the simple linear regression model of A_s versus β_1^{fb} and A_s versus β_1^{mb}, respectively. The dramatic decrease of the optimal CV score ratio $\mathscr{R}_{opCV}$ after introducing the first component presented in Figure 3.1 demonstrates that the interstellar dust reddening is the major contributor to the inhomogeneity of the observed supernova spectra [36, 49, 47]. One subtle but crucial observation is that an obvious nonlinear trend can be found in the first panel where the relation between the first score of the fb-FPCA model and the color parameter A_s from SNEMO2 is shown. Although the fb-FPCA and the mb-FPCA models

---

[1]See [50] and [63].

appear to be equally well correlated to the color indicators of SNEMO2 and SALT2, noticeable deviations at relatively large values of A_s and β_1^{fb} are seen in Figure 3.6; such deviations are less significant for the β_1^{mb} model. This is understandable as the effect of interstellar dust extinction can be approximated as a linear process in the flux scale only when the extinction effect is low.

The second row of Figure 3.6 shows the scatterplots of the second score β_2^{fb} and β_2^{mb} against the c_1 from SNEMO2 or x_2 from SALT2. It's not surprising that the levels of correlation are low, because the means and the first components of all models are affected by their own color corrections which are mixtures of the effects of interstellar dust extinction and intrinsic color diversities. The separation of these effects will facilitate the construction of a method of data-driven extinction correction and spectral quantification, as will be shown in the next section.
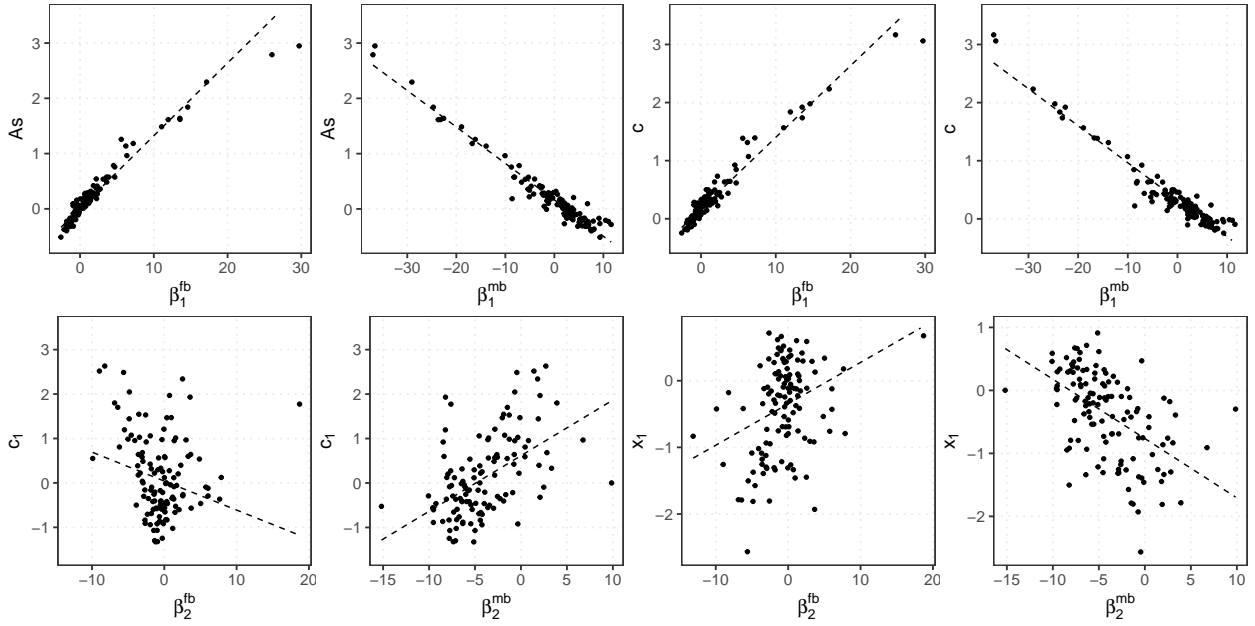
Figure 3.6: The scores of two-component fb-FPCA and mb-FPCA model versus parameters of SALT2 and SNEMO2. The color parameter and the coefficient of the first component of SNEMO2 are denoted with A_s and c_1, and the counterparts of SALT2 are denoted using c and x_1. The scores of the fb-FPCA model and the mb-FPCA model are represented using $\{\beta_r^{fb}\}_{r=1}^2$ and $\{\beta_r^{mb}\}_{r=1}^2$, respectively. From upper left to bottom right, the scatterplots correspond to A_s versus β_1^{fb}, A_s versus β_1^{mb}, c versus β_1^{fb}, c versus β_1^{mb}, c_1 versus β_2^{fb}, c_1 versus β_2^{mb}, x_1 versus β_2^{fb}, and x_1 versus β_2^{mb}. The data points are fitted by the robust linear regression (the black dashed lines).

3.5 The Dust Extinction and the Intrinsic Property
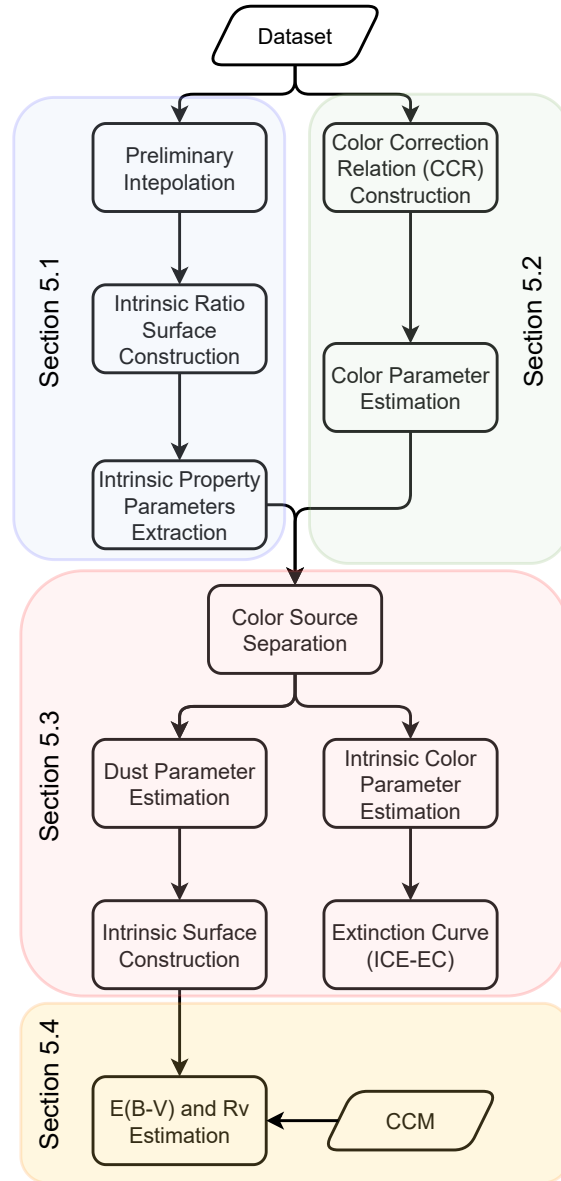
Figure 3.7: An illustrative flowchart of the **Independent Component Estimate of SNe Ia**.

In Section 3.5, we describe the procedure of Independent Component Estimate (ICE) that separates the effect due to dust extinction and intrinsic SN properties (Figure 3.7). In Section 3.5.1, we construct an interstellar dust extinction free surface called *flux ratio surface* to extract super-

nova intrinsic property parameters. In Section 3.5.2, we present two heuristic CCRs and calculate the supernova *color parameter* which quantifies the mixed effect of interstellar dust reddening and supernova intrinsic color information. By investigating the relation between the color parameter and intrinsic property parameters, we formulate two hypotheses on the supernova intrinsic color models and thereafter separate intrinsic color from the dust reddening effect in Section 3.5.3. Finally, we derive the dust extinction curve based on the SN data. With the intrinsic SED surfaces of SNe Ia we have constructed, we also carry out parameter studies of the extinction properties based on the CCM extinction law [2]. Hereafter, we use $y^{(s)}(\lambda, t)$ and $\epsilon^{(s)}(\lambda, t)$ to represent the irregular spectrophotometric training data $\{y_{i,j}^{(s)}\}$ and its corresponding uncertainties with $i \in \{1, \ldots, I_j^{(s)}\}$ and $j \in \{1, \ldots, J^{(s)}\}$ for SN Ia s to simplify the notations.

3.5.1 Flux Ratio Surface

The exploration of the SN Ia intrinsic features relies on taking away the dust reddening effect from the observed measurements and constructing a intrinsic property spectrophotometric time series. In this section, we propose a dust reddening free flux ratio surface for each SN Ia called *flux ratio surface* whose construction follows a two-step procedure: interpolate the original SED surface to estimate the spectrum at the phase of B-band peak magnitude $f_0(\lambda, t_{max})$ and take the ratio of the spectral time series to the estimated spectrum $f_0(\lambda, t_{max})$ to derive the surface of spectral ratio.

The behavior of interstellar dust extinction is known to be a smooth function of wavelength and time-independent or only very weakly time dependent (see however, [57], [87], and [88] for potential time dependent extinctions). Therefore, we construct the flux ratio surface by assuming that dust reddening is an time-independent multiplier towards the flux such that

$$f(\lambda, t) = A(\lambda; \alpha) f_0(\lambda, t),$$

where $A(\lambda; \alpha)$ is time-independent extinction law with α being supernova-specific extinction parameter and $f_0(\lambda, t)$ denotes the intrinsic SED surface of SNe Ia. Therefore, the spectrum at the

phase of B-band maximum t_{max} can be represented as $f(\lambda, t_{max}) = A(\lambda; \alpha) f_0(\lambda, t_{max})$. With these notations, we bring forward the definition of extinction-independent flux ratio surface as

$$fr(\lambda, t) := \frac{f(\lambda, t)}{f(\lambda, t_{max})} = \frac{f_0(\lambda, t)}{f_0(\lambda, t_{max})},$$

which is essentially the ratio between the spectra observed at any time t with the spectra at B-band maximum. Despite the fact that no B-band maximum spectra observation is available, we can perform a preliminary interpolation using ten-component comprehensive fb-FPCA model to derive the $f(\lambda, t_{max})$ estimation. Two randomly selected examples SNe Ia SN0 and SN24 are presented in Figure 3.8 to exhibit the goodness of the predicted spectra at the B-band maximum. Different panel corresponds to the spectra observed at different phases. The gray points are observations and black solid lines are fitted spectra. The red solid lines are the predicted spectra at the phase of B-band maximum. As the time approaches the phase of B-band maximum, the black solid lines tend to converge to the predicted B-band maximum spectra.

After eliminating the effect of interstellar dust by constructing the flux ratio surface, we can analyze SN Ia intrinsic properties based on the scores of the fb-FPCA model of ratio surfaces. We refer these scores as *intrinsic property parameters*, denoted as $\{\gamma_1, \gamma_2, \ldots, \gamma_R\}$. Because the strength of the spectral lines are physically related to the temperature of the SN photosphere, it is reasonable to assume that the information about the intrinsic color is encoded in these parameters. To determine the number of components (R) for the flux ratio surfaces, we use the optimal CV score ratio $\mathscr{R}_{opCV}$ as discussed in Section 3.4.1. The $\mathscr{R}_{opCV}$ is calculated to be 100%, 66.8%, 59.2%, $56.7\%, 54.1\%$, 52.1%, and 49.3% for $R = 0, \ldots, 6$ and finally stabilizes at around 40% when R approaches 25. For the current stage, we focus on the score of the dominant component, i.e., γ_1.

3.5.2 Color Correction Relation

The FPCA components derived in the previous section, although showing properties that are related to dust extinction, are also affected by the intrinsic colors of SNe Ia. In this section, we

49
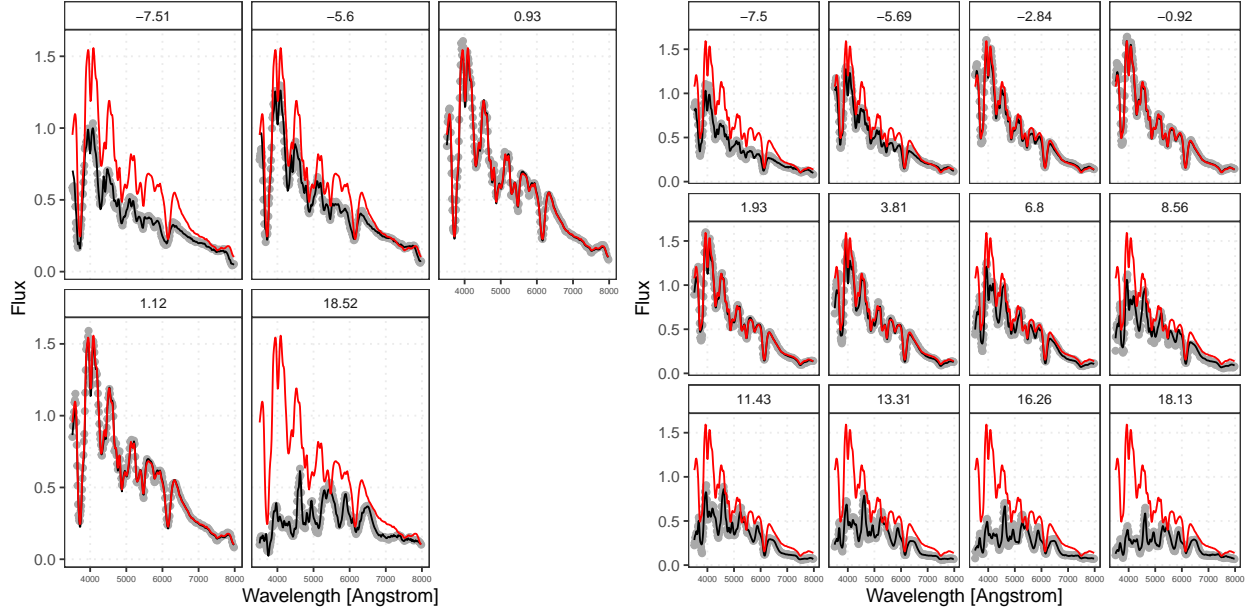
Figure 3.8: Examples of the fb-FPCA fitted spectra at B-band maximum of SN0 (left) and SN24 (right). The gray points are the observed spectral data at different phases. The phases of the data are shown at the top of each panel. The black lines are fitted spectra at different observed phases. The red lines are the predicted spectra at the phase of B-band maximum.

extract a smooth wavelength-dependent and time-independent component from the SN data which enables the separation of the dust reddening from the intrinsic color of SNe Ia. Two CCR's will be constructed, one is denoted by $\psi(\lambda)$ and is based on the surface of the one-component mb-FPCA model, and the other is denoted by $\psi_{ica}(\lambda)$ and is based on the results of independent component analysis (ICA).

3.5.2.1 *Independent Component Analysis*

As shown in Figure 3.3, besides a dust extinction law like trend, the extracted first eigenspectra also contains a considerable amount of spectral variations, suggesting that the component is contaminated by the intrinsic spectral features of the SNe and therefore also likely by their intrinsic colors. In this regard, we employ a signal processing tool called independent component analysis [70, 71, ICA] to explore a direct separation of the extinction from the localized spectral information. ICA is a multivariate statistical method to decompose a signal into several independent

components by maximizing the statistical independence of the estimated components. In contrast, the orthogonality between the eigenfunctions required in FPCA ususally does not guarantee their statistical independence and, in turn, the minimization of the signal mixing.

With the assumption that the effects of dust extinction and the SNe intrinsic properties originate from distinct independent nongaussian physical processes, the two processes can be statistically separated by maximizing their non-Gaussianity using FastICA [89, 90, 91]. To facilitate the calculation of ICA, we impute the irregular spectrophotometric data using ten-component mb-FPCA model and evaluate on an uniform grid in wavelength and phase. The data is centered using the mean surface $\phi_0(\lambda, t)$ defined in the mb-FPCA model (3.3) before performing ICA. We assume that there exists two latent independent components, one accounts for the dust extinction effect and the other depicts SN Ia intrinsic features. For consistency, they are denoted with $\phi_{ica,1}(\lambda, t)$ and $\phi_{ica,2}(\lambda)$. The decomposition of the SED surface using ICA can be expressed as,

$$\log\left(f^{(s)}(\lambda, t)\right) = \phi_0(\lambda, t) + \beta^{(s)}_{ica,1}\phi_{ica,1}(\lambda, t) + \beta^{(s)}_{ica,2}\phi_{ica,2}(\lambda, t),$$

where $\phi_{ica,1}(\lambda, t)$ and $\phi_{ica,2}(\lambda, t)$ are statistically independent. Figure 3.9 illustrates the behaviour of $\phi_{ica,1}(\lambda, t)$ and $\phi_{ica,2}(\lambda, t)$. The first ICA components are more featureless compared with eigenspectra slices of $\phi_1(\lambda, t)$ of the two-component mb-FPCA model shown in Figure 3.3.

3.5.2.2   The Color Correlation Relation

In this section, we present a heuristic two-step procedure to construct the time-independent wavelength-smoothed CCRs $\psi(\lambda)$ and $\psi_{ica}(\lambda)$ for the FPCA and ICA decompositions, respectively. They are used to depict the behavior of intrinsic color effect of the SNe and the properties of the dust extinction on the observed spectral shapes and time evolution. We take the construction of $\psi(\lambda)$ as an illustrative example and $\psi_{ica}(\lambda)$ can be established following the same procedure.

The construction of $\psi(\lambda)$ is based on the principal component $\phi_1(\lambda, t)$ of mb-FPCA model by averaging over its the time dependence and smoothing the localized spectral features. The detailed procedure is summarized below,
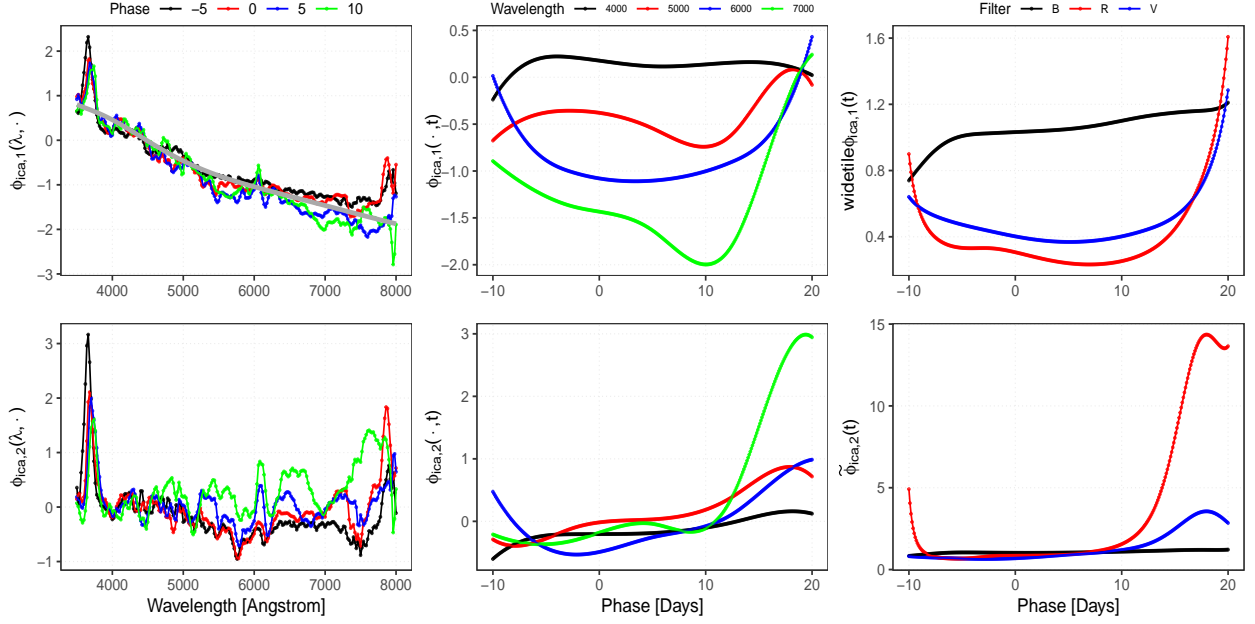
Figure 3.9: Same as Figure 3.3 but for ICA model.

Step 1. Obtain a time-free function $\langle \phi_1(\lambda, t) \rangle_t$ by taking the average of $\phi_1(\lambda, t)$ w.r.t the phase argument.

Step 2. Approximate $\langle \phi_1(\lambda, t) \rangle_t$ with a much smoother function $\psi(\lambda)$ which is expressed by cubic B-splines with interior knots located at the effective wavelengths of the B, V and R filters.

Note that $\langle \cdot \rangle_t$ denotes the operation of taking average over the time domain. The left column of Figure 3.10 shows the elements used or produced during the CCR's construction. The principal component $\phi_1(\lambda, t)$ and the independent component $\phi_{ica,1}(\lambda, t)$ evaluated at different phases ranging from -8 to 16 days with respect to the B-band maximum are shown by the blue lines of different heaviness. The black dashed lines represent $\langle \phi_1(\lambda, t) \rangle_t$ and $\langle \phi_{ica,1}(\lambda, t) \rangle_t$. The red solid curves represent the CCR $\psi(\lambda)$ and $\psi_{ica}(\lambda)$ after the B-spline smoothing. The dotted black lines at the bottom illustrate the B-splines we use to construct $\psi(\lambda)$ and $\psi_{ica}(\lambda)$.

Once we have determined the CCR $\psi(\lambda)$, the supernova-specific *color parameter* α is defined by projecting the spectrophotometric data $y^{(s)}(\lambda, t)$ onto the space spanned by $\psi(\lambda)$ after removing

the effect of the averaged SED surface $\phi_0(\lambda, t)$. The color parameter α for a specific SN Ia s (denoted by $\alpha^{(s)}$) can be estimated by the method of least squares, i.e.,

$$\alpha^{(s)} = \underset{\alpha^{(s)}}{\arg\min}\{\log\left(y^{(s)}(\lambda, t)\right) - \left[\phi_0(\lambda, t) - \alpha^{(s)}\psi(\lambda)\right]\}^2.$$

This definition ensures that all information associated with the CCR is extracted from the data and this information is quantified by the color parameter α. The corresponding color parameter of ψ_{ica} is denoted by α_{ica}. The color parameters α and α_{ica} can be compared with similar quantities derived by other methods including c from SALT2 and A_s from SNEMO2. These parameters are found to be strongly correlated, as shown in the middle and right panels of Figure 3.10.

3.5.3 Dust Reddening and Intrinsic Spectral Features Separation

The CCRs we have derived accounts for the effects of both dust reddening and intrinsic SN color. By investigating the correlation between the supernova color parameters (α and α_{ica}) derived in Section 3.5.2 and the intrinsic property parameter γ_1 as derived from the fb-FPCA of the flux ratios, we may decompose α or α_{ica} into two parts, each part of which accounts for a different source of the color variations, including intrinsic color diversities and interstellar dust reddening (Section 3.5.3.1). After removing the supernova intrinsic color diversities, we may derive an average *dust extinction curve (ICE-EC)* using a sample of SNe showing significant amount of dust reddening (Section 3.5.3.2). This also allows us to derive an extinction corrected flux manifold which is a representation of the intrinsic properties of of SNe Ia (Section 3.5.3.3).

3.5.3.1 *Intrinsic Color Hypotheses*

We formulate two hypotheses about the composition of the SNe. They consider two extreme cases where the observed colors are attributed to intrinsic color diversities and interstellar dust property variations described as below,

Hypothesis 1: All SNe Ia share an identical intrinsic color independent of their the spectral features, the difference in the observed colors is caused by the difference in the amount of
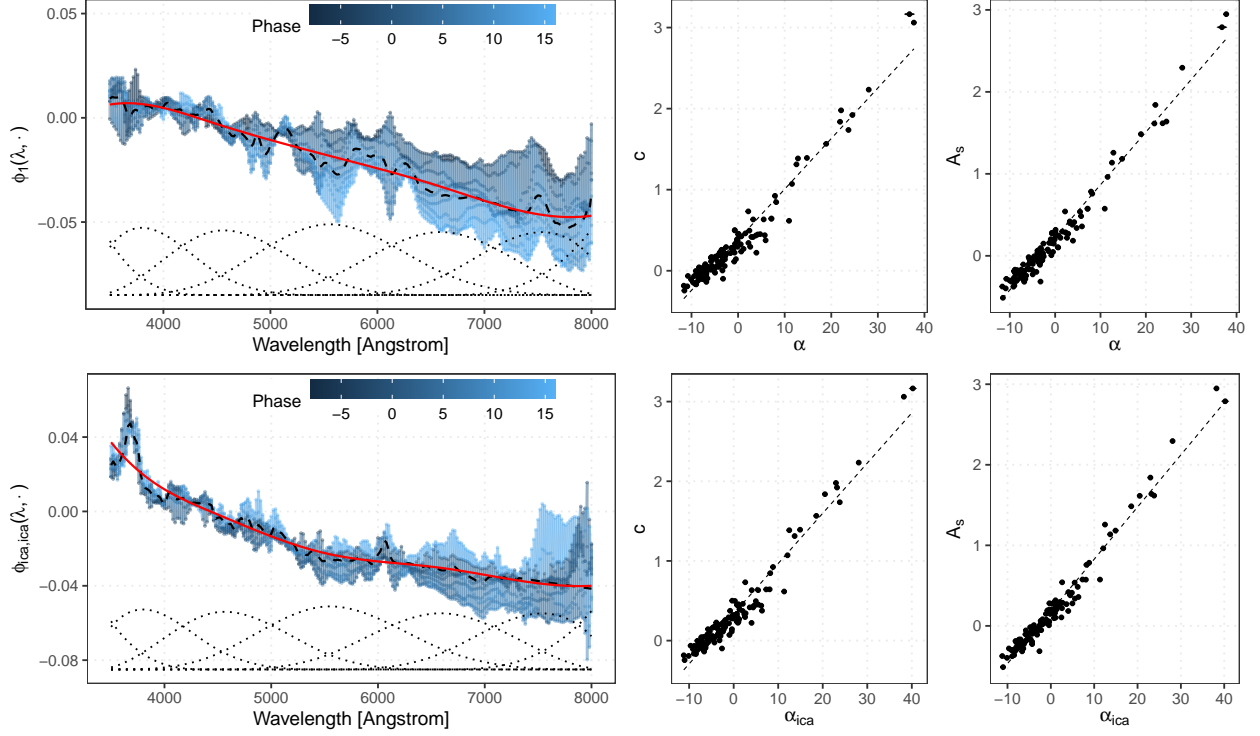
Figure 3.10: The construction of $\psi(\lambda)$ and $\psi_{ica}(\lambda)$ and the comparisons of α and α_{ica} with the color parameters from SALT2 and SNEMO2. Left: The blueish transparent lines represent $\phi_1(\lambda, t)$ (top) and $\phi_{ica,1}(\lambda, t)$ (bottom) at phases from day -8 to 16 w.r.t. B-band maximum. The dashed black lines represent $\langle\phi_1(\lambda, t)\rangle_t$ (top) and $\langle\phi_{ica,1}(\lambda, t)\rangle_t$ (bottom). The $\psi(\lambda)$ and $\psi_{ica}(\lambda)$ are presented using red solid lines. The dotted black lines in both panels are B-splines used to approximate the target functions. Middle and Right: The black points show the comparisons between the c of SALT2 and the A_s of SNEMO2 with the corresponding color parameters derived from the mb-FPCA (top) and ICA (bottom) decomposition. The black dashed lines are regression lines.

interstellar dust along the line of sights towards the SNe.

Hypothesis 2: The interstellar dust properties of all the SNe Ia share a common physical origin, but their intrinsic colors are dependent on their intrinsic spectral properties.

The SNEMO2 formulation of the problem virtually assumes Hypothesis 1. SALT3 explicitly requires its color index c to be independent of the the parameter x_1 which describes the intrinsic properties of SNe Ia. SALT3 is thus constructed with the goal of being consistent with Hypothesis 2. However, as shown in Figure 3.11, the c of SALT3 is found to be correlated with γ_1. The SNEMO A_s values are tightly correlated with those of SALT2 [63], and show similar correlations

with γ_1 (Figure 3.11). In fact, all the eigenvectors and their scores of SALT2/SALT3 are affected by the amount of dust extinction in the data sample and the requirement of c and x_1 being uncorrelated is insufficient to guarantee consistency with Hypothesis 2.

The intrinsic color as an important feature of SNe Ia may be natually encoded in the parameters describing their intrinsic properties. The FPCA decomposition using flux ratios in Section 3.5.1 can be employed to establish the correlation between intrinsic color and the fb-FPCA scores. The scatterplot of the color parameters α versus the intrinsic property parameter is shown in Figure 3.11, together with two Hypothetical models corresponding to the two hypotheses above. The blue solid horizontal line represents the naive lower envelope corresponding to the line $y = \min\{\alpha^{(s)}\}$. The red solid line depicts the lower boundary of the scatterplot which is estimated by 10% quantile regression. The slop and intercept of the red line are calculated by $(b, a) = \arg\min_{b,a} \sum_s \left(\alpha^{(s)} - f(\gamma_1^{(s)}) \right) \cdot \left(0.1 - 1_{\{x<0\}} (\alpha^{(s)} - f(\gamma_1^{(s)}) < 0) \right)$ with $f(\gamma_1) = b\gamma_1 + a$, where the Heaviside indicator function $1_{\{x<0\}}(u)$ is equal to 1 when $u < 0$ and 0 otherwise. This line is used to define the γ_1 dependent intrinsic color. The vertical distance between the color parameter α and the intrinsic color α_{intr} is then assumed to be due to interstellar dust reddening, denoted as $\alpha_{dust} = \alpha - \alpha_{intr}$. An illustrative example of intrinsic color and interstellar dust property under Hypothesis 2 are presented using the red dot and black segment in Figure 3.11. Similarly, we construct the intrinsic color model of ICA under the same hypotheses and it is presented in Supplemental Materials 3.6.2. The corresponding intrinsic color and dust extinction parameters are given as $\alpha_{ica,intr}$ and $\alpha_{ica,dust}$, respectively.

The proposed hypotheses consider two extreme cases of the source of SN Ia color. The naive horizontal line is defined as the intrinsic color model under Hypothesis 1 such that α_{intr} remain identical for all SNe Ia regardless of their various spectral feature properties. The α is attributed entirely to the interstellar dust property diversities, i.e., the statistical distribution of α_{dust} varies for different sub-Types of SNe Ia. The red edge curve serves as the intrinsic color model under Hypothesis 2 where the intrinsic color is assumed to be a function of SN intrinsic spectral properties, whereas the statistical distribution of α_{dust} under Hypothesis 2 is independent of the intrinsic
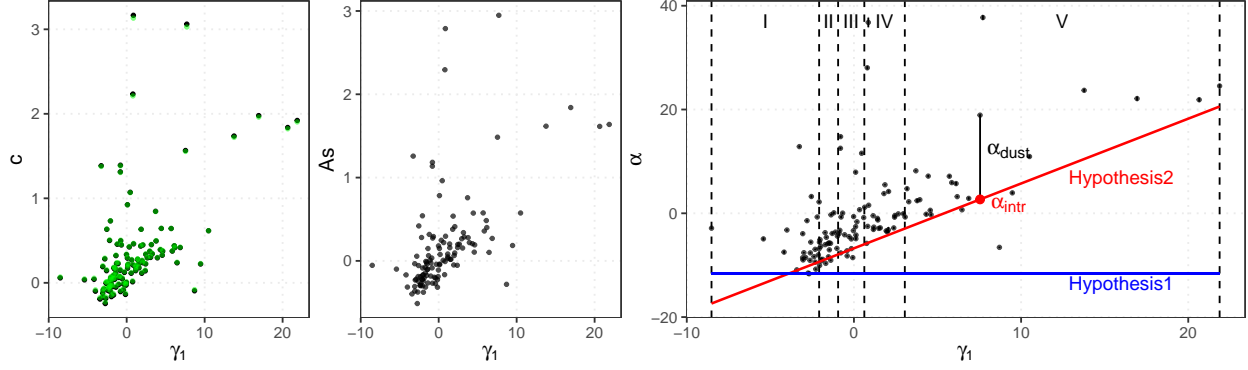
Figure 3.11: The color parameters versus intrinsic property parameter γ_1. Left: SALT2 (black) and SALT3 (green) color parameters c versus γ_1. Note that the differences are small and the majority of the SALT2 and SALT3 points overlap in this figure. Middle: SNEMO2 color parameter A_s versus γ_1. Right: the mb-FPCA color index α versus γ_1. The blue solid horizontal line and the red solid line correspond to the intrinsic color α_{intr} as a function of intrinsic property parameter γ_1 under Hypothesis 1 and 2, respectively. The definitions of α_{intr} and α_{dust} under Hypothesis 2 are shown. The vertical distance between the color parameter α and α_{intr} (the black solid line) defines the parameter of dust extinction, α_{dust}. The SNe Ia are split into five subgroups according to the quantile levels $0 \sim 20\%, 20 \sim 40\%, 40 \sim 60\%, 60 \sim 80\%, 80 \sim 100\%$ of γ_1, shown by the vertical dashed lines, to test the consistency of their statistical distribution.

properties of SNe Ia.

We examine statistical distribution of α_{dust} of SNe Ia with different intrinsic properties under the two hypotheses using Kolmogorov-Smirnov test. To overcome the problem caused by limited sample size, we split the SNe Ia into five subgroups according to the quantile levels $0 \sim 20\%, 20 \sim 40\%, 40 \sim 60\%, 60 \sim 80\%, 80 \sim 100\%$ of the intrinsic property parameter γ_1 (Figure 3.11) and treat the SNe Ia in each subgroup as repeated observations of one object since they share similar intrinsic properties. The distribution of α_{dust} for five subgroups under two hypotheses are shown in Figure 3.12. The corresponding figure of $\alpha_{ica,dust}$ can be found in Supplemental Materials 3.6.2.

The p-values of the Kolmogorov-Smirnov test of α_{dust} are reported in Table 3.1 for each pair of subgroups under two hypotheses. The upper triangular part (shaded cells) and lower triangular part of the table report the p-values for the pairwise comparisons between the five subgroups under Hypothesis 1 and 2, respectively. Under Hypothesis 1, the distribution of each pair of α_{dust} being identical is rejected at a significance level of 0.05 in most cases, suggesting different dust property

56
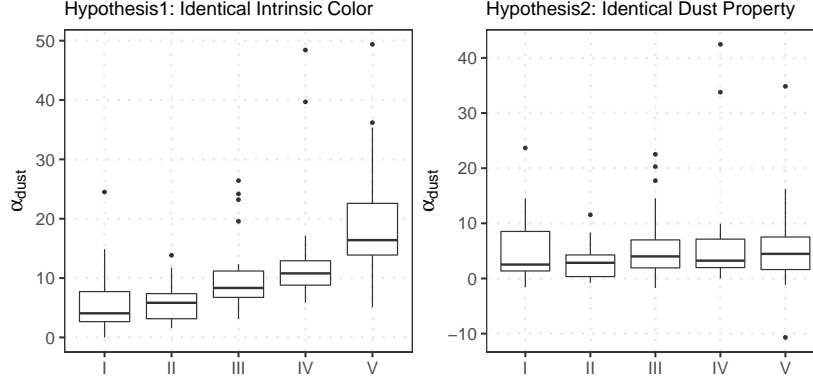
Figure 3.12: The distribution of α_{dust} for different subgroups under Hypothesis 1 (left) and Hypothesis 2 (right). The lower and upper bounds of each box show the 25% and 75% of the distribution, respectively. The thick horizontal lines show the median values of each box. The outliers are given by black dots.

for different SNe Ia, whereas hypothesis 2 is in agreement with the statistical test. The corresponding table of the Kolmogorov-Smirnov test for the ICA is reported in Supplemental Materials 3.6.2.

Table 3.1: The p-value of Kolmogorov–Smirnov test for two hypotheses. The shaded cells contain the p-values under Hypothesis 1 and the lower triangular part reports the results under Hypothesis 2. The values in bold font are less than the significance level 0.05.

	I	II	III	IV	V
I	1.000	0.156	**0.004**	**0.000**	**0.000**
II	0.379	1.000	**0.012**	**0.000**	**0.000**
III	0.475	0.263	1.000	**0.030**	**0.000**
IV	0.475	0.449	0.902	1.000	**0.001**
V	0.710	0.127	0.996	0.463	1.000

3.5.3.2 *Dust Extinction Curve from Intrinsic-Color-Removed Surface*

As discussed above, the wavelength dependence of the CCRs ($\psi(\lambda)$ and $\psi_{ica}(\lambda)$) are caused by a mixture of the effects of dust extinction and intrinsic color. Under Hypothesis 2, we can remove the intrinsic color dependence quantified by $\alpha_{intr}^{(s)}\psi(\lambda)$ and $\alpha_{ica,intr}^{(s)}\psi_{ica}(\lambda)$ from the data to enhance the effect of the dust. This leads to the intrinsic-color-corrected spectra, i.e., $\log\left(y_{dust}^{(s)}(\lambda,t)\right) =$

$\log\left(y^{(s)}(\lambda, t)\right) + \alpha_{intr}^{(s)}\psi(\lambda)$ (or $\alpha_{ica,intr}^{(s)}\psi_{ica}(\lambda)$). The FPCA and ICA can be retrained using the intrinsic-color-corrected data to deliver a better approximation of the wavelength dependence of dust extinction. This training process leads to revised CCRs denoted by $\psi_{dust}(\lambda)$ and $\psi_{ica,dust}(\lambda)$ which capture the wavelength dependence of the *dust extinction curves (ICE-CE)*.

Figure 3.13 compares $\psi(\lambda)$, $\psi_{ica}(\lambda)$, $\psi_{dust}(\lambda)$, and $\psi_{ica,dust}(\lambda)$ with the CCM extinction law [2] with $R_V = 1.5$, 5 and 7, with $\psi(\lambda)$ (or $\psi_{ica}(\lambda)$, $\psi_{dust}(\lambda)$, $\psi_{ica,dust}(\lambda)$) re-scaled and vertically shifted to match the scale of the CCM extinction law with various R_V values for a fixed A_V of 1. As shown in Figure 3.13, the wavelength dependence of the functions $\psi_{ica}(\lambda)$ and $\psi_{ica,dust}(\lambda)$ match very well with for $R_V = 1.5$, but the extinction curves deduced from $\psi(\lambda)$ and $\psi_{dust}(\lambda)$ require R_V that is significantly larger than 1.5. In fact, $\psi(\lambda)$ and $\psi_{dust}(\lambda)$ show good consistency with the CCM extinction laws with $R_V = 7$ and $R_V = 5$, respectively.
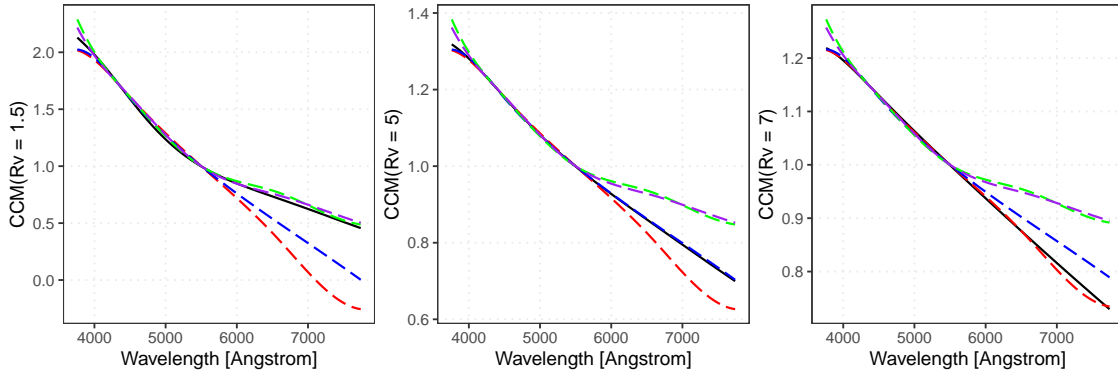
Figure 3.13: The CCRs $\psi(\lambda)$ and $\psi_{ica}(\lambda)$ and their corresponding ICE-ECs $\psi_{dust}(\lambda)$ and $\psi_{ica,dust}(\lambda)$. The black solid lines show the CCM extinction laws [2] with fixed $A_V = 1$ and various $R_V = 1.5$ (left), 5 (middle), and 7 (right). The red, blue, green, and purple dashed lines correspond $\psi(\lambda)$, $\psi_{dust}(\lambda)$, $\psi_{ica}(\lambda)$ and $\psi_{ica,dust}(\lambda)$, respectively. All the CCRs are scaled and shifted vertically to match the CCM curves at 5500Å and 4350Å.

3.5.3.3 *Intrinsic SED Surface after Dust Extinction Correction*

A representation of SNe Ia intrinsic SED surface can be constructed to facilitate supernova studies and their cosmological application. The procedures above can effectively separate the

effect due to dust extinction and intrinsic properties and thereby allow us to construct the 2D extinction free manifold in a straightforward way. This is done by simply applying the CCRs but with only the portion due to dust extinction to the original data sample to derive a dust-free dataset, i.e. $\log\left(y_{intr}^{(s)}(\lambda, t)\right) = \log\left(y^{(s)}(\lambda, t)\right) + \alpha_{dust}^{(s)}\psi(\lambda)($ or $\alpha_{ica,dust}^{(s)}\psi_{ica}(\lambda))$, and retrain the mb-FPCA model. The formula of the mb-FPCA model for the extinction-free dataset is expressed as

$$\log\left(y_{intr}^{(s)}(\lambda, t)\right) = \phi_{0,intr}(\lambda, t) + \sum_{r=1}^{R} \beta_{intr,r}^{(s)}\phi_{intr,r}(\lambda, t) + \epsilon_{intr}^{(s)}(\lambda, t), \qquad (3.8)$$

where $\{\phi_{r,intr}(\lambda, t)\}_{r=0}^{R}$ represent the universal averaged and principal component surfaces of the extinction-free SED data, $\{\beta_{r,intr}^{(s)}\}_{r=1}^{R}$ are SN specific scores and $\epsilon_{intr}^{(s)}(\lambda, t)$ is the noise term. See the discussions following equation (3.3) and (3.4) for the meaning and properties of the notations in equation (3.8). In principle, the data set thus corrected by the CCRs may still be contaminated by the effect of dust extinction because the mean component and the eigenvectors from the PCA and ICAs are affected by interstellar dust. An iterative procedure may be devised to alleviate the remaining effect of dust extinction. However, owning to the small size of the data sample we do not pursue this further. Nonetheless, as shown in Figure 3.14, the scores of the first PCA component of the intrinsic surfaces ($\beta_{1,intr}$) do show strong correlation with γ_1 and weak correlations with α_{dust}, which indicates that the intrinsic SED surface we have constructed to be reasonable approximations already.

Note that since we apply two different ICE-ECs to correct for the effect due to dust extinction, we finally derive two different intrinsic SED surface training sets and accordingly two different averaged and principal component spectrophotometric time series. The mean surfaces and eigensurfaces of the dust-free spectrophotometric data for data set corrected by the CCRs of the FPCA and ICA are available for download from the online document.

3.5.4 Final Dust Extinction Relation

With the intrinsic SED and its representitave templates, we can derive the dust extinction of each SN Ia by matching any observed spectra with the intrinsic SED manifold. As the wavelength
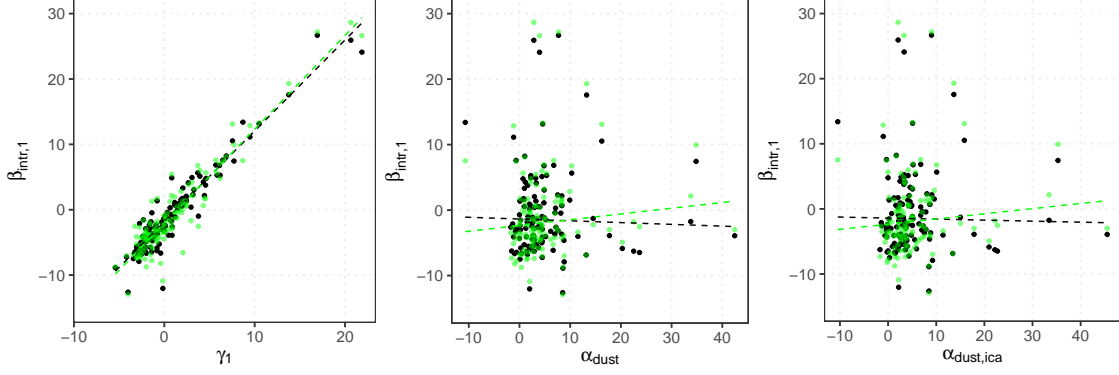
Figure 3.14: The first mb-FPCA score of the intrinsic SED surfaces $\beta_{intr,1}$ versus γ_1, α_{dust} and $\alpha_{ica,dust}$. The black and green points represent $\beta_{intr,1}$ associated with $\psi(\lambda)$ corrected and $\psi_{ica}(\lambda)$ corrected intrinsic SED surfaces, respectively. The black and green dashed lines are regression lines. The score of the first PCA component is found to be tightly correlated with the intrinsic spectral property measure γ_1, and is uncorrelated with the dust extinction indices deduced for each SN.

dependence of the extinction curves we have derived agrees with the CCM extinction law, we will adopt the CCM extinction law [2] with its R_V and A_V parameterization for this purpose. The formula is given as

$$
\begin{aligned}
(R_V^{(s)}, A_V^{(s)}) = \underset{R_V^{(s)}, A_V^{(s)}, c^{(s)}}{\arg\min} \Big\{ \log\left(y^{(s)}(\lambda, t)\right) - 0.4 \log_{10} e \cdot A_V^{(s)} \text{CCM}(\lambda, R_V^{(s)}) - \\
\phi_{0,intr}(\lambda, t) - \Sigma_r \beta_{r,intr}^{(s)} \phi_{r,intr}(\lambda, t) + c^{(s)} \Big\}^2,
\end{aligned}
\tag{3.9}
$$

where the factor $\log_{10} e$ accounts for the difference between $\log$ and $\log_{10}$, an additional offset $c^{(s)}$ is allowed for each SN. The intrinsic surfaces $\{\phi_{r,intr}\}_{r=0}^R$ are taken to be $\psi(\lambda)$ corrected ones or $\psi_{ica}(\lambda)$ corrected ones and thereby we have two groups of R_V and A_V estimation. When solving the optimization problem (3.9), we restrict the parameter R_V lying within the range $(0, 8)$ and reject the samples that deliver an estimation on or beyond boundaries. The distributions of A_V and R_V are shown in Figure 3.15. The two intrinsic surfaces lead to two different estimates of R_V and A_V, this difference is largely caused by the systematic errors involved in the ICE-ECs, especially at wavelengths redder than ~ 7000Å. The two ICE-ECs can be discriminated by examining the values of some very well observed SNe Ia with significant extinction and with

extensive wavelength and time coverage that allow for the A_V and R_V values to be determined directly from observational data. For example, the R_V value is equal to 1.4±0.1 for SN 2014J with $A_V = 1.85 \pm 0.11$ based on optical to NIR observations [92]. More examples of low R_V values can be found in e.g., [93, 94, 95, 96]. It is extremely encouraging that the R_V values derived from the ICA decomposition is consistent with these observations. If the ICA decomposition does provide a reasonable way of measuring the extinction properties of SNe Ia, we may also infer from Figure 3.15 that there is a general trend that the R_V values tend to be larger for SNe with larger A_V. Such a trend, if borne out with more data, is certainly important for cosmological applications of SNe Ia.
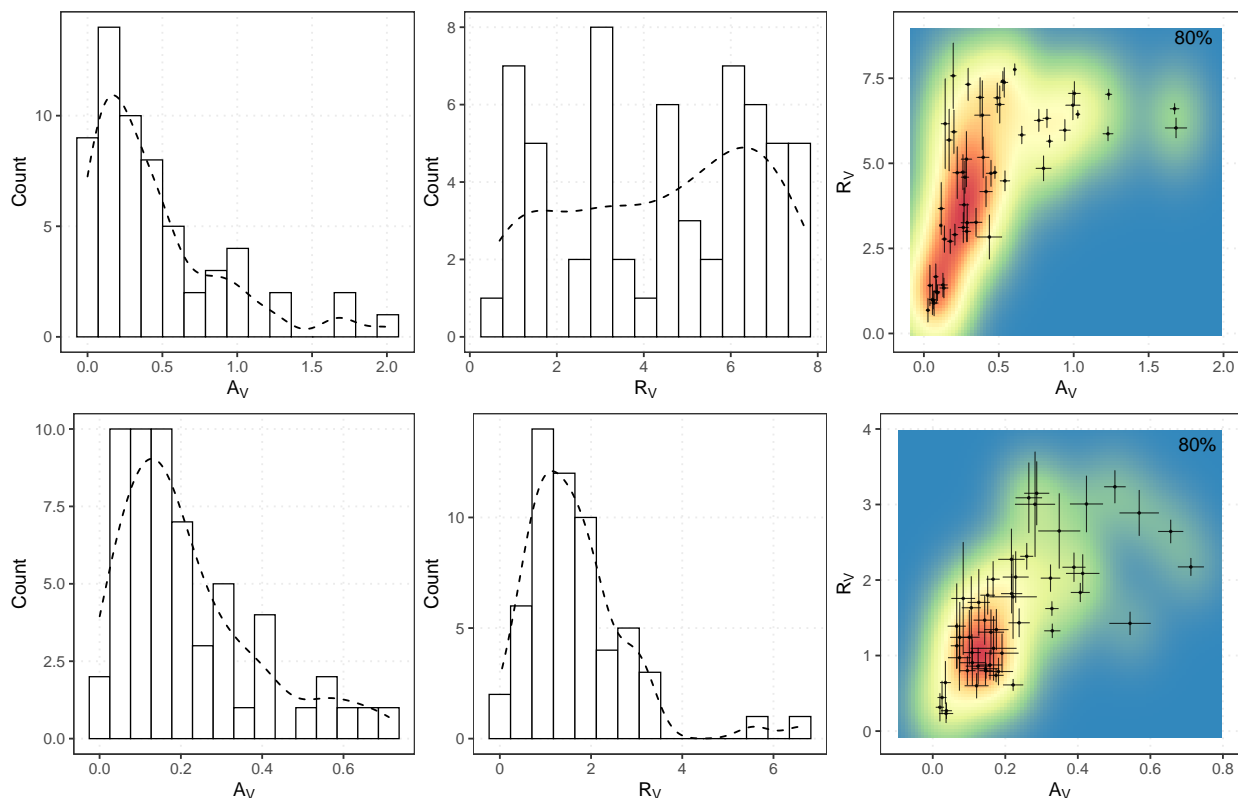
Figure 3.15: The distribution of A_V and R_V estimated with equation (3.9) using the $\psi(\lambda)$ corrected intrinsic SED surface (top) and $\psi_{ica}(\lambda)$ corrected intrinsic SED surface (bottom). From left to right, the columns are histograms of A_V, R_V, and the density plots of R_V versus A_V, only the SNe with dust parameters $\alpha_{dust}^{(s)}$ (or $\alpha_{ica,dust}^{(s)}$) above the 20% quantile level of their respective distributions are drawn. The scaled density curves are overplotted on histograms using dashed curves.

3.6   Supplemental Materials

3.6.1   FPCA Model Training Algorithm

This section presents the details of the model training algorithm in Section 3.3.2 for the flux-based functional principal component analysis (fb-FPCA). Similar implementation applies to the mb-FPCA model.

3.6.1.1   Learning the Mean Surface

For the spectrophotometric observation of the s-th supernova at time (phase) $t_j^{(s)}$, let $\mathbf{W}_{j,I_j^{(s)}\times p}^{(s)} = \left(\mathbf{W}(\lambda_1^{(s)}), \mathbf{W}(\lambda_2^{(s)}), \cdots, \mathbf{W}(\lambda_{I_j^{(s)}}^{(s)})\right)^T$ be the $I_j^{(s)} \times p$ wavelet basis matrix evaluated at the sequence of observation wavelength $(\lambda_1^{(s)}, \dots, \lambda_{I_j^{(s)}}^{(s)})$ and $\mathbf{B}_{j,q\times 1}^{(s)} = \mathbf{B}(t_j^{(s)})$ be the spline basis vector. The fb-FPCA model (3.5) for this observation in matrix format is

$$\mathbf{y}_j^{(s)} = \mathbf{W}_j^{(s)}\Theta_0\mathbf{B}_j^{(s)} + \sum_{r=1}^{R} \beta_r^{(s)}\mathbf{W}_j^{(s)}\Theta_r\mathbf{B}_j^{(s)} + \mathbf{V}_j^{(s)}\boldsymbol{\epsilon}_j^{(s)}, \tag{3.10}$$

where $\mathbf{y}_j^{(s)}$ is a spectra observation vector with length $I_j^{(s)}$, $\boldsymbol{\epsilon}_j^{(s)}$ is a mean-zero random vector of length $I_j^{(s)}$ with identity covariance matrix, and $\mathbf{V}_j^{(s)} = \text{diag}\{\sigma_1^{(s)}, \cdots, \sigma_{I_j^{(s)}}^{(s)}\}$. According to the construction of the functional principal component model (3.5), the last two terms in Equation (3.10) have zero expectations. The (weighted) least squares criterion guarantees a consistent estimation of the coefficient matrix Θ_0 of the mean surface and is presented below,

$$\min_{\Theta_0} \underbrace{\sum_{s=1}^{S}\sum_{j=1}^{J^{(s)}} \frac{1}{I_j^{(s)}}\|(\mathbf{V}_j^{(s)})^{-1}\left(\mathbf{y}_j^{(s)} - \mathbf{W}_j^{(s)}\Theta_0\mathbf{B}_j^{(s)}\right)\|_2^2}_{\text{I}} + \underbrace{\sum_{m=1}^{p}\eta\|\Theta_0^m\|}_{\text{II}} + \underbrace{\rho\,\text{tr}\left(\Theta_0\Omega\Theta_0^\top\right)}_{\text{III}}, \tag{3.11}$$

where Θ_0^m represents the m-th row of the matrix Θ_0 and $\Omega_{q\times q} = \int \mathbf{B}''(t)\mathbf{B}''(t)^T dt$. The term I is the weighted least square criterion. The term II is the group lasso penalty which encourages row sparsity of the coefficient matrix Θ_0 and thereby facilitates the selection of the over-complete wavelet basis. With this regularization, important spectral features can be identified, whereas the noise gets filtered out. The term III is the roughness penalty to encourage the smooth evolution of

the mean surface along the time (phase) dimension. The smoothness is achieved by controlling the norm of the second-order derivative of the mean function along the time dimension.

The optimization problem (3.11) is solved by the fast iterative shrinkage-thresholding algorithm [80, FISTA]. The tunning parameters η and ρ are selected by K-fold cross-validation. More details of the algorithm is discussed below. For the simplicity of notation, we assume the weighting matrix $\mathbf{V}_j^{(s)-1}$ is absorbed by $\mathbf{y}_j^{(s)}$ and $\mathbf{W}_j^{(s)}$.

Let $\mathscr{L}(\Theta_0) = \mathrm{I} + \mathrm{III}$ be the summation of the smooth and differentiable terms of (3.11), the gradient of $\mathscr{L}(\Theta_0)$ with respect to matrix Θ_0 is given by

$$\nabla_{\Theta_0}\mathscr{L}(\Theta_0) = -2\sum_{s=1}^{S}\sum_{j=1}^{J^{(s)}}\frac{1}{I_j^{(s)}}\mathbf{W}_j^{(s)\top}\left(\mathbf{y}_j^{(s)} - \mathbf{W}_j^{(s)}\Theta_0\mathbf{B}_j^{(s)}\right)\mathbf{B}_j^{(s)\top} + 2\rho\Theta_0\mathbf{\Omega}. \qquad (3.12)$$

FISTA updates Θ_0 along the negative gradient by $\Theta_0 = \Theta_0 - \gamma\nabla_{\Theta_0}\mathscr{L}(\Theta_0)$, and meanwhile induces sparsity (incurred by II) via applying the soft-thresholding operator to the updated Θ_0. The rows of Θ_0 with relatively small norm are set to be exact zero, and the rows with large norm shrinks towards zero. The expression for the soft-thresholding operator is

$$\Theta_0^m = \left(1 - \frac{c}{\|\Theta_0^m\|}\right)_+ \Theta_0^m, \qquad (3.13)$$

where c is a constant depending on the algorithm step size and the tunning parameter η. The notation $(\cdot)_+$ standards for the operation that $(x)_+ = x$ if $x > 0$ and $(x)_+ = 0$ otherwise. The algorithm for estimating the mean surface is presented in Algorithm 3.

3.6.1.2 Learning the Principal Component Surfaces

Suppose $\widehat{\Theta}_0$ is the optimizer of the optimization problem (3.11), the corresponding mean surface is $\widehat{\phi}_0(\lambda, t) = \mathbf{W}(\lambda)\widehat{\Theta}_0\mathbf{B}(t)$. We subtract the mean surface from the observed spectra $\mathbf{y}_j^{(s)}$ and denote the residual spectra as $\widetilde{\mathbf{y}}_j^{(s)}$, i.e. $\widetilde{\mathbf{y}}_j^{(s)} = \mathbf{y}_j^{(s)} - \mathbf{W}_j^{(s)}\widehat{\Theta}_0\mathbf{B}_j^{(s)}$. The principal component surfaces are estimated based on the residual spectra. With a fixed R, the coefficient matrices $\{\Theta_r\}_{r=1}^{R}$ and the principal component scores $\{\boldsymbol{\beta}^{(s)}\}_{s=1}^{S}$ can be estimated by solving the optimization prob-

Algorithm 3: Learning the Mean Surface $\phi_0(\lambda, t)$.

Require: Stepsize: γ. Weight paramerter $k_1 = 1$.

Require: Initialization: the coefficient matrix of mean spectrophotometric time series Θ and the intermediate parameter Θ'. (We suppress the subscript for notation simplicity.)

1: **while** not converged **do**
2:      Update iteration $t \leftarrow t + 1$.
3:      Calculate gradient at iteration t using equation (3.12): $g_t \leftarrow \nabla_\Theta \mathscr{L}(\Theta_t')$.
4:      Update matrix $\Theta_t = \Theta_t' + \gamma \cdot g_t$ and apply soft-thresholding operator to Θ_t using equation (3.13).
5:      $k_{t+1} \leftarrow \left(1 + \sqrt{1 + 4k_t^2}\right)/2$.
6:      Update the intermediate parameter $\Theta_{t+1}' \leftarrow \Theta_t + (k_t - 1)/k_{t+1} \cdot (\Theta_t - \Theta_{t-1})$.
7: **end while**
8: **return** Θ. The mean function is then calculated by $\phi_0(\lambda, t) = \mathbf{W}(\lambda)^\top \Theta \mathbf{B}(t)$.

lem, which is based on (3.6),

$$
\min_{\{\Theta_r\}_{r=1}^R, \{\boldsymbol{\beta}^{(s)}\}_{s=1}^S} \sum_{s=1}^S \sum_{j=1}^{J^{(s)}} \frac{1}{I_j^{(s)}} \|\widetilde{\mathbf{y}}_j^{(s)} - \mathbf{W}_j^{(s)} \otimes \mathbf{B}_j^{s\top} \Theta \boldsymbol{\beta}^{(s)}\|_2^2 + \mathscr{P}_{\eta'}(\Theta) + \rho' \mathrm{tr}[\Theta^\top (\mathbf{I} \otimes \boldsymbol{\Omega})\Theta],
$$

$$
\text{subject to } \Theta^\top \Theta = \mathbf{I}_R.
$$

$$(3.14)$$

where $\Theta_{pq \times R} = \left(\mathrm{vec}(\Theta_1), \ldots, \mathrm{vec}(\Theta_R)\right)$ and $\boldsymbol{\beta}^{(s)} = \left(\beta_1^{(s)}, \ldots \beta_R^{(s)}\right)$. The penalty term $\mathscr{P}_{\eta'}(\Theta)$ encourages sparse wavelet representation of spectral features. Specifically, $\mathscr{P}_{\eta'}(\Theta) = \sum_{m=1}^p \eta' \|\Theta^m\|$ with Θ^m being the block matrix corresponding to the $(m-1)q+1, \cdots, mq$-th rows of Θ. A roughness penalty $\mathrm{tr}[\Theta^\top (\mathbf{I} \otimes \boldsymbol{\Omega})\Theta]$ is also included to encourage smooth time evolution of the component surfaces. The tuning parameter η' is selected by K-fold cross-validation.

Solving the optimization (3.14) is challenging due to the following reasons. The algorithm could potentially have a high computational cost at each iteration, especially for a large dataset, because the wavelet basis is over-complete and high-dimensional. Meanwhile, the algorithm should respect the orthonormality constraint $\Theta^\top \Theta = \mathbf{I}_R$ and the sparsity-inducing penalty $\mathscr{P}_{\eta'}$.

For efficient optimization searching, we use an efficient first-order stochastic gradient algorithm called *Adam* [28]. The search is combined with a manifold optimization technique [97] to respect

the orthonormality constraint $\mathbf{\Theta}^{\top}\mathbf{\Theta} = \mathbf{I}_R$. The sparse-inducing soft-thresholding (3.13) will also be applied.

The algorithm proceeds in a way that the parameters $\{\mathbf{\Theta}_r\}_{r=1}^R$ and $\{\boldsymbol{\beta}^{(s)}\}_{s=1}^S$ get updated alternatively until convergence. Given $\{\boldsymbol{\beta}^{(s)}\}_{s=1}^S$, a random supernova is drawn to compute the gradient with respect to $\mathbf{\Theta}$. Suppose the sampled supernova is indexed by s, and $\mathscr{J}(\mathbf{\Theta})$ is the summation of the first and third (smooth and differentiable) term of (3.14). Its gradient with respect to $\mathbf{\Theta}$ is

$$\nabla_{\mathbf{\Theta}}\mathscr{J}(\mathbf{\Theta}) = \sum_{j=1}^{J^{(s)}} \frac{1}{I_j^{(s)}} \left(\mathbf{W}_j^{(s)} \otimes \mathbf{B}_j^{(s)\top}\right)^{\top} \left(\widetilde{\mathbf{y}}_j^{(s)} - \mathbf{W}_j^{(s)} \otimes \mathbf{B}_j^{(s)\top}\mathbf{\Theta}\boldsymbol{\beta}^{(s)}\right)\boldsymbol{\beta}^{(s)\top} + 2\rho'(\mathbf{I} \otimes \mathbf{\Omega})\mathbf{\Theta}.$$

(3.15)

In Adam, this stochastic gradient will be adjusted by historical momentum and volatility for each element of the parameter matrix $\mathbf{\Theta}$. The adjusted gradient matrix is denoted as $\mathbf{D}$, and its negative will be employed as an update direction for $\mathbf{\Theta}$. Due to the orthonormal constraint on $\mathbf{\Theta}$, the matrix $\mathbf{\Theta}$ belongs to a special matrix manifold called Stiefel manifold. The geometrical structure of the manifold can be further exploited for efficient update. We project the update direction $\mathbf{D}$ onto the tangent space of the Stiefel manifold via

$$\mathrm{Proj}\left(\mathbf{D}\right) = \mathbf{D} - \mathbf{\Theta}\frac{\mathbf{\Theta}^{\top}\mathbf{D} + \mathbf{D}^{\top}\mathbf{\Theta}}{2}.$$

(3.16)

The matrix $\mathbf{\Theta}$ gets updated along the negative of the projected direction, and then the sparsity inducing soft-thresholding operator (3.13) in applied. Finally, the updated $\mathbf{\Theta}$ is retracted to the Stiefel manifold using QR decomposition, which ensures $\mathbf{\Theta}^{\top}\mathbf{\Theta} = \mathbf{I}_R$ at the end of each iteration.

Given $\mathbf{\Theta}$, the scores $\{\boldsymbol{\beta}^{(s)}\}_{s=1}^S$ are updated according to the weighted least squares criterion, i.e.

$$\widehat{\boldsymbol{\beta}}^{(s)} = \arg\min \sum_{j=1}^{J^{(s)}} \frac{1}{I_j^{(s)}} \|\widetilde{\mathbf{y}}_j^{(s)} - \mathbf{W}_j^{(s)} \otimes \mathbf{B}_j^{(s)\top}\mathbf{\Theta}\boldsymbol{\beta}^{(s)}\|^2.$$

(3.17)

The algorithm iteratively updates the two groups of parameters until convergence. Suppose $\widehat{\mathbf{\Theta}}$ is the solution of the optimization problem (3.14), the resulting principal component surfaces are

$\widehat{\phi}_r(\lambda, t) = \mathbf{W}(\lambda)^\top \otimes \mathbf{B}(t)^\top \widehat{\boldsymbol{\Theta}}_r$, where $\widehat{\boldsymbol{\Theta}}_r$ is the rth column of the matrix $\widehat{\boldsymbol{\Theta}}$ for $r = 1, \ldots, R$. We summarize the procedure in Algorithm 4.

Algorithm 4: Learning the Principal Component Surfaces $\phi_r(\lambda, t)$, $r = 1, \ldots, R$.

Require: Stepsize: γ; Exponential decay rates: $\beta_1 = 0.9$ and $\beta_2 = 0.999$; $\epsilon = 10^{-8}$.
Require: Initialization: $\boldsymbol{\Theta}_0$, m_0 and v_0.
1: **while** not converged **do**
2:    Update iteration $t \leftarrow t + 1$.
3:    Calculate unbiased gradient at iteration t using equation (3.15): $g_t \leftarrow S \cdot \nabla_{\boldsymbol{\Theta}} \mathcal{J}(\boldsymbol{\Theta}_{t-1})$.
4:    Update the biased first and second raw moment estimate:

$$m_t \leftarrow m_{t-1} + (1 - \beta_1) \cdot g_t,$$
$$v_t \leftarrow v_{t-1} + (1 - \beta_2) \cdot g_t \odot g_t \quad \text{(elementwise product)}.$$

5:    Correct the bias for the estimates above

$$m'_t \leftarrow m_t / (1 - \beta_1^t),$$
$$v'_t \leftarrow v_t / (1 - \beta_2^t).$$

6:    Define $\mathbf{D}_t = m'_t / \left(\sqrt{v'_t} + \epsilon \right)$ (with elementwise square-root and division) and project it onto the tagent space of $\boldsymbol{\Theta}_{t-1}$ using equation (3.16).
7:    Update parameters $\boldsymbol{\Theta}_t \leftarrow \boldsymbol{\Theta}_{t-1} - \gamma \cdot \text{Proj}(\mathbf{D}_t)$.
8:    Apply soft-thresholding operator towards $\boldsymbol{\Theta}_t$ using equation (3.13) and then retract the resulting parameters to Stiefel manifold using QR decomposition.
9:    Update scores $\{\boldsymbol{\beta}_t^{(s)}\}_{s=1}^S$ using equation (3.17).
10: **end while**
11: **return** $\boldsymbol{\Theta}_t$ and $\{\boldsymbol{\beta}^{(s)}\}_{s=1}^S$

3.6.2 Independent Component Analysis Related Results

In Section 3.5.3, we formulated two hypothesis about the supernova color composition and presented the intrinsic color models under two hypotheses based on CCR $\psi(\lambda)$. This section presents the parallel results of the development of the two intrinsic color models and hypothesis testing results related with $\psi_{ica}(\lambda)$.

The definition of *color parameter* corresponding to $\psi_{ica}(\lambda)$ is given by

$$\alpha_{ica}^{(s)} = \underset{\alpha_{ica}^{(s)}}{\arg\min} \left\{ \log\left(y^{(s)}(\lambda,t)\right) - \left[\phi_0(\lambda,t) - \alpha_{ica}^{(s)}\psi_{ica}(\lambda)\right] \right\}^2.$$

Figure 3.16 shows the scatterplot of the color parameter α_{ica} versus intrinsic property parameter γ_1 and the correspondingly developed two hypotheses are highlighted with blue and red solid lines.
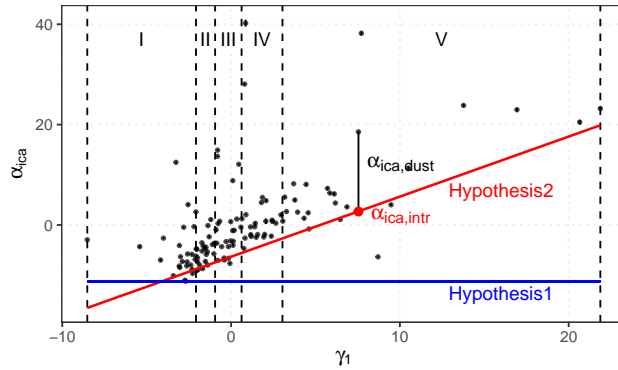
Figure 3.16: Same as in Figure 3.11, but for ICA

The distributions of $\alpha_{ica,dust}$ for five subgroups splited by the quantile levels of γ_1 under two hypotheses are plotted in Figure 3.17.
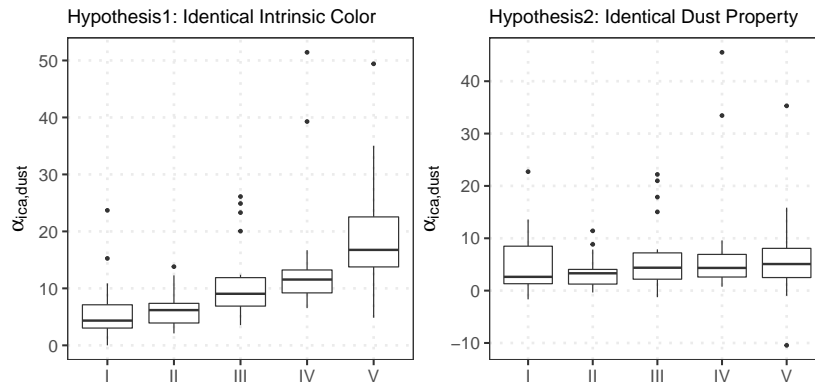
Figure 3.17: Same as in Figure 3.12, but for ICA.

The p-values of Kolmogorov-Smirnov test performed between each pair of five subgroups under two hypotheses are reported in Table 3.2.

Table 3.2: The p-value of Kolmogorov–Smirnov test for two hypotheses of $\alpha_{ica,dust}$. Same as in Table 3.1, but for ICA.

	I	II	III	IV	V
I	1.000	0.163	**0.005**	**0.000**	**0.000**
II	0.616	1.000	**0.012**	**0.000**	**0.000**
III	0.426	0.140	1.000	**0.012**	**0.000**
IV	0.230	0.068	0.686	1.000	**0.000**
V	0.285	0.062	0.942	0.450	1.000

4. SUMMARY AND CONCLUSIONS

Functional data analysis is concerned with the development of theoretical and applicable tools for statistical analysis of the functional data.

As the core concept in functional data analysis, covariance operator generalizes the definition of covariance matrix to depict the interdependence and the computation cost to construct its empirical estimate grows overwhelming as the data size increases. In the first part of this thesis, we studied the randomized algorithms to estimate the covariance operator by forming a sketch of much smaller size by subsampling according to some probability distributions. The optimal sampling probability that minimizes the expected squared Hilbert-Schmidt norm of the subsampling error is determined by the norm of each function. The concentration bounds are developed for both sampling probabilities, showing that under some regularity conditions, the target precision can be guaranteed with high probability. The mathematical accuracy and computational efficiency have been elaborated in numerical experiments and real word problems.

In the second part, we presented two empirical statistical models of SN Ia spectrophotometric data: flux-based FPCA (fb-FPCA) and magnitude-based FPCA (mb-FPCA). The construction of the models utilizes FPCA and the entire process is developed in a purely data-driven fashion and the use of the stochastic first-order gradient descent optimization method improves the computational efficiency. With these models, the spectral-temporal surface of SNe Ia with non-uniformly sampled observations can be evaluated at any desired wavelength and phase, and it can be expressed by a linear combination of multiple components whose scores are regarded as SN Ia low-rank representation.

To distinguish interstellar dust extinction from the intrinsic color properties of SNe Ia and to quantify them separately, a novel procedure is developed to extract the independent physical processes involved in the observed spectral data of SNe Ia. The derivation of the parameters of the intrinsic property relies on the scores of the extinction free flux ratios, whose construction involves the estimation of the spectrum at B-band maximum of each SN and division of the spectropho-

tometric time series by the spectrum at maximum. The flux ratio surface is independent of the effect of interstellar dust extinction and enables quantitative analyses of the intrinsic properties of SN Ia. We constructed two CCRs by taking the time-average and smoothing the wavelength dependence of the first components of the FPCA and ICA decompositions. They are used to define the wavelength dependence of interstellar dust extinction and to derive the color parameters of the SNe.

The color parameters are separated into a component of SN Ia intrinsic color and a component of interstellar dust reddening. This is achieved by investigating the relation between the color parameters and intrinsic property parameters derived from flux ratios. By removing the supernova intrinsic color dependence, we derived two ICE-ECs to account for the interstellar dust properties and the ICE-EC of the ICA shows excellent agreement with the CCM extinction law [2] with $R_V \lesssim 2.0$ but the counterpart of FPCA shows larger R_V value inconsistent with well observed SNe with multi-band data. Finally, after correcting the effect due to dust extinction, we are able to construct the intrinsic SED surfaces of SNe Ia. This intrinsic SED surface can be used to derive the essential extinction properties of interstellar dust along the line-of-sight as well as the principle components of the intrinsic properties of the SNe Ia.

For the future work, more statistical inference properties such as asymptotic behavior of randomized estimators remain to be investigated for the static data. In the dynamic scenario where streaming data continuously arrives, the design of the randomized algorithm should takes more practical limitation into consideration. The idea of randomization can also be applied to more functional data analysis methods and large-scale astronomical dataset.

REFERENCES

[1] S. He and X. Yan, "Randomized estimation of functional covariance operator via subsampling," *Stat*, vol. 9, no. 1, p. e311, 2020.

[2] J. A. Cardelli, G. C. Clayton, and J. S. Mathis, "The Relationship between Infrared, Optical, and Ultraviolet Extinction," *The Astrophysical Journal*, vol. 345, p. 245, Oct. 1989.

[3] J. Ramsay and B. W. Silverman, "Functional data analysis," 2001.

[4] G. Zhao, Y.-H. Zhao, Y.-Q. Chu, Y.-P. Jing, and L.-C. Deng, "Lamost spectral survey—an overview," *Research in Astronomy and Astrophysics*, vol. 12, no. 7, p. 723, 2012.

[5] D. J. Eisenstein, D. H. Weinberg, E. Agol, H. Aihara, C. A. Prieto, S. F. Anderson, J. A. Arns, É. Aubourg, S. Bailey, E. Balbinot, *et al.*, "Sdss-iii: Massive spectroscopic surveys of the distant universe, the milky way, and extra-solar planetary systems," *The Astronomical Journal*, vol. 142, no. 3, p. 72, 2011.

[6] P. Liu, L. Di, Q. Du, and L. Wang, "Remote sensing big data: Theory, methods and applications," *Remote Sensing*, vol. 10, no. 5, p. 711, 2018.

[7] T. S. Tian, "Functional data analysis in brain imaging studies," *Frontiers in Psychology*, vol. 1, pp. 35–35, 2010.

[8] N. B. Turkbrowne, "Functional interactions as big data in the human brain," *Science*, vol. 342, no. 6158, pp. 580–584, 2013.

[9] P. Z. Hadjipantelis and H.-G. Müller, "Functional data analysis for big data: A case study on california temperature trends," in *Handbook of Big Data Analytics*, pp. 457–483, Springer, 2018.

[10] P. Drineas, R. Kannan, and M. W. Mahoney, "Fast monte carlo algorithms for matrices i: Approximating matrix multiplication," *SIAM Journal on Computing*, vol. 36, no. 1, pp. 132–157, 2006.

[11] P. Drineas, M. Magdon-Ismail, M. W. Mahoney, and D. P. Woodruff, "Fast approximation of matrix coherence and statistical leverage," *Journal of Machine Learning Research*, vol. 13, no. Dec, pp. 3475–3506, 2012.

[12] P. Ma, M. W. Mahoney, and B. Yu, "A statistical perspective on algorithmic leveraging," *The Journal of Machine Learning Research*, vol. 16, no. 1, pp. 861–911, 2015.

[13] P. Drineas, M. W. Mahoney, and S. Muthukrishnan, "Subspace sampling and relative-error matrix approximation: Column-row-based methods," in *Algorithms - ESA 2006, 14th Annual European Symposium, Zurich, Switzerland, September 11-13, 2006, Proceedings*, 2006.

[14] H. Wang, R. Zhu, and P. Ma, "Optimal subsampling for large sample logistic regression," *Journal of the American Statistical Association*, vol. 113, no. 522, pp. 829–844, 2018.

[15] H. Wang, "More efficient estimation for logistic regression with optimal subsamples," *Journal of Machine Learning Research*, vol. 20, no. 132, pp. 1–59, 2019.

[16] T. Zhang, Y. Ning, and D. Ruppert, "Optimal sampling for generalized linear models under measurement constraints," *Journal of Computational and Graphical Statistics*, vol. 30, no. 1, pp. 106–114, 2021.

[17] P. Drineas, M. W. Mahoney, S. Muthukrishnan, and T. Sarlós, "Faster least squares approximation," *Numerische mathematik*, vol. 117, no. 2, pp. 219–249, 2011.

[18] S. Wang, A. Gittens, and M. W. Mahoney, "Sketched ridge regression: Optimization perspective, statistical perspective, and model averaging," *The Journal of Machine Learning Research*, vol. 18, no. 1, pp. 8039–8088, 2017.

[19] P. Drineas, R. Kannan, and M. W. Mahoney, "Fast monte carlo algorithms for matrices ii: Computing a low-rank approximation to a matrix," *SIAM Journal on computing*, vol. 36, no. 1, pp. 158–183, 2006.

[20] P. Drineas, R. Kannan, and M. W. Mahoney, "Fast monte carlo algorithms for matrices iii: Computing a compressed approximate matrix decomposition," *SIAM Journal on Computing*, vol. 36, no. 1, pp. 184–206, 2006.

[21] M. Pilanci and M. J. Wainwright, "Randomized sketches of convex programs with sharp guarantees," *IEEE Transactions on Information Theory*, vol. 61, no. 9, pp. 5096–5115, 2015.

[22] M. Pilanci and M. J. Wainwright, "Newton sketch: A near linear-time optimization algorithm with linear-quadratic convergence," *SIAM Journal on Optimization*, vol. 27, no. 1, pp. 205–245, 2017.

[23] N. Halko, P. Martinsson, and J. A. Tropp, "Finding structure with randomness: Probabilistic algorithms for constructing approximate matrix decompositions," *Siam Review*, vol. 53, no. 2, pp. 217–288, 2011.

[24] Y. Yang, M. Pilanci, M. J. Wainwright, *et al.*, "Randomized sketches for kernels: Fast and optimal nonparametric regression," *The Annals of Statistics*, vol. 45, no. 3, pp. 991–1023, 2017.

[25] P. Drineas and M. W. Mahoney, "Lectures on randomized numerical linear algebra," *The Mathematics of Data*, vol. 25, p. 1, 2018.

[26] D. P. Woodruff, "Sketching as a tool for numerical linear algebra," *arXiv preprint arXiv:1411.4357*, 2014.

[27] G. Raskutti and M. W. Mahoney, "A statistical perspective on randomized sketching for ordinary least-squares," *Journal of Machine Learning Research*, vol. 17, no. 1, pp. 7508–7538, 2016.

[28] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[29] L. Horváth and P. Kokoszka, *Inference for functional data with applications*, vol. 200. Springer Science & Business Media, 2012.

[30] T. Cai and M. Yuan, "Nonparametric covariance function estimation for functional and longitudinal data," *Technical Report*, 2010.

[31] L. Xiao, V. Zipunnikov, D. Ruppert, and C. M. Crainiceanu, "Fast covariance estimation for high-dimensional functional data," *Statistics and Computing*, vol. 26, no. 1, pp. 409–421, 2016.

[32] J. A. Tropp *et al.*, "An introduction to matrix concentration inequalities," *Foundations and Trends® in Machine Learning*, vol. 8, no. 1-2, pp. 1–230, 2015.

[33] V. Koltchinskii, K. Lounici, *et al.*, "Asymptotics and concentration bounds for bilinear forms of spectral projectors of sample covariance," in *Annales de l'Institut Henri Poincaré, Probabilités et Statistiques*, vol. 52, pp. 1976–2013, Institut Henri Poincaré, 2016.

[34] S. Minsker, "On some extensions of bernstein's inequality for self-adjoint operators," *Statistics & Probability Letters*, vol. 127, pp. 111–119, 2017.

[35] L. H. Dicker, D. P. Foster, D. Hsu, *et al.*, "Kernel ridge vs. principal component regression: Minimax bounds and the qualification of regularization operators," *Electronic Journal of Statistics*, vol. 11, no. 1, pp. 1022–1047, 2017.

[36] A. G. Riess, W. H. Press, and R. P. Kirshner, "A Precise Distance Indicator: Type IA Supernova Multicolor Light-Curve Shapes," *The Astrophysical Journal*, vol. 473, p. 88, Dec. 1996.

[37] S. Perlmutter, G. Aldering, G. Goldhaber, R. A. Knop, P. Nugent, P. G. Castro, S. Deustua, S. Fabbro, A. Goobar, D. E. Groom, I. M. Hook, A. G. Kim, M. Y. Kim, J. C. Lee, N. J. Nunes, R. Pain, C. R. Pennypacker, R. Quimby, C. Lidman, R. S. Ellis, M. Irwin, R. G. McMahon, P. Ruiz-Lapuente, N. Walton, B. Schaefer, B. J. Boyle, A. V. Filippenko, T. Matheson, A. S. Fruchter, N. Panagia, H. J. M. Newberg, W. J. Couch, and T. S. C. Project, "Measurements of Ω and Λ from 42 High-Redshift Supernovae," *The Astrophysical Journal*, vol. 517, pp. 565–586, June 1999.

[38] B. P. Schmidt, N. B. Suntzeff, M. M. Phillips, R. A. Schommer, A. Clocchiatti, R. P. Kirshner, P. Garnavich, P. Challis, B. Leibundgut, J. Spyromilio, A. G. Riess, A. V. Filippenko, M. Hamuy, R. C. Smith, C. Hogan, C. Stubbs, A. Diercks, D. Reiss, R. Gilliland, J. Tonry,

J. Maza, A. Dressler, J. Walsh, and R. Ciardullo, "The High-Z Supernova Search: Measuring Cosmic Deceleration and Global Curvature of the Universe Using Type IA Supernovae," *The Astrophysical Journal*, vol. 507, pp. 46–63, Nov. 1998.

[39] Y. P. Pskovskii, "Photometric classification and basic parameters of type I supernovae," *Soviet Astronomy*, vol. 28, pp. 658–664, Dec. 1984.

[40] M. M. Phillips, "The Absolute Magnitudes of Type IA Supernovae," *The Astrophysical Journal l*, vol. 413, p. L105, Aug. 1993.

[41] LSST Dark Energy Science Collaboration, "Large Synoptic Survey Telescope: Dark Energy Science Collaboration," *arXiv e-prints*, p. arXiv:1211.0310, Nov. 2012.

[42] D. Spergel, N. Gehrels, C. Baltay, D. Bennett, J. Breckinridge, M. Donahue, A. Dressler, B. S. Gaudi, T. Greene, O. Guyon, C. Hirata, J. Kalirai, N. J. Kasdin, B. Macintosh, W. Moos, S. Perlmutter, M. Postman, B. Rauscher, J. Rhodes, Y. Wang, D. Weinberg, D. Benford, M. Hudson, W. S. Jeong, Y. Mellier, W. Traub, T. Yamada, P. Capak, J. Colbert, D. Masters, M. Penny, D. Savransky, D. Stern, N. Zimmerman, R. Barry, L. Bartusek, K. Carpenter, E. Cheng, D. Content, F. Dekens, R. Demers, K. Grady, C. Jackson, G. Kuan, J. Kruk, M. Melton, B. Nemati, B. Parvin, I. Poberezhskiy, C. Peddie, J. Ruffa, J. K. Wallace, A. Whipple, E. Wollack, and F. Zhao, "Wide-Field InfrarRed Survey Telescope-Astrophysics Focused Telescope Assets WFIRST-AFTA 2015 Report," *arXiv e-prints*, p. arXiv:1503.03757, Mar. 2015.

[43] B. W. Rust, "The Use of Supernovae Light Curves for Testing the Expansion Hypothesis and Other Cosmological Relations.," Jan. 1974.

[44] I. P. Pskovskii, "Light curves, color curves, and expansion velocity of type I supernovae as functions of the rate of brightness decline," *Soviet Astronomy*, vol. 21, p. 675, Dec. 1977.

[45] S. Perlmutter, S. Gabi, G. Goldhaber, A. Goobar, D. E. Groom, I. M. Hook, A. G. Kim, M. Y. Kim, J. C. Lee, R. Pain, C. R. Pennypacker, I. A. Small, R. S. Ellis, R. G. McMahon, B. J. Boyle, P. S. Bunclark, D. Carter, M. J. Irwin, K. Glazebrook, H. J. M. Newberg, A. V.

Filippenko, T. Matheson, M. Dopita, and W. J. Couch, "Measurements of the Cosmological Parameters Ω and Λ from the First Seven Supernovae at z >= 0.35," *The Astrophysical Journal*, vol. 483, pp. 565–581, July 1997.

[46] C. R. Burns, M. Stritzinger, M. M. Phillips, E. Y. Hsiao, C. Contreras, S. E. Persson, G. Folatelli, L. Boldt, A. Campillay, S. Castellón, W. L. Freedman, B. F. Madore, N. Morrell, F. Salgado, and N. B. Suntzeff, "The carnegie supernova project: Intrinsic colors of type ia supernovae," *The Astrophysical Journal*, vol. 789, p. 32, Jun 2014.

[47] M. M. Phillips, P. Lira, N. B. Suntzeff, R. A. Schommer, M. Hamuy, and J. Maza, "The Reddening-Free Decline Rate Versus Luminosity Relationship for Type IA Supernovae," *The Astronomical Journal*, vol. 118, pp. 1766–1776, Oct. 1999.

[48] L. Wang, G. Goldhaber, G. Aldering, and S. Perlmutter, "Multicolor Light Curves of Type Ia Supernovae on the Color-Magnitude Diagram: A Novel Step toward More Precise Distance and Extinction Estimates," *The Astrophysical Journal*, vol. 590, pp. 944–970, June 2003.

[49] R. Tripp, "A two-parameter luminosity correction for Type IA supernovae," *The Astronomy and Astrophysics*, vol. 331, pp. 815–820, Mar. 1998.

[50] J. Guy, P. Astier, S. Baumont, D. Hardin, R. Pain, N. Regnault, S. Basa, R. G. Carlberg, A. Conley, S. Fabbro, D. Fouchez, I. M. Hook, D. A. Howell, K. Perrett, C. J. Pritchet, J. Rich, M. Sullivan, P. Antilogus, E. Aubourg, G. Bazin, J. Bronder, M. Filiol, N. Palanque-Delabrouille, P. Ripoche, and V. Ruhlmann-Kleider, "SALT2: using distant supernovae to improve the use of type Ia supernovae as distance indicators," *The Astronomy and Astrophysics*, vol. 466, pp. 11–21, Apr. 2007.

[51] W. D. Kenworthy, D. O. Jones, M. Dai, R. Kessler, D. Scolnic, D. Brout, M. R. Siebert, J. D. R. Pierel, K. G. Dettman, G. Dimitriadis, R. J. Foley, S. W. Jha, Y. C. Pan, A. Riess, S. Rodney, and C. Rojas-Bravo, "SALT3: An Improved Type Ia Supernova Model for Measuring Cosmic Distances," *The Astrophysical Journal*, vol. 923, p. 265, Dec. 2021.

[52] A. Conley, R. G. Carlberg, J. Guy, D. A. Howell, S. Jha, A. G. Riess, and M. Sullivan, "Is There Evidence for a Hubble Bubble? The Nature of Type Ia Supernova Colors and Dust in External Galaxies," *The Astrophysical Journal l*, vol. 664, pp. L13–L16, July 2007.

[53] E. L. Fitzpatrick and D. Massa, "An Analysis of the Shapes of Interstellar Extinction Curves. V. The IR-through-UV Curve Morphology," *The Astrophysical Journal*, vol. 663, pp. 320–341, July 2007.

[54] F. Patat, S. Taubenberger, N. L. J. Cox, D. Baade, A. Clocchiatti, P. Höflich, J. R. Maund, E. Reilly, J. Spyromilio, L. Wang, J. C. Wheeler, and P. Zelaya, "Properties of extragalactic dust inferred from linear polarimetry of Type Ia Supernovae," *The Astronomy and Astrophysics*, vol. 577, p. A53, May 2015.

[55] Y. Yang, L. Wang, D. Baade, P. J. Brown, A. Cikota, M. Cracraft, P. A. Höflich, J. R. Maund, F. Patat, W. B. Sparks, J. Spyromilio, H. F. Stevance, X. Wang, and J. C. Wheeler, "Mapping Circumstellar Matter with Polarized Light: The Case of Supernova 2014J in M82," *The Astrophysical Journal*, vol. 854, p. 55, Feb. 2018.

[56] L. Wang, M. Strovink, A. Conley, G. Goldhaber, M. Kowalski, S. Perlmutter, and J. Siegrist, "Nonlinear Decline-Rate Dependence and Intrinsic Variation of Type Ia Supernova Luminosities," *The Astrophysical Journal*, vol. 641, pp. 50–69, Apr. 2006.

[57] L. Wang, "Dust around Type Ia Supernovae," *The Astrophysical Journal l*, vol. 635, pp. L33–L36, Dec. 2005.

[58] F. Patat, "Reflections on reflexions - I. Light echoes in Type Ia supernovae," *Monthly Notices of the Royal Astronomical Society*, vol. 357, pp. 1161–1177, Mar. 2005.

[59] A. Goobar, "Low R_V from Circumstellar Dust around Supernovae," *The Astrophysical Journal l*, vol. 686, p. L103, Oct. 2008.

[60] M. Bulla, A. Goobar, and S. Dhawan, "Shedding light on the Type Ia supernova extinction puzzle: dust location found," *Monthly Notices of the Royal Astronomical Society*, vol. 479, pp. 3663–3674, Sept. 2018.

[61] M. Hu, L. Wang, and X. Wang, "The Effects of Circumstellar Dust Scattering on the Light Curves and Polarizations of Type Ia Supernovae," *arXiv e-prints*, p. arXiv:2109.05504, Sept. 2021.

[62] E. Y. Hsiao, A. Conley, D. A. Howell, M. Sullivan, C. J. Pritchet, R. G. Carlberg, P. E. Nugent, and M. M. Phillips, "K-Corrections and Spectral Templates of Type Ia Supernovae," *The Astrophysical Journal*, vol. 663, pp. 1187–1200, July 2007.

[63] C. Saunders, G. Aldering, P. Antilogus, S. Bailey, C. Baltay, K. Barbary, D. Baugh, K. Boone, S. Bongard, C. Buton, J. Chen, N. Chotard, Y. Copin, S. Dixon, P. Fagrelius, H. K. Fakhouri, U. Feindt, D. Fouchez, E. Gangler, B. Hayden, W. Hillebrandt, A. G. Kim, M. Kowalski, D. Küsters, P. F. Leget, S. Lombardo, J. Nordin, R. Pain, E. Pecontal, R. Pereira, S. Perlmutter, D. Rabinowitz, M. Rigault, D. Rubin, K. Runge, G. Smadja, C. Sofiatti, N. Suzuki, C. Tao, S. Taubenberger, R. C. Thomas, M. Vincenzi, and T. Nearby Supernova Factory, "SNEMO: Improved Empirical Models for Type Ia Supernovae," *The Astrophysical Journal*, vol. 869, p. 167, Dec. 2018.

[64] H. Akaike, "A new look at the statistical model identification," *IEEE transactions on automatic control*, vol. 19, no. 6, pp. 716–723, 1974.

[65] X. Chen, L. Hu, and L. Wang, "Artificial Intelligence-Assisted Inversion (AIAI) of Synthetic Type Ia Supernova Spectra," *The Astrophysical Journal s*, vol. 250, p. 12, Sept. 2020.

[66] B. E. Stahl, J. Martínez-Palomera, W. Zheng, T. de Jaeger, A. V. Filippenko, and J. S. Bloom, "deepSIP: linking Type Ia supernova spectra to photometric quantities with deep learning," *Monthly Notices of the Royal Astronomical Society*, vol. 496, pp. 3553–3571, Aug. 2020.

[67] L. Hu, X. Chen, and L. Wang, "Spectroscopic Studies of Type Ia Supernovae Using LSTM Neural Networks," *arXiv e-prints*, p. arXiv:2202.02498, Feb. 2022.

[68] W. E. Kerzendorf and S. A. Sim, "A spectral synthesis code for rapid modelling of supernovae," *Monthly Notices of the Royal Astronomical Society*, vol. 440, pp. 387–404, May 2014.

[69] J. O. Ramsay, "Functional data analysis," *Encyclopedia of Statistical Sciences*, vol. 4, 2004.

[70] J. Herault and C. Jutten, "Space or time adaptive signal processing by neural network models," vol. 151, no. 1, pp. 206–211, 1986.

[71] P. Comon, "Independent component analysis, a new concept?," *Signal processing*, vol. 36, no. 3, pp. 287–314, 1994.

[72] G. Aldering, G. Adam, P. Antilogus, P. Astier, R. Bacon, S. Bongard, C. Bonnaud, Y. Copin, D. Hardin, F. Henault, D. A. Howell, J.-P. Lemonnier, J.-M. Levy, S. C. Loken, P. E. Nugent, R. Pain, A. Pecontal, E. Pecontal, S. Perlmutter, R. M. Quimby, K. Schahmaneche, G. Smadja, and W. M. Wood-Vasey, "Overview of the Nearby Supernova Factory," in *Survey and Other Telescope Technologies and Discoveries* (J. A. Tyson and S. Wolff, eds.), vol. 4836 of *Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series*, pp. 61–72, Dec. 2002.

[73] G. Aldering, P. Antilogus, C. Aragon, S. Bailey, C. Baltay, S. Bongard, K. Boone, C. Buton, N. Chotard, Y. Copin, S. Dixon, H. K. Fakhouri, U. Feindt, D. Fouchez, E. Gangler, B. Hayden, W. Hillebrandt, A. G. Kim, M. Kowalski, D. Küsters, P. F. Léget, Q. Lin, S. Lombardo, F. Mondon, J. Nordin, R. Pain, E. Pecontal, R. Pereira, S. Perlmutter, K. Ponder, M. Pruzhinskaya, D. Rabinowitz, M. Rigault, D. Rubin, K. Runge, C. Saunders, L. P. Says, G. Smadja, N. Suzuki, C. Tao, S. Taubenberger, R. C. Thomas, M. Vincenzi, B. Weaver, and Nearby Supernova Factory Collaboration, "The SNEMO and SUGAR Companion Data Sets," *Research Notes of the American Astronomical Society*, vol. 4, p. 63, May 2020.

[74] P. F. Léget, E. Gangler, F. Mondon, G. Aldering, P. Antilogus, C. Aragon, S. Bailey, C. Baltay, K. Barbary, S. Bongard, K. Boone, C. Buton, N. Chotard, Y. Copin, S. Dixon, P. Fagrelius, U. Feindt, D. Fouchez, B. Hayden, W. Hillebrandt, A. Kim, M. Kowalski, D. Kuesters, S. Lombardo, Q. Lin, J. Nordin, R. Pain, E. Pecontal, R. Pereira, S. Perlmutter, K. A. Ponder, M. V. Pruzhinskaya, D. Rabinowitz, M. Rigault, K. Runge, D. Rubin, C. Saunders, L. P. Says, G. Smadja, C. Sofiatti, N. Suzuki, S. Taubenberger, C. Tao, and R. C. Thomas, "SUGAR: An

improved empirical model of Type Ia supernovae based on spectral features," *The Astronomy and Astrophysics*, vol. 636, p. A46, Apr. 2020.

[75] S. He, L. Wang, and J. Z. Huang, "Characterization of Type Ia Supernova Light Curves Using Principal Component Analysis of Sparse Functional Data," *The Astrophysical Journal*, vol. 857, p. 110, Apr. 2018.

[76] A. Wagers, L. Wang, and S. Asztalos, "Quantifying Spectral Features of Type Ia Supernovae," *The Astrophysical Journal*, vol. 711, pp. 711–730, Mar. 2010.

[77] R. Tibshirani, "Regression shrinkage and selection via the lasso," *Journal of the Royal Statistical Society: Series B (Methodological)*, vol. 58, no. 1, pp. 267–288, 1996.

[78] C. K. Chui and J.-z. Wang, "On compactly supported spline wavelets and a duality principle," *Transactions of the American Mathematical Society*, vol. 330, no. 2, pp. 903–915, 1992.

[79] M. Yuan and Y. Lin, "Model selection and estimation in regression with grouped variables," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 68, no. 1, pp. 49–67, 2006.

[80] A. Beck and M. Teboulle, "A fast iterative shrinkage-thresholding algorithm for linear inverse problems," *SIAM journal on imaging sciences*, vol. 2, no. 1, pp. 183–202, 2009.

[81] B. Efron and R. J. Tibshirani, *An introduction to the bootstrap*. CRC press, 1994.

[82] M. Stone, "Cross-validatory choice and assessment of statistical predictions," *Journal of the royal statistical society: Series B (Methodological)*, vol. 36, no. 2, pp. 111–133, 1974.

[83] M. Stone, "An asymptotic equivalence of choice of model by cross-validation and akaike's criterion," *Journal of the Royal Statistical Society: Series B (Methodological)*, vol. 39, no. 1, pp. 44–47, 1977.

[84] K. P. Murphy, *Machine learning: a probabilistic perspective*. MIT press, 2012.

[85] L. Wang, D. Baade, P. Höflich, A. Khokhlov, J. C. Wheeler, D. Kasen, P. E. Nugent, S. Perlmutter, C. Fransson, and P. Lundqvist, "Spectropolarimetry of SN 2001el in NGC 1448:

Asphericity of a Normal Type Ia Supernova," *The Astrophysical Journal*, vol. 591, pp. 1110–1128, July 2003.

[86] D. Kasen, P. Nugent, L. Wang, D. A. Howell, J. C. Wheeler, P. Höflich, D. Baade, E. Baron, and P. H. Hauschildt, "Analysis of the Flux and Polarization Spectra of the Type Ia Supernova SN 2001el: Exploring the Geometry of the High-Velocity Ejecta," *The Astrophysical Journal*, vol. 593, pp. 788–808, Aug. 2003.

[87] F. Patat, P. Chandra, R. Chevalier, S. Justham, P. Podsiadlowski, C. Wolf, A. Gal-Yam, L. Pasquini, I. A. Crawford, P. A. Mazzali, A. W. A. Pauldrach, K. Nomoto, S. Benetti, E. Cappellaro, N. Elias-Rosa, W. Hillebrandt, D. C. Leonard, A. Pastorello, A. Renzini, F. Sabbadin, J. D. Simon, and M. Turatto, "Detection of Circumstellar Material in a Normal Type Ia Supernova," *Science*, vol. 317, p. 924, Aug. 2007.

[88] X. Wang, J. Chen, L. Wang, M. Hu, G. Xi, Y. Yang, X. Zhao, and W. Li, "The Cold and Dusty Circumstellar Matter around Fast-expanding Type Ia Supernovae," *The Astrophysical Journal*, vol. 882, p. 120, Sept. 2019.

[89] A. Hyvärinen and E. Oja, "A fast fixed-point algorithm for independent component analysis," *Neural computation*, vol. 9, no. 7, pp. 1483–1492, 1997.

[90] A. Hyvarinen, "Fast and robust fixed-point algorithms for independent component analysis," *IEEE transactions on Neural Networks*, vol. 10, no. 3, pp. 626–634, 1999.

[91] A. Hyvärinen and E. Oja, "Independent component analysis: algorithms and applications," *Neural networks*, vol. 13, no. 4-5, pp. 411–430, 2000.

[92] R. Amanullah, A. Goobar, J. Johansson, D. P. K. Banerjee, V. Venkataraman, V. Joshi, N. M. Ashok, Y. Cao, M. M. Kasliwal, S. R. Kulkarni, P. E. Nugent, T. Petrushevska, and V. Stanishev, "The Peculiar Extinction Law of SN 2014J Measured with the Hubble Space Telescope," *The Astrophysical Journal l*, vol. 788, p. L21, June 2014.

[93] K. Krisciunas, N. C. Hastings, K. Loomis, R. McMillan, A. Rest, A. G. Riess, and C. Stubbs, "Uniformity of (V-Near-Infrared) Color Evolution of Type Ia Supernovae and Implications

for Host Galaxy Extinction Determination," *The Astrophysical Journal*, vol. 539, pp. 658–674, Aug. 2000.

[94] N. Elias-Rosa, S. Benetti, E. Cappellaro, M. Turatto, P. A. Mazzali, F. Patat, W. P. S. Meikle, M. Stehle, A. Pastorello, G. Pignata, R. Kotak, A. Harutyunyan, G. Altavilla, H. Navasardyan, Y. Qiu, M. Salvo, and W. Hillebrandt, "Anomalous extinction behaviour towards the Type Ia SN 2003cg," *Monthly Notices of the Royal Astronomical Society*, vol. 369, pp. 1880–1900, July 2006.

[95] N. Elias-Rosa, S. Benetti, M. Turatto, E. Cappellaro, S. Valenti, A. A. Arkharov, J. E. Beckman, A. di Paola, M. Dolci, A. V. Filippenko, R. J. Foley, K. Krisciunas, V. M. Larionov, W. Li, W. P. S. Meikle, A. Pastorello, G. Valentini, and W. Hillebrandt, "SN 2002cv: a heavily obscured Type Ia supernova," *Monthly Notices of the Royal Astronomical Society*, vol. 384, pp. 107–122, Feb. 2008.

[96] X. Wang, A. V. Filippenko, M. Ganeshalingam, W. Li, J. M. Silverman, L. Wang, R. Chornock, R. J. Foley, E. L. Gates, B. Macomber, F. J. D. Serduke, T. N. Steele, and D. S. Wong, "Improved Distances to Type Ia Supernovae with Two Spectroscopic Subclasses," *The Astrophysical Journal l*, vol. 699, pp. L139–L143, July 2009.

[97] P.-A. Absil, R. Mahony, and R. Sepulchre, *Optimization algorithms on matrix manifolds*. Princeton University Press, 2009.