Some Theoretical Essays on Functional Data Classification

Agne KAZAKEVICIUTE

A dissertation submitted in partial fulfillment

of the requirements for the degree of

Doctor of Philosophy

of

University College London.

Department of Statistical Science

University College London

August 15, 2017

I, Agne KAZAKEVICIUTE, confirm that the work presented in this thesis is my own. Where information has been derived from other sources, I confirm that this has been indicated in the work.

Abstract

Functional data analysis is a fast-growing research area in statistics, dealing with statistical analysis of infinite-dimensional (functional) data. For many pattern recognition problems with finite-dimensional data there usually exists a solid theoretical foundation, for example, it is known under which assumptions various classifiers have desirable theoretical properties, such as consistency. Therefore, a natural interest is to extend the theory to the setting of infinite-dimensional data.

The thesis is written in two directions: one is when we observe full curves, and the other is when we observe sparse and irregular curves. In the first direction, the main goal is to give a justification for a logistic classifier, where only the projection of the parameter function on some subspace is estimated via maximum quasi-likelihood and the rest of its coordinates are set to zero. This is preceded with studying the problem of detecting sample point separation in logistic regression–the case in which the maximum quasi-likelihood estimate of the model parameter does not exist or is not unique. In the other direction, a problem of extending sparsely and irregularly sampled functional data to full curves is considered so that potentially the theory from the first research direction could be applied in the future.

There are several contributions of this thesis. First, it is proved that the separating hyperplane can be found from a finite set of candidates, and an upper bound of the probability of point separation is given. Second, the assumptions under which the logistic classifier is consistent are established, although simulation studies reveal that some assumptions are not necessary and may be relaxed. Thirdly, the thesis proposes a collaborative curve extension method, which is proven to be consistent under certain assumptions.

Acknowledgements

After having red the masterpiece acknowledgement section of Francesco Donat's PhD thesis, I had a feeling that I do not want to write my acknowledgements at all–there is no chance I can put together something as poetic as that! But then I had a second thought on it. Aknowledgements are the only part of the thesis that is not examined and the only part that myself I will be reading ten years from now. It does not have to be perfect, it simply has to be worth remembering.

My first thank you goes to Dr Giampiero Marra. That was summer of 2013 that I came to UCL to talk about my MSc grades when I accidentally bumped into Dr Giampiero in the corridor and ended up in his office talking about possible funding for a PhD. What a day! It may sound unbelievable but it is absolutely true: if you had not encouraged me to apply for a PhD, I would have never applied for it.

My second thank you goes to Dr Jinghao Xue. First of all, for selecting me for the PhD even though I may have been the only candidate in the set of candidates for you to choose from (probably nobody wanted to go to Singapore). Second, for providing me with extensive criticism on my work throughout the PhD which caused me at least one heart attack every time I red an e-mail message from you and an increased blood pressure for at least six remaining days (until your next e-mail message). On the other hand, I must admit that this pushed me to work harder, perform better and write less nonsense in my weakly reports. Third, for allowing me to study mathematical statistics even though this was not your main research interest and for referring me to several books on functional data analysis. Fourth, for polishing every sentence of this thesis and especially for helping me to put together a great abstract.

Here I want to squeeze a brief thank you note also for prof. Simon Arridge and Dr Ben Cox with whom I worked in the first year of my PhD studies. I still remember the quite impressive meeting with prof. Simon Arridge during which he spotted my mistake in 5 seconds without even looking at my code. Even though I have changed my research topic both of your inputs to my work at that time have helped me to pass the PhD upgrade exam without which I would have died in London.

My third huge thank you goes to prof. Malini Olivo. The support you have given me during my attachment with A*STAR is tremendous. You gave me all the time in the world to work on my PhD project but at the same time I had the possibility to be involved into interdisciplinary collaborations, team projects, team training and various company events, seminars, talks and meetings. You have supported me throughout my first paper, my first conference and on whatever decision I made. You have personally supported me both through the sad times and through the happy times. You basically were my family when I was far away from mine. I do not know a better leader than you are.

I also would like to thank prof. Tom Fearn for visiting me in Singapore and for telling me that nobody can forbid me from trying to prove theorems, if I want to, and for telling me that you do not see the reason why I could not submit my thesis on time which at that time seemed a big issue for me. Also, for caring about my research progress and even calling me over Skype despite the eight-hour time difference and technical sound issues which you have originally solved by using your landline telephone.

I would also like to acknowledge my A*STAR colleague Dr Chris Jun Hui Ho for supporting me with real data and proofreading my first-ever paper on photoacoustic imaging which later became my proof to myself that I am capable of achieving something and proof to others that sometimes even incremental contributions can be published in good journals. Also for inviting me to participate in hackathons which has broaden my horizons and awaken my entrepreneurship spirit.

Finally I would like to thank the sharpest mind I have ever met and my biggest mathematical authority my father prof. Vytautas Kazakevicius for referring me to

many useful books and teaching me that writing a single correct sentence is more than writing several wrong ones. People consider you a mathematician but to me you are more than that, you are an artist, a philosopher. If I have ever learned to see objects through numbers like in The Matrix movie, I learned it from you. I remember you told me there exist structures walking through which you become a reversed version of yourself. I also remember you told me that if God spoke any language that language would be mathematics. You are the reason why up to this day I still hold a passion and admiration for mathematics–the world of ideas and the world of truth.

I remember in the beginning of my PhD (and later in Singapore) I saw a very funny illustrated guide to a PhD by Matt Might, where he described a PhD as the following: during your PhD studies you keep moving to the boundary of knowledge in your research area, then you push at the boundary for a few years and that small dent that you make after those years is called a PhD. I do not know how it worked out for him but myself I have never reached that boundary. It seems that the boundary kept moving away from me faster than I moved towards it. In fact, after all these years instead of having a small dent I ended up with a very scary question whether or not that boundary exists in the first place. So what do we actually call a PhD?

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Prologue

This thesis concerns challenges arising when classifying functional data both in theory and in practice. The main goal of the thesis is to investigate which assumptions lead to consistent classification of functional data. Based on this, two directions are investigated in the thesis: one is on assuming that we observe full curves and the other is on assuming that we observe curves sparsely and irregularly.

In the first direction, the main goal is to give a justification for a logistic classifier, where data come from an abstract Hilbert space but only the projection of the parameter function on some subspace is estimated via maximum-likelihood and the rest of the coordinates of the parameter function are set to zero. The goal is achieved in two steps. The first step involves calculating the probability that a maximum quasi-likelihood estimate exists and is unique and investigating under which assumptions this probability tends to 1. This is shown to be deviating to another research area of sample point separation in logistic regression. The second step involves investigating assumptions on the distribution of data and on the dimension for projection that are needed to obtain a consistent resulting logistic estimate of the parameter function. The subspaces in this step are assumed to be non-random, even though some guidelines on how they should look like to yield consistency are given, based on the distribution of data which in practice is unknown. Based on these guidelines, the subspaces could be selected adaptively (that is, depending on

data) in the future, potentially by using principal component analysis (PCA).

In the second direction, I study a common case of functional data appearing in practice, where the data are sampled sparsely and irregularly. To link the two directions of this thesis, I investigate the ways of extending observed data to full curves so that theoretical results from the first research direction could be applied in the future. I propose a consistent way to estimate the reference maximum and minimum curves in a collaborative fashion and then to predict the unobserved function values by interpolation shifted vertically. Under certain assumptions, I then prove the consistency of the proposed curve extension approach.

This Introduction is structured as follows. I first discuss the meaning of functional data and its differences to vectorial data in Section 1.2. I then describe the binary classification task in Section 1.3. Finally, I state the contributions and structure of this thesis in Section 1.4.

1.2 The nature of functional data

In the functional data setting, the data come from a functional space E instead of a finite-dimensional space $\mathbb{R}^k$. In this thesis I will discuss the case where the sample data are independent identically distributed observations drawn from the same distribution as some E-valued random element X.

There are two common choices for functional space E [4]:

1. A separable Banach space (complete normed vector space). For example, $\mathrm{C}[0,1]$ – the space of real continuous functions $x : [0,1] \to \mathbb{R}$ endowed with the norm $\|x\| = \sup_t |x(t)|$.

2. A separable Hilbert space (an abstract vector space with the defined inner product and complete with respect to the induced norm). For example, $L^2[0,1]$ – the space of square integrable real functions on $[0,1]$ endowed with the usual inner product $\langle x_1, x_2 \rangle = \int_0^1 x_1(t)x_2(t)\mathrm{d}t$. Note, however, that any Hilbert space is also a Banach space with the norm $\|x\| = \sqrt{\langle x, x \rangle}$. For exam-

ple, the norm in $L^2[0,1]$ space is defined by

$$\|x\|_{L_2}^2 = \int_0^1 x^2(t)\mathrm{d}t.$$

In this thesis, I will refer to the data from E as *curves*, *functions*, *functional observations* or *elements of functional space* E, while the data from the finite-dimensional $\mathbb{R}^k$ space as *finite-dimensional vectors*, *vectorial observations* or *elements of* $\mathbb{R}^k$. I will say that functional observations are observations of the process in time and that vectorial observations are observations of variables, even though this is only for the differentiation of the two.

Sometimes, we can borrow techniques from the $\mathbb{R}^k$ setting and apply them to solve related problems in the functional data setting. In theory, we can always assume that we observe full curves $X_i(t), i = 1,\ldots,n$. Then the only fundamental difference between a functional observation and a vectorial observation is that the functional observation is infinite-dimensional, while the vectorial observation is finite-dimensional. In the $\mathbb{R}^k$ setting, a special attention is recently given to the so-called *high-dimensional* case where the number of observations n in the sample is less than the number of variables k. Therefore, in high-dimensional case, various pattern recognition tasks, such as classification, involve estimating the parameter vector whose length k is greater than the number of observations n. From general algebra, it is known that any system of equations has a non-unique solution if the number of variables is larger than the number of equations. That is, estimating parameter vector in high-dimensional case results in non-uniqueness of the solution. Moreover, a curse of dimensionality and overfitting are also common problems when working with high-dimensional data [5]. To avoid these problems, usually a dimensionality reduction step is included which projects the high-dimensional observation into some k_n-dimensional subspace, where $k_n < n$. The open problem is then how to select good dimensions and to select a good k_n for projection. The same principle can be used also for functional data, where the same problem needs to be tackled.

In practice, however, we never observe full curves as we observe them at some finite number of time points. Moreover, the time points are likely to differ from observation to observation, that is, we observe $X_i(T_{i1}),\dots,X_i(T_{iM_i}), i=1,\dots,n$, where M_i are, for example, independent copies of some random variable M. Depending on how data come to a researcher, the functional data are usually classified into *densely observed curves*, where the distribution of M does depend on n in a way that the number of time points at which we observe the ith function diverges together with n, and *sparsely observed curves*, where the distribution of M does not depend on n [6]. The problem is then how to process such data as we cannot apply the standard techniques used in $\mathbb{R}^k$.

1.3 Binary classification for functional data

In this thesis I study the logistic classifier which is a binary classifier. A task of binary classification is to attach every x from a functional space E to one of the two groups, 0 or 1 (sometimes, -1 and 1). Formally, a *binary classifier* is a Borel function $h : E \to \{0,1\}$ [7]. The requirement of h to be a Borel function guarantees that $h(X)$ is a random variable. The pair (x,y), where $y \in \{0,1\}$ is the true group of x, is considered as a realization of a random vector (X,Y). Suppose the distribution of X is μ, and the conditional probability of $Y=1$, given $X=x$, is $p(x)$. The function p is an element of $\mathrm{L}^1(E,\mu)$, the space of all μ-integrable functions (meaning that such a function is measurable and that the integral of the function w.r.t. μ is defined) endowed with the semi-metric

$$d(p_1,p_2) = \int_E \big|p_1(x) - p_2(x)\big| \mathrm{d}\mu. \tag{1.1}$$

Naturally, semi-metric (1.1) tells us how distant the functions p_1 and p_2 are in the space $\mathrm{L}^1(E,\mu)$.

Choosing this semi-metric is a common practice and was previously used for classification of functional data (see, e.g. [8]). Choosing this semi-metric is also common for theoretical inference, e.g. to measure how close or far the estimated conditional probability $\hat{p}$ is, when compared with the true conditional probability

p_0. However, in practice, the distribution μ of X and the true conditional probability p_0 are unknown. They can be estimated either parametrically or non-parametrically from the training set $(X_1, Y_1), \ldots, (X_n, Y_n)$. Each estimator $\hat{p}$ induces a class (u-class) of classifiers of the form

$$\hat{h}_u(x) = \begin{cases} 1, & \text{if } \hat{p}(x) > u, \\ 0, & \text{if otherwise,} \end{cases} \tag{1.2}$$

where u is a pre-selected threshold. In other words, different values of a threshold u induce different classifiers $\hat{h}_u$. The choice of u depends on a researcher's needs to control Type I and Type II errors (a.k.a. false positives and false negatives) with the usual choice being $u = 1/2$ which means that the cost of making Type I error is the same as that of making Type II error. However, other choices for u are also possible, such as that of setting u to be the rate of responses $Y = 0$ in the training set [9].

1.4 Epilogue

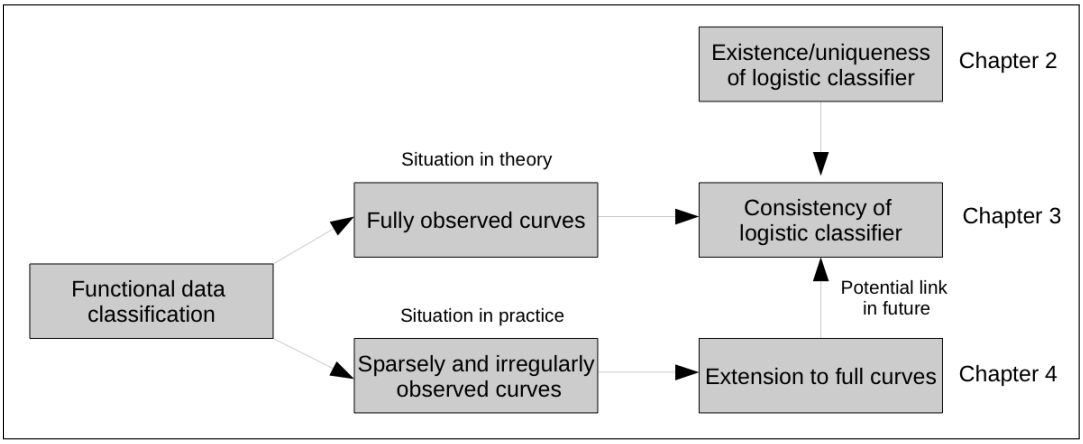

Figure 1.1: Structure of the main body of the thesis.

There are several contributions of this thesis that can be outlined (see Figure 1.1). The thesis begins with investigating what is the probability that the maximum quasi-likelihood estimate of the parameter function in logistic regression exists and is unique. Since it is already known that it exists and is unique if and only if there

is no separation of sample points, this leads to investigating the probability of point separation in logistic regression for functional variables in Chapter 2. Here two contributions are made. The first involves proving that the separating hyperplane can be found from a finite set of candidate hyperplanes, a result that has not yet been proved in literature. The second contribution involves giving an upper bound of the probability that a sample is separable in expression of which the dimension for projection k_n is included. This allows to directly derive the assumption on k_n so that the upper bound tends to 0.

In Chapter 3 I investigate under which assumptions on the distribution of X and on the dimension for projection k_n the logistic estimate is consistent which is the open problem described in Section 1.2. The consistency for generalized linear models when data come from the Hilbert space was already investigated in [10]. However, one of their assumptions in proving consistency of maximum quasi-likelihood estimate of model parameters is not valid in the case of logistic regression model. The main contribution of this Chapter is therefore proving the consistency of a logistic classifier for Hilbert space-valued random variables.

Finally, in Chapter 4 I study a common situation of functional data appearing in practice, where data are observed sparsely and irregularly as described in Section 1.2. I propose a consistent way to estimate reference maximum and minimum curves in a collaborative fashion, similarly as the mean function was estimated in [11]. I then propose a method for predicting unobserved function values by interpolation shifted vertically based on the estimated reference functions. The main contribution of this Chapter is that, under certain data model, I prove the consistency of the proposed method.

To conclude, the contributions of this thesis are summarized in the following papers:

- 'Point Separation in Logistic Regression on Hilbert Space-Valued Variables', *Published in Statistics & Probability Letters* (with prof. M. Olivo).
- 'Consistency of Logistic Classifier in Abstract Hilbert Spaces', *To be submitted* (with prof. M. Olivo).

- 'Extending Sparsely and Irregularly Sampled Functional Data Using Collaborative Prediction', *To be submitted*.

Chapter 2

Probability of Point Separation in Logistic Regression for Functional Variables

We study point separation for the logistic regression model for Hilbert space-valued variables. It is known that in the case of sample point separation, the maximum quasi-likelihood estimate of parameter function does not exist or exists but is not unique. As a consequence, there is no strict definition for the logistic estimate for such data arrangement which leads to problems when proving its consistency. To mitigate the negative effects of such data arrangement, we investigate assumptions under which the probability of point separation tends to 0. We achieve this by proving that the separating hyperplane can be found from a set containing a finite number of candidates and giving an upper bound for the probability of point separation.

2.1 Introduction to the problem

The problem of point separation in logistic regression has been studied since as early as in [12] and more than 700 papers have cited [12] since then. In [12] the authors established the conditions on the maximum-likelihood estimate of the parameter vector in logistic regression model to exist when data come from the $\mathbb{R}^k$ space. Three scenarios of the arrangement of the data points were introduced: complete separation, quasi-complete separation and overlap. The authors proved that in the

first two scenarios, the maximum-likelihood estimate of parameter vector does not exist, or exists but is not unique, while in the third (overlap) scenario the maximum-likelihood estimate exists and is unique. The authors also suggested an iterative algorithm to be used when checking, whether or not the data points are in quasi-complete separation. Other methods on detecting overlap have been established as well (see, e.g. [13]).

The majority of papers in this research area are devoted to proposing new parameter estimates that would exist and would have desirable theoretical properties in the case where the data are already known to be in complete or quasi-complete separation. For example, the penalized maximum-likelihood estimator was introduced by [14] and asymptotically investigated by [15], while [16] proposed a hidden logistic regression model to overcome the problem of non-uniqueness of the parameter estimate. Based on the recent activity in the field (see, e.g. [17] or [18], where they investigated which methods work well in quasi-complete separation, or [19], where they proposed adaptive prior weighting to avoid complete separation), we believe that various results on the problem of point separation in logistic regression in the $\mathbb{R}^k$ setting are still of a great interest.

Moreover, with the recent expansion of functional data analysis (FDA) (see [20], [21] for an overview of the topic), the functional logistic regression models have been widely studied. The logistic estimate in abstract Hilbert spaces can be called a naïve approach because the dimensionality reduction is achieved by simply cutting the infinite-dimensional observation after some $k_n < n$ time point, where n is the number of sample points. In such a way the first k_n parameter values are estimated via maximum-likelihood and the rest are set to zero. This approach is avoided in literature for various reasons. For example, [22] argued that the naïve approach in the context of functional data introduces multicollinearity (strong dependence among predictors) which in turn causes inaccurate parameter estimates and increases their variance. Therefore, the standard approaches include dimensionality reduction based on Principal Component Analysis (PCA) or Partial Least Squares (PLS) (see, e.g. [23], [24], [25], [26]) or by basis expansion with some

added penalty (see e.g. [27] or [28]). In none of these cases the consistency of functional logistic regression model parameter was established, mainly because the optimal rule for selecting the number of principal components or basis functions has not been established. The closest attempt to provide the theoretical justification of such a rule was done in [10]. However, in their work the authors approximated infinite-dimensional model by a finite-dimensional one without proving that the error of such an approximation tends to 0.

There are two theoretical contributions of this Chapter. First is that we provide a theorem which transforms the problem of finding the separating hyperplane from the set of infinitely many elements into a feasible problem of finding it from the finite set of candidate hyperplanes and we describe how to construct such a set. We believe this theorem could speed up various established algorithms used by practitioners for determining whether or not a maximum-likelihood estimate exists or is unique for given datasets. The second contribution is that we provide an upper bound of the probability of the event that a sample is in quasi-complete separation. As a corollary of the latter result, we derive the minimal requirements on the selection of the dimension k_n for projection of the data such that the consistency of the resulting functional logistic estimate could be expected. We will use this result in Chapter 3.

2.2 Logistic estimate in abstract Hilbert spaces

Let E be a separable Hilbert space with the inner product $\langle \cdot, \cdot \rangle$. Let $X \in E$ be a Hilbert space-valued random variable and Y a random variable, gaining values -1 and 1, with conditional probabilities (w.r.t. X), $1 - p_{\theta_0}(X)$ and $p_{\theta_0}(X)$, respectively. Here $\theta_0 \in E$ is an unknown parameter and

$$p_\theta(x) = \frac{1}{1 + \mathrm{e}^{-\langle \theta, x \rangle}}, \ \theta, x \in E.$$

For example, if $E = \ell^2$, the space of all square-summable sequences, then $\langle \theta, x \rangle = \sum_{k=1}^{\infty} \theta_k x_k$. If $E = L^2([0,1])$, then $\langle \theta, x \rangle = \int_0^1 \theta(t)x(t)\mathrm{d}t$. Since E can be any Hilbert space, we will work with the general notation $\langle \theta, x \rangle$ instead.

Let $(X_1,Y_1),\dots,(X_n,Y_n)$ be a random sample from the distribution of (X,Y). For $\theta,x\in E$ and $y\in\{-1,1\}$ define

$$m_\theta(x,y)=\log(1+\mathrm{e}^{-y\langle\theta,x\rangle})$$

and denote

$$M_n(\theta)=\overline{m_\theta(X,Y)}=\frac{m_\theta(X_1,Y_1)+\dots+m_\theta(X_n,Y_n)}{n},\quad M(\theta)=\mathsf{E}m_\theta(X,Y).$$

Note that

$$M_n(\theta)=\frac{1}{n}\sum_{i=1}^n\log(1+\mathrm{e}^{-Y_i\langle\theta,X_i\rangle})=\frac{1}{n}\log\prod_{i=1}^n(1+\mathrm{e}^{-Y_i\langle\theta,X_i\rangle})=-\frac{1}{n}\log\prod_{i=1}^n q_\theta(X_i,Y_i),$$

where

$$q_\theta(X_i,Y_i)=\frac{1}{1+\mathrm{e}^{-Y_i\langle\theta,X_i\rangle}}.$$

Obviously, $q_\theta(X_i,1)=p_\theta(X_i)$ and $q_\theta(X_i,-1)=1-p_\theta(X_i)$. Also, for any bounded f,

$$\mathsf{E}f(X,Y)=\int f(x,y)q_\theta(x,y)\mu(\mathrm{d}x)\nu(\mathrm{d}y)=\int f(x,1)q_\theta(x,1)\mu(\mathrm{d}x)+\int f(x,-1)q_\theta(x,-1)\mu(\mathrm{d}x),$$

where ν is a counting measure in the set $\{-1,1\}$. Therefore $q_\theta(x,y)$ is a density of (X,Y) w.r.t. the measure $\mu\times\nu$. Hence, since μ is unknown, $M_n(\theta)$ can be interpreted as the logarithm of the quasi-likelihood function, multiplied by $-1/n$.

Naturally, for various practical tasks it is of great interest to provide an estimate of p_θ.

Let (E_k) be some fixed sequence of the linear subspaces of the space E such that the following conditions are satisfied: (1) $\dim E_k=k$ for all k, (2) $E_k\subset E_{k+1}$ for all k, and (3) $\overline{\bigcup_k E_k}=E$. For any k and n define

$$\hat\theta_{kn}=\arg\min_{\theta\in E_k}M_n(\theta).\tag{2.1}$$

Note that taking $\theta \in E_k$ in the above expression introduces some approximation error. To force this error to tend to 0 as n diverges, fix some sequence (k_n) and set

$$\hat{\theta} = \hat{\theta}_{k_n n} \quad \text{and} \quad \hat{p} = p_{\hat{\theta}}. \tag{2.2}$$

We will call $\hat{p}$ the *logistic estimate* of the conditional probability p_{θ_0}. For example, let $E = L^2(T)$ with the usual inner product

$$\langle \theta, x \rangle = \int_T \theta(t)x(t)\mathrm{d}t,$$

where $T \subset \mathbb{R}$ is an interval and L^2 is defined in Section 1.2. The standard method for obtaining logistic estimate from a given sample $(X_1, Y_1), \ldots, (X_n, Y_n)$ is expanding X and θ via selected basis functions $\{e_j\}$

$$X_i(t) = \sum_{j=1}^{\infty} X_{ij} e_j(t), \quad \theta(t) = \sum_{j=1}^{\infty} \theta_j e_j(t),$$

choosing $k = k_n$ and then using (2.1), where

$$E_k = \left\{ \sum_{j=1}^{k} c_j e_j \mid c_1, \ldots, c_k \in \mathbb{R} \right\}.$$

The number k_n of basis functions to be used is usually selected less than n so that the parameter vector could be estimable. However, there are two open problems. First is that (as discussed before) the estimate (2.1) does not exist or is not unique, if sample points are separable. This results in convergence to a false estimate which causes biased results. Second problem is that it is not clear how to select k_n with respect to n so that the resulting estimate would be consistent, for example. In Section 2.3 we solve the first problem, where we describe how separation of points can be checked against in practice. In Section 2.4 we partially solve the second problem, where we give the minimal requirements for k_n so that consistency of the resulting estimate (2.1) could be expected.

Remark 2.1. If $\theta \in E_k$, then $\langle \theta, X \rangle = \langle \theta, X^{(k)} \rangle$, where $X^{(k)}$ is the orthogonal pro-

jection of X on the space E_k. Therefore, $\hat{\theta}_{kn}$ is obtained only from $X_i^{(k)}$, $i = 1,\ldots,n$. One could get a wrong idea that then the data are from $\mathbb{R}^k$ and we do not need to consider the general case when calculating the probability of point separation. However, the situation is more difficult than this. While the conditional probability of $Y = 1$, w.r.t. X, is denoted by $p_\theta(X)$ and has a nice expression, the same conditional probability w.r.t. $X^{(k)}$ is not $p_\theta(X^{(k)})$ but $\mathsf{E}^{X^{(k)}} p_\theta(X)$, where $\mathsf{E}^{X^{(k)}}$ is the conditional expectation w.r.t. $X^{(k)}$.

2.3 Separability of sample points

Let $(x_1,y_1),\ldots,(x_n,y_n)$ be n vectors from $E_k \times \{-1,1\}$. We will call them *sample points*. Let $a \neq 0$ be another vector from E_k. We will say that a vector a *separates sample points* if, for all i,

$$y_i\langle a,x_i\rangle \geq 0.$$

We say that sample points are *separable*, if there exists some $a \neq 0$ that separates them. Note that this definition is equivalent to the definition of quasi-complete separation in the $\mathbb{R}^k$ case, established by [12].

Obviously, if some vector a separates sample points, then vector ca with any $c > 0$ also separates them. However, $-ca$ with any $c > 0$ does not separate them. The separability of sample points has also a geometric interpretation. Any nonzero vector a corresponds to a hyperplane H_a which is defined by the equation $\langle a,x\rangle = 0$ (note that 0 is used in this equation due to the fact that in this thesis we consider the logistic model without an intercept term). The vector a is then a normal of a hyperplane H_a. The subsets of E, defined by inequalities $\langle a,x\rangle \geq 0$ and $\langle a,x\rangle \leq 0$, are then called *half-spaces* of E. If we change a to ca with $c > 0$, the associated hyperplane as well as the associated half-spaces will not change. If we change a to $-ca$ with $c > 0$, the associated hyperplane will not change but the associated half-spaces will have the reversed order. If a' is not proportional to a, the associated hyperplanes differ. Therefore, a hyperplane defines a normal to a precision up to a constant c. Moreover, a hyperplane uniquely defines the pair of half-spaces, rather than individual half-spaces. If we want a hyperplane to define a normal to a precision up

to a positive constant c, we have to introduce an *oriented hyperplane*. Formally speaking, an oriented hyperplane is a hyperplane with a fixed unit length normal. An oriented hyperplane uniquely defines individual half-spaces, and we can call one of the two half-spaces *an upper half-space*, and another one *a lower half-space*. For example, the upper half-space is defined by the equation $\langle a,x\rangle \geq 0$, where a is that fixed normal. If a separates sample points and H is the corresponding hyperplane, we can say that points from different groups fall into different half-spaces. Of course, one has to keep in mind that those half-spaces overlap, that is, points on the hyperplane belong to both half-spaces. If H is an oriented hyperplane and $a/\|a\|$ is its fixed normal, then points from the group $y = 1$ belong to the upper half-space, while the rest belong to the lower half-space.

Denote by $X_i^{(k)}$ the projection of the point X_i on the space E_k. We will say that the sample $(X_1,Y_1),\ldots,(X_n,Y_n)$ is *k-separable*, if the random sample points $(X_1^{(k)},Y_1),\ldots,(X_n^{(k)},Y_n)$ are separable. The latter definition defines some subset of the event space Ω that consists of $\omega \in \Omega$ for which the sample points

$$(X_1^{(k)}(\omega),Y_1(\omega)),\ldots,(X_n^{(k)}(\omega),Y_n(\omega)) \tag{2.3}$$

are separable. It is well-known that if the sample is k-separable, then the maximum quasi-likelihood estimate of θ does not exist or is not unique [12].

When searching for a separating hyperplane, there are infinitely many candidate hyperplanes to consider. This fact makes the theoretical investigation of the probability that the sample is separable harder since the sums of infinitely many possible separating hyperplanes are involved in the calculations. In practice the search area of an algorithm for finding the possible separating hyperplane is restricted to some set of finite number of candidate hyperplanes that is guaranteed to contain the true separating hyperplane. However, this has not been proved yet. In the following Section, we give a proof for this.

Let $(e_1,\ldots,e_k)$ be the orthonormal basis in E_k. For any $x_1,\ldots,x_k \in E_k$, we will

denote

$$\det[x_1,\dots,x_k] = \begin{vmatrix} c_{11} & \dots & c_{k1} \\ \vdots & \ddots & \vdots \\ c_{1k} & \dots & c_{kk} \end{vmatrix},$$

where c_{ij} is the jth coordinate of the ith covariate, that is,

$$x_i = c_{i1}e_1 + \dots + c_{ik}e_k.$$

. Obviously, det is a k-linear antisymmetric form.

Since $\det[x_1,\dots,x_{k-1},x]$ is a linear function w.r.t. x, it is of the form $\langle a,x\rangle$ with some a. In other words, there exists a unique a such that, for all x, $\det[x_1,\dots,x_{k-1},x] = \langle a,x\rangle$. Obviously, a is a function of $x_1,\dots,x_{k-1}$.

If $x_1,\dots,x_{k-1}$ are linearly dependent, the determinant is equal to 0 for all x, that is, $a = 0$. Conversely, if $a = 0$, then $x_1,\dots,x_{k-1}$ are linearly dependent (otherwise we could find x_k for which $x_1,\dots,x_k$ are linearly independent which would imply that the determinant is nonzero, that is, $a \neq 0$).

There is an intrinsic relationship between a determinant and a hyperplane. If $x_1,\dots,x_{k-1}$ are linearly independent, then $a \neq 0$ defines some hyperplane H_a. This hyperplane has the special property that points $x_1,\dots,x_{k-1}$ belong to it (because determinant is equal to 0 when any two columns in it are equal). In fact, it is the unique hyperplane that contains these points because all a that are perpendicular to all $x_1,\dots,x_{k-1}$ are proportional.

Suppose $n \geq k$. We will prove that when checking the separability of sample points it is enough to sort out a finite number of potential vectors a that possibly separate the sample. This will allow us the correct use of (A.1) in Appendix A.2. Note that the set of such possible vectors is random. For any family of distinct indexes $(i_1,\dots,i_{k-1}) \subset \{1,\dots,n\}$ denote by $Z_{i_1\dots i_{k-1}}$ a random vector from E_k such that, for all $x \in E_k$,

$$\det[X^{(k)}_{i_1},\dots,X^{(k)}_{i_{k-1}},x] = \langle Z_{i_1\dots i_{k-1}},x\rangle.$$

Let

$$S = \{\pm Z_{i_1 \ldots i_{k-1}} \mid (i_1, \ldots, i_{k-1}) \subset \{1, \ldots, n\}\}.$$

Note that the set S is finite and the number of elements in it is

$$|S| = 2\binom{n}{k-1}.$$

Theorem 2.1: Separability criteria

If $n \geq k$, then the sample is k-separable if and only if the points $X_1^{(k)}, \ldots, X_n^{(k)}$ can be separated by some vector from the set S.

Remark 2.2. If $n \leq k$, the points are always k-separable. If $n = k$, any properly oriented hyperplane passing through $k-1$ point separates the sample points. If $n = k-1$, there is only one hyperplane passing through all the sample points, and it separates the sample points, regardless of its orientation. If $n < k-1$, then there are infinitely many hyperplanes passing through the sample points, and all of them separates the sample points, regardless of their orientation.

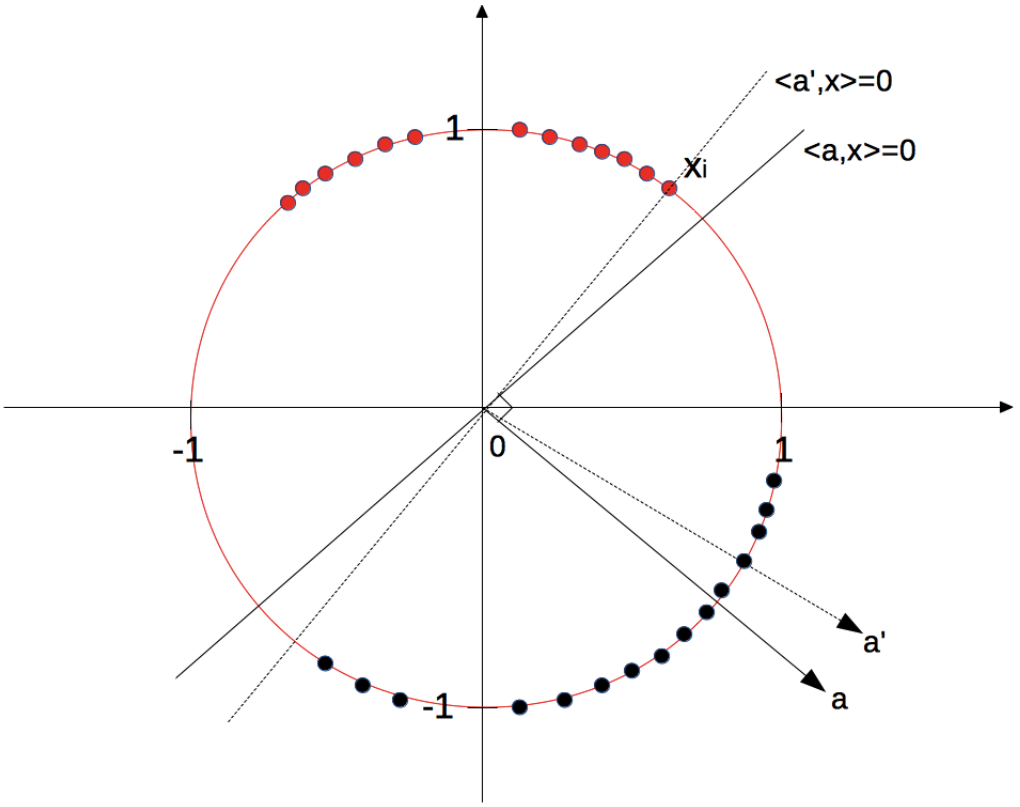

Figure 2.1: Conceptual illustration of Theorem 2.1, where $k = 2$. If sample points are k-separable by some vector a, there exists vector a' that passes through $k-1$ sample points and also separates the sample points.

2.4 Probability that sample is separable

Theorem 2.1 implies that the sample is k-separable if and only if, for some distinct $i_1, \ldots, i_{k-1}$,

$$\forall i \; Y_i \det[X_{i_1}^{(k)}, \ldots, X_{i_{k-1}}^{(k)}, X_i^{(k)}] \geq 0 \tag{2.4}$$

or

$$\forall i \; Y_i \det[X_{i_1}^{(k)}, \ldots, X_{i_{k-1}}^{(k)}, X_i^{(k)}] \leq 0. \tag{2.5}$$

Let q_{kn} be the probability that sample is k-separable. We will need the following assumption on the distribution of X:

(FR) We will say that the distribution of X is *of full rank*, if $\mathsf{P}(\langle \theta, X \rangle = 0) = 0$, for all $\theta \neq 0$.

Theorem 2.2: Probability of point separation

If (FR) holds and $n \geq k$, then with some $q < 1$ that does not depend on n or on k,

$$q_{kn} \leq 2\binom{n}{k-1} q^{n-k+1}.$$

Theorem 2.2 gives an upper bound of the probability that sample points are k-separable. It may not be the lowest upper bound but it gives a good understanding about what sequence (k_n) should be chosen for projection of X so that we could expect estimate (2.1) to be consistent. The following Corollary summarizes this.

Corollary 2.1: Existence/uniqueness of maximum-likelihood estimate

If $k_n/n \to 0$, then $q_{k_n n} \to 0$.

For example, if we take $k_n = \lfloor \sqrt{n} \rfloor$, the probability that the logistic estimate exists and is unique is close to 1, for n large enough.

2.5 Discussion

The results presented in this Chapter can be directly used for the theoretical investigations of the properties of logistic classifier in abstract Hilbert spaces, such as consistency in Chapter 3, for example. When working with functional data, an infinitely-dimensional parameter vector cannot be uniquely estimated only from the finite number of observations. Therefore, a common practice is to 'cut' the parameter vector θ after, say, the kth coordinate, and set the remaining coordinates to zero. However, this approach is avoided in literature, mainly due to the fact that the quantitative rule of selecting such k in a way that the resulting estimate would have desirable theoretical properties has not been established yet. Theorem 2.2 contributes to the understanding of what a good rule for selecting k could possibly be. Corollary 2.1 tells us that at least $k_n/n \to 0$ should be required so that we could expect a maximum quasi-likelihood estimate in logistic regression models in abstract Hilbert spaces to have desirable theoretical properties.

Chapter 3

Consistency of Logistic Classifier in Abstract Hilbert Spaces

We study the asymptotic behavior of the logistic classifier in an abstract Hilbert space and require realistic conditions on the distribution of data for its consistency. The number k_n of estimated parameters via maximum quasi-likelihood is allowed to diverge so that $k_n/n \to 0$ and $n\tau_{k_n}^4 \to \infty$, where n is the number of observations and τ_{k_n} is the variance of the last principal component of data used for estimation. This is the only result on the consistency of the logistic classifier we know so far when the data are assumed to come from a Hilbert space.

3.1 Introduction to the problem

Most of classifiers assign an observation to the class with the largest estimated posterior probability. Consistency of such a classifier is then implied by the consistency of the estimate of that probability. If it depends on a finite number of unknown parameters, as in the logistic model in $\mathbb{R}^k$, then it suffices to consistently estimate all the parameters. For example, in the $\mathbb{R}^k$ case the logistic classifier has been proved to be consistent, strongly consistent (see, e.g. [29]) and even uniformly consistent [30].

The situation becomes more complicated if conditional probability is modelled by the infinite number of parameters, as in the logistic model in an infinite-dimensional Hilbert space E. In this case we are given independent observa-

tions $(X_1,Y_1),\ldots,(X_n,Y_n)$ of (X,Y), where X is E-valued random variable and $Y \in \{-1,1\}$ is its associated class label. Usually, then the following 3-step procedure is used: (1) some orthonormal basis in E is chosen and the observations are replaced by their coefficients in that basis (a finite number, say, l of coefficients are retained), (2) the principal component analysis of the obtained $n \times l$ array of data is performed and the first k principal components are retained, (3) the usual logistic regression on the new $n \times (k+1)$ array of data is performed. From the mathematical point of view this means that we replace the original observations by their orthogonal projections in some k-dimensional subspace $E_k \subset E$ and find the estimate $\hat{\theta}_{kn}$ of the unknown parameter $\theta_0 \in E$, which maximizes the quasi-likelihood over all $\theta \in E_k$. Of course, if we want to analyze asymptotic properties of such an estimator (and of the corresponding classifier, based on that estimator), we should also assume that k depends on n, that is, the final estimator to be analyzed is $\hat{\theta}_{k_n n}$ for some sequence $k_n \to \infty$.

Note that if E_k is obtained by the procedure described above, then it is a random subspace of E (it depends on data). This makes the analysis of $\hat{\theta}_{k_n n}$ rather complicated. Therefore in this Chapter we will analyze the simpler case where E_k are non random. Formally, this means that we omit the step of principal component analysis. This approach (call it naïve) is also known in the literature, but in some cases is not recommended for practical use. For example, [22] argued that the naïve approach in the context of functional data introduces multicollinearity (strong dependence among predictors) which in turn causes inaccurate parameter estimates and increases their variance. However, the asymptotic results in the case where E_k are non random in some situations are good, as we show later. Moreover, they show what can be expected in the general case because some required assumptions are likely to remain also in the general setting.

In this Chapter we establish the consistency of the logistic classifier under the two sets of conditions. The first set consists of three conditions on the distribution of X that are rather simple and, nevertheless, sufficiently general. All three conditions are satisfied if X has a normal distribution in Hilbert space with zero mean and posi-

tive definite covariance form. The second set of conditions bound the growth rate of k_n: we require that $k_n/n \to 0$ and $n\tau_{k_n}^4 \to \infty$, where $\tau_k = \min_{\theta \in E_k, \|\theta\|=1} C(\theta,\theta)$ and C is the moment form of X defined by (3.1). As we later discuss, τ_k can be interpreted as the variance of the kth theoretical principal component. The first condition requires k to be asymptotically less than n diverges which is almost necessary. The second condition suggests that the variance of the last theoretical principal component tends to 0 slower than $1/n^{-1/4}$, as $n \to \infty$. However, this condition can be relaxed, as our simulation study shows.

In the literature, there are limited attempts to study asymptotic behavior of logistic estimate when dimensionality k_n of data used for estimation diverges together with the sample size. For example, [31], [32] and [33] studied related but slightly different problems, that is, models that include some kind of penalty on parameter vector, such as lasso. At first look it could seem that a very close attempt to solve the described problem was the one of [34], where asymptotic normality of the parameter estimate under mild conditions is proved. However, the fundamental difference between their work and ours is that they did not consider covariates X to be random, while we do. In principle, the results for the model with nonrandom data can be applied also to the case where the data are random, provided that the assumptions used for nonrandom data are satisfied for each realization of random data. However, we cannot apply their result to solve our problem because one of their assumptions translates as $\inf_k \tau_k > 0$ which is not the case if data come from a Hilbert space and follow normal distribution in Hilbert space. In such situation we can always select basis system $\{e_j\}$ such that the coordinates of X are uncorrelated. Then $\sum_{j=1}^{\infty} C(e_j,e_j) = \sum_{j=1}^{\infty} \mathsf{E}X_j^2 = \mathsf{E}\|X\|^2 < \infty$. If E_k are such as required in Chapter 2, then $\tau_k = C(e_k,e_k)$ and thus $\inf_k \tau_k = 0$.

The results nearest to ours are achieved in [10]. In the paper, the authors studied generalized linear models with no penalty and established asymptotic normality for a properly scaled distance between the estimated and the true parameters. However, they assume (see assumption (M1)) that if $\mathrm{Var}^X Y = \sigma^2(\mathsf{E}^X Y)$ (where E^X, Var^X denote the conditional mean and variance, given X) then the function σ is bounded

away from 0: $\sigma^2(\mu) \geq \delta > 0$ for all μ. This is not the case for logistic regression model, where $\sigma^2(\mu) = \mu(1-\mu)$. This means that the results in [10] cannot be applied to prove consistency of logistic classifier as considered in this Chapter. Moreover, [10] approximated infinite-dimensional model by a finite-dimensional one, that is they assumed that the distribution of Y depends on the projection of θ_0 onto some subspace E_k rather than on full $\theta_0 \in E$, and assumed that the error of such an approximation tends to 0. However, we could not find any proof of the latter rather complicated statement. No such approximation is involved in this Chapter.

This Chapter is organized as follows. In Section 3.2 we describe the statistical problem considered in this Chapter, explicitly state the assumptions, give some discussion on them, and state our main result. In Section 3.3 we provide a simulation study and we end this Chapter with a brief discussion in Section 3.4.

3.2 Consistency

Let E be a separable infinite-dimensional Hilbert space with the inner product $\langle \cdot, \cdot \rangle$. Let X be a random vector from E, and Y a random variable, gaining values -1 and 1, with conditional probabilities (w.r.t. X) $1 - p_{\theta_0}(X)$ and $p_{\theta_0}(X)$, respectively. Here $\theta_0 \in E$ is an unknown parameter and

$$p_\theta(x) = \frac{1}{1 + \mathrm{e}^{-\langle \theta, x \rangle}} \quad \text{for } \theta, x \in E.$$

We consider the following statistical task. We want to estimate the unknown conditional probability p_{θ_0}, given the sample $(X_1, Y_1), \ldots, (X_n, Y_n)$ from the distribution of (X, Y). The quality of the estimate $\hat{p}$ is assessed by the risk $\mathsf{E}|\hat{p}(X) - p_{\theta_0}(X)|$. If the risk tends to 0, the estimate $\hat{p}$ is called *consistent*. It is well known that if $\hat{p}$ is consistent, then the empirical classifier, which assigns x to the class 1 whenever $\hat{p}(x) > 1/2$, is also consistent (see, e.g., [7]). Here we will consider the same logistic estimate (2.2) as in Chapter 2, where we suppose $\hat{\theta}_{kn} = 0$ if the minimum is not attained or is not unique.

Recall that any family of random variables (Z_s) is called *uniformly integrable*,

if

$$\sup_s \mathsf{E}|Z_s|1_{\{|Z_s|>c\}} \xrightarrow[c\to\infty]{} 0.$$

The consistency of the logistic estimate will be proved under the following assumptions on the distribution of X:

(FR) The distribution of X is of full rank.

(M) $\mathsf{E}\|X\|^4 < \infty$.

(UI) The family of random variables $(\langle\theta,X\rangle^2/\mathsf{E}\langle\theta,X\rangle^2 \mid \|\theta\| = 1)$ is uniformly integrable.

Assumption (M) implies that the mean of X and the second moment form of X are correctly defined. The mean is the only such vector $\mathsf{E}X$ from E that $\langle\theta,\mathsf{E}X\rangle = \mathsf{E}\langle\theta,X\rangle$ for all $\theta \in E$. The second moment form is defined by

$$C(\theta_1,\theta_2) = \mathsf{E}\langle\theta_1,X\rangle\langle\theta_2,X\rangle. \tag{3.1}$$

If $\mathsf{E}X = 0$ it is called the *covariance form*. For example, if $E = L^2([0;1])$, then

$$C(\theta_1,\theta_2) = \mathsf{E}\int_0^1 \theta_1(s)X(s)\mathrm{d}s\int_0^1 \theta_2(t)X(t)\mathrm{d}t = \int_0^1 \mathrm{d}s\int_0^1 \theta_1(s)\theta_2(t)\tilde{C}(s,t)\mathrm{d}t,$$

where $\tilde{C}(s,t) = \mathsf{E}X(s)X(t)$ is a covariance function of the process X. If $E = \ell^2$ and x_i denote the coordinates of $x \in \ell^2$, then

$$C(\theta_1,\theta_2) = \mathsf{E}\sum_{i=1}^{\infty}\theta_{1i}X_i\sum_{j=1}^{\infty}\theta_{2j}X_j = \sum_{i,j}\theta_{1i}\theta_{2j}c_{ij},$$

where (c_{ij}) is a covariance matrix of the random vector X. Since E can be any abstract Hilbert space, we will work with the general notation $C(\theta_1,\theta_2)$.

The second moment form is a continuous bilinear form on E. Moreover, it is symmetric and positive semi-definite, that is, for all θ,

$$C(\theta,\theta) = \mathsf{E}\langle\theta,X\rangle^2 \geq 0.$$

Obviously, $C(\theta,\theta)=0$ if and only if $\mathsf{P}(\langle\theta,X\rangle=0)=1$. This implies that $C(\theta,\theta)>0$ if and only if $\mathsf{P}(\langle\theta,X\rangle=0)<1$. Recall that assumption (FR) is $\mathsf{P}(\langle\theta,X\rangle=0)=0$. Hence assumption (FR) is slightly stronger than requirement of C being positive definite.

The conditions we require are realistic and hold for a variety of real-life settings. For example, all three assumptions hold, if X is a normally distributed random vector with zero mean and positive definite covariance form. Indeed, then $\mathsf{E}\|X\|^s<\infty$, for all s, and

$$\sup_{\|\theta\|=1}\mathsf{E}\frac{\langle\theta,X\rangle^2}{\mathsf{E}\langle\theta,X\rangle^2}1_{\left\{\frac{\langle\theta,X\rangle^2}{\mathsf{E}\langle\theta,X\rangle^2}>c\right\}}=\mathsf{E}Z^2 1_{\{Z^2>c\}}\xrightarrow[c\to\infty]{}0.$$

Here Z is a random variable with the standard normal distribution.

Denote

$$\tau_k=\min_{\substack{\theta\in E_k\\ \|\theta\|=1}}C(\theta,\theta). \tag{3.2}$$

Here C is the moment form of X, defined by (3.1). For example, if $E=\ell^2$, E_k are as defined in Chapter 2, $\mathsf{E}X=0$, the coordinates of X are uncorrelated and the variances of them decrease, then τ_k is the variance of the kth coordinate. In other words, τ_k is the variance of the kth theoretical principal component.

Our main result is the following Theorem.

Theorem 3.1: Consistency of logistic estimate (no intercept)

Suppose that assumptions (FR), (M) and (UI) hold. Moreover, suppose

$$k_n\to\infty,\quad \frac{k_n}{n}\to 0\quad\text{and}\quad n\tau_{k_n}^4\to\infty.$$

Then the logistic estimate is consistent.

Note that the condition $n\tau_{k_n}^4\to\infty$ requires that the data are such that the variance of the last principal component tends to 0 slower than $1/n^{-1/4}$, as $n\to\infty$. This in turn suggests that the data need to be such that it cannot be sufficiently explained

only by a few principal components. For example, if data are such that 99% of its first 3 dimensions are explained by the first 2 principal components and adding every other dimension does not influence the cumulative variance explained by these first 2 principal components, then the theoretical results will not be valid for such data. Of course, such example is only an interpretation of the theoretical asymptotic result.

In statistics, the logistic model with an intercept is usually preferred over the one without it because useful model information might be incorporated in the intercept term. Theorem 3.1 implies the analogous result on the logistic estimate, when the model with an intercept is considered, that is, when the conditional probability that $Y = 1$, given $X = x$, is defined by

$$p_{\alpha,\theta}(x) = \frac{1}{1+\mathrm{e}^{-\alpha-\langle\theta,x\rangle}} \quad \text{for } \alpha \in \mathbb{R} \text{ and } \theta, x \in E. \tag{3.3}$$

In this case, the assumption (FR) should be changed to

(FR') $\mathsf{P}(\langle\theta,X\rangle = \alpha) = 0$ for all $\theta \neq 0$ and $\alpha \in \mathbb{R}$.

We call $p_{\hat{\alpha},\hat{\theta}}$ the logistic estimate of (3.3), if

$$(\hat{\alpha},\hat{\theta}) = \arg\min_{(\alpha,\theta)\in\mathbb{R}\times E_{k_n}} M_n(\alpha,\theta), \tag{3.4}$$

where

$$M_n(\alpha,\theta) = \overline{m_{\alpha,\theta}(X,Y)}, \quad m_{\alpha,\theta}(X,Y) = \log(1+\mathrm{e}^{-Y(\alpha+\langle\theta,X\rangle)}).$$

We say that the logistic estimate is consistent, if $\mathsf{E}|p_{\hat{\alpha},\hat{\theta}}(X) - p_0(X)| \to 0$, as $n \to \infty$, where $p_0(x) = p_{\alpha_0,\theta_0}(x)$ in this case. As before, τ_k is defined by (3.2), where C is the covariance form of X. Our last result is the following Theorem.

Theorem 3.2: Consistency of logistic estimate (with intercept)

Suppose assumptions (FR'), (M) and (UI) hold, and $\mathsf{E}X = 0$. Moreover, suppose

$$k_n \to \infty, \quad \frac{k_n}{n} \to 0 \quad \text{and} \quad n\tau_{k_n}^4 \to \infty.$$

Then the logistic estimate is consistent.

3.3 Simulation study

To illustrate the established assumptions, we conducted a simulation study. We will give the two examples: one, where all assumptions hold, and another one, where the assumption $n\tau_k^4 \to \infty$ does not hold.

Example 1. Since $X_i(t) = \sum_{j=1}^{\infty} C_{ij} e_j(t)$ for any selected basis system, it is enough to generate coefficients C_{ij}. To go in line with the (UI) assumption, we will generate C_{ij} as independent and normally distributed variables with zero mean and variance $\sigma_j^2 = 1/(1.1^j)$. Then $\tau_k = \sigma_k^2$. If we want that $n\tau_k^4 = n1.1^{-4k}$ tend to ∞, we have to take $k = \lceil c \log n \rceil$ with $c < 1/(4\log 1.1) \approx 2.62$. In this example, we will take $c = 2$, so that $n\tau_k^4 \to \infty$ and all assumptions hold.

We took θ_0 with $\theta_{0i} = 1/(1.1^i)$ and calculated $p_{\theta_0}(X_i)$ up to precision $\varepsilon = 10^{-4}$. To this end we generated additional coordinates X_{ij} for $j \le l$, where l was the first index with $|\theta_{0l} X_{il}| < \varepsilon$.

We generated $300, 500, 1000, 1500$ and 2000 observations, respectively, over 100 independent runs for each setting, and each time we approximated the distance

$$d(\hat{p}, p_0) = f(\hat{\theta}, \theta_0),$$

where

$$f(\theta, \theta_0) = \mathsf{E}|1/(1+\mathrm{e}^{-U_1}) - 1/(1+\mathrm{e}^{-U_2})|$$

with $U = (U1, U2)$ distributed according to the normal law with zero mean and

covariance matrix

$$\Sigma = \begin{pmatrix} \sum_i \theta_i^2 \sigma_i^2 & \sum_i \theta_i \theta_{0i} \sigma_i^2 \\ \sum_i \theta_i \theta_{0i} \sigma_i^2 & \sum_i \theta_{0i}^2 \sigma_i^2 \end{pmatrix}.$$

We calculated f using the Monte Carlo method. We simulated 10000 independent copies of U, which gives, as preliminary testing shows, approximate 0.01 precision for d. We also reported misclassification rate

$$\text{MCR} = \frac{1}{n} \sum_{i=1}^{n} 1_{\{\hat{y}_i \neq y_i\}},$$

where we set $\hat{y}_i = 1$, if $\hat{p}(x_i) >= 1/2$. Moreover, we reported the Bayes risk, where the probability of misclassification was calculated by

$$\mathsf{E} \min(p_0(X), 1 - p_0(X)) = \mathsf{E} \frac{1}{1 + \mathrm{e}^{|U|}}, \tag{3.5}$$

where $U \sim N(0, 1/(1.1^3 - 1))$. Again, we used Monte Carlo method to calculate (3.5). Figure 3.1 illustrates the simulated data as well as the true and estimated conditional probabilities. The x axis in plots (a)-(c) in Figure 3.1 represents the coefficient number j which stops after the kth value is generated. The y axis in plots (a)-(c) in Figure 3.1 represents the values of C_{ij}. As we can see from plots (a)-(c) the C_{ij} are distributed normally with mean 0 and their variance decreases as j increases. Plots (d)-(f) in Figure 3.1 shows the true and estimated conditional probabilities p_0 and $\hat{p}$, respectively, as functions of x. The x axis represents the observation number i and the y-axis shows the values of p_0 and $\hat{p}$ at the $x = x_i, i = 1, \ldots, n$. We can see that the true and estimated conditional probabilities are close to each other for every observation suggesting that the average difference between the two is small. This is further confirmed by $\hat{d}(\hat{p}, p_0)$ values in Table 3.1. Numerical results, averaged over 100 independent runs, are displayed in Table 3.1. As we can see from Table 3.1, the assumption $n\tau_k^4 \to \infty$ holds and $\hat{d}(\hat{p}, p_0) \to 0$, as expected.

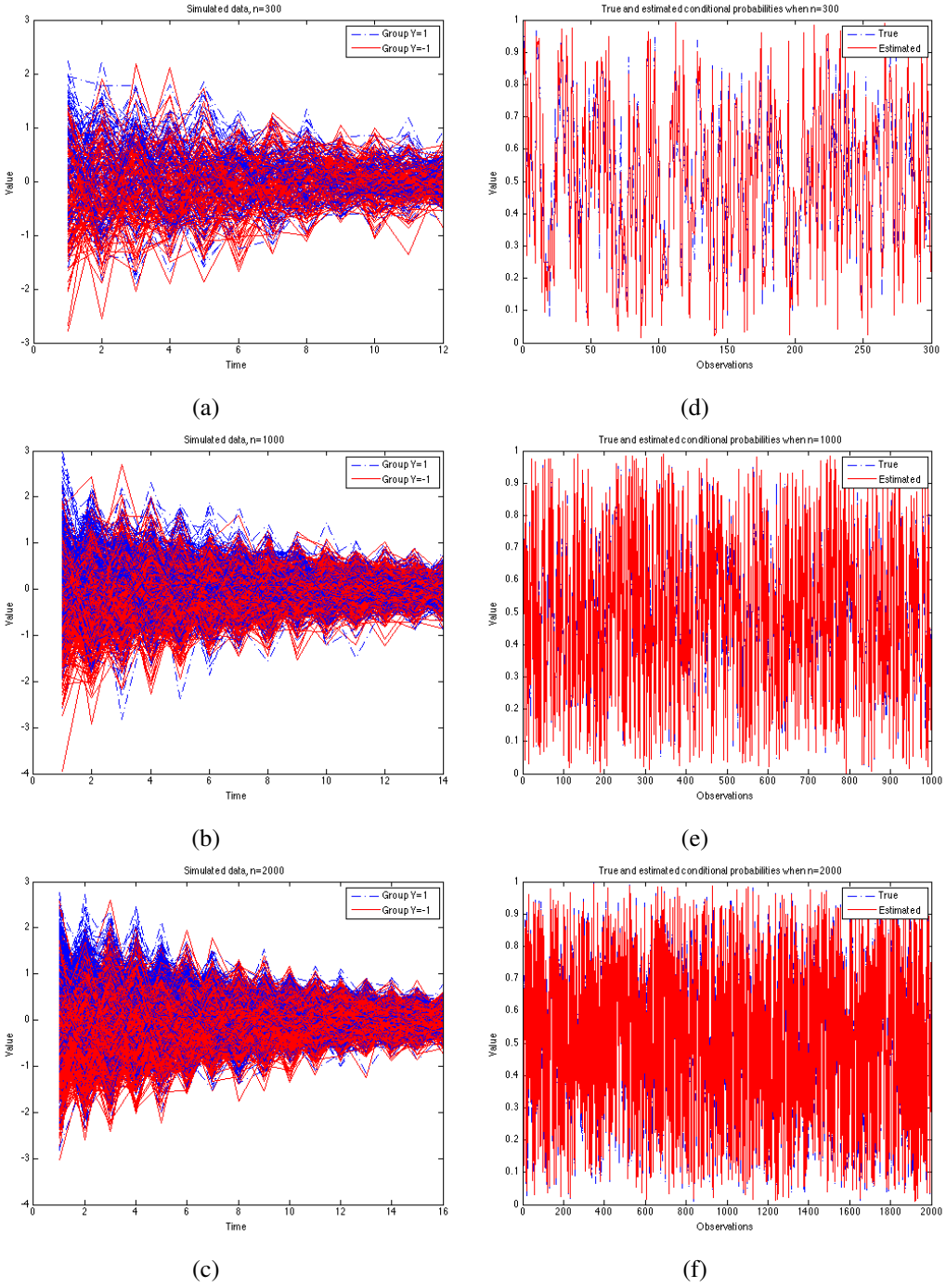

Figure 3.1: Illustration of simulated data for Example 1. (a)-(c) Simulated data for $n = 300, 1000$ and 2000, respectively. (d)-(f) True conditional probability p_0 and estimated conditional probability $\hat{p}$, evaluated for the generated observations.

Table 3.1: Numerical results for Example 1, averaged over 100 independent runs

n	300	500	1000	1500	2000
k	12	13	14	15	16
$n\tau_k^4$	3.1	3.5	4.8	4.9	4.5
$\hat{d}(\hat{p}, p_0)$ ($\pm$ sd)	0.095 ($\pm$ 0.017)	0.078 ($\pm$ 0.013)	0.061 ($\pm$ 0.008)	0.054 ($\pm$ 0.007)	0.048 ($\pm$ 0.007)
MCR (%, $\pm$ sd)	26.08 ($\pm$ 2.7)	26.35 ($\pm$1.8)	26.33 ($\pm$ 1.41)	26.76 ($\pm$ 1.15)	26.55 ($\pm$ 0.91)
Bayes (%, $\pm$ sd)	24.32 ($\pm$ 0.16)	24.32 ($\pm$ 0.16)	24.32 ($\pm$ 0.16)	24.32 ($\pm$ 0.16)	24.32 ($\pm$ 0.16)

Example 2. Let us consider the same settings as for Example 1, except that now we will take $c = 6$, so that $n\tau_k^4 \to 0$ and even $n\tau_k^2 \to 0$. Figure 3.2 illustrates the simulated data as well as the true and estimated conditional probabilities. Numerical

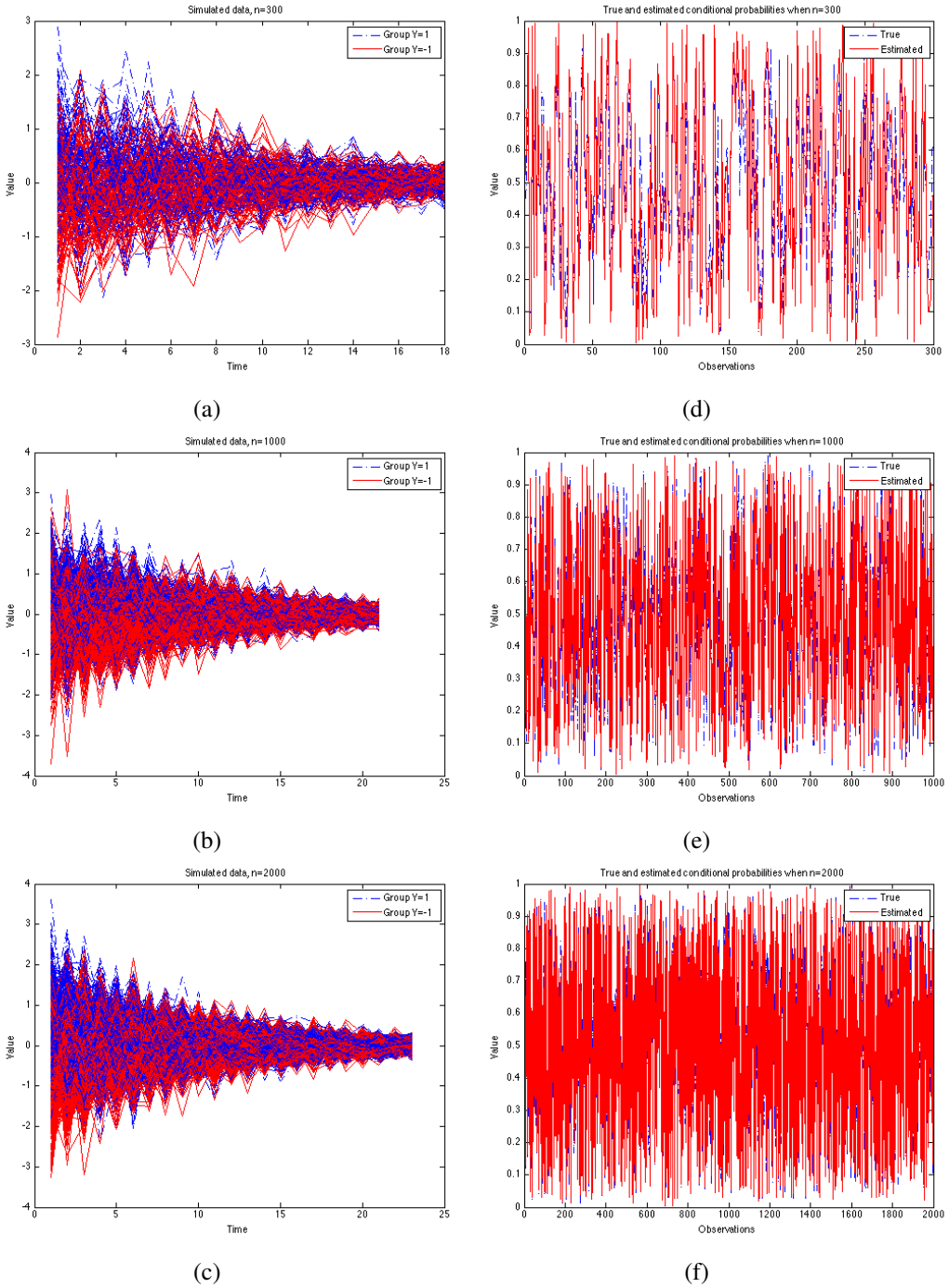

Figure 3.2: Illustration of simulated data for Example 2. (a)-(c) Simulated data for $n = 300, 1000$ and 2000, respectively. (d)-(f) True conditional probability p_0 and estimated conditional probability $\hat{p}$, evaluated for the generated observations.

results, averaged over 100 independent runs, are displayed in Table 3.2.

Table 3.2: Numerical results for Example 2, averaged over 100 independent runs

n	300	500	1000	1500	2000
k	18	19	21	22	23
$n\tau_k^2$	0.3	0.4	0.3	0.3	0.3
$n\tau_k^4$	$5*10^{-4}$	$2.5*10^{-4}$	$1.1*10^{-4}$	$7.8*10^{-5}$	$4.2*10^{-5}$
$\hat{d}(\hat{p}, p_0)$ ($\pm$ sd)	0.127 ($\pm$ 0.024)	0.102 ($\pm$ 0.015)	0.082 ($\pm$ 0.014)	0.069 ($\pm$ 0.011)	0.065 ($\pm$ 0.01)
MCR (%, $\pm$ sd)	24.56 ($\pm$ 2.31)	25.35 ($\pm$ 1.94)	26.01 ($\pm$ 1.28)	26.49 ($\pm$ 1.21)	26.55 ($\pm$ 0.94)
Bayes (%, $\pm$ sd)	24.32 ($\pm$ 0.16)	24.32 ($\pm$ 0.16)	24.32 ($\pm$ 0.16)	24.32 ($\pm$ 0.16)	24.32 ($\pm$ 0.16)

As we can see from Table 3.1, the assumption $n\tau_k^4 \to \infty$ (and even weaker assumption $n\tau_k^2 \to \infty$) is violated but $\hat{d}(\hat{p}, p_0) \to 0$. This suggests that the assumption $n\tau_k^4 \to \infty$ might be not needed to establish the consistency of logistic estimate and could be relaxed in future investigations.

3.4 Discussion

As we noted in the previous Section, assumption $n\tau_{k_n}^4 \to \infty$ does not seem to be necessary for our main result to hold. It is interesting that the analogous assumption (M3) in [10] translates into $n\tau_{k_n}^2/k_n^2 \to \infty$. However, our simulation study shows (see Example 2) that even assumption $n\tau_{k_n}^2 \to \infty$ is not necessary. At the moment it is not clear what is the true asymptotic lower bound for τ_{k_n}, and how Theorem 3.1 can be proved under assumption, weaker than $n\tau_{k_n}^2 \to \infty$.

Chapter 4

An Approach to Extending Sparsely and Irregularly Sampled Functional Data

We consider a problem of extending sparsely and irregularly sampled functional data to a common time interval. We suggest a consistent way to construct two reference functions from the data which are then used to predict missing values by using interpolation shifted vertically. Under certain assumptions, we establish the consistency of the proposed curve extension method which is then illustrated on real and simulated data.

4.1 Introduction to the problem

In Chapter 3 the consistency of logistic classifier for functional data was established. However, in Chapter 3 observing full functional data was assumed which is the case that exists only in theory. As discussed in Introduction of the thesis, in practice functional data can be observed only at some finite number of time points. Moreover, those time points as well as the number of them can differ amongst observations. This makes the application of the logistic classifier in Chapter 3 to practical classification of functional data difficult.

In this Chapter, we consider a problem of extending sparsely and irregularly sampled curves which can then be used for statistical analysis such as, for exam-

ple, classification. We are given a collection of (random) curves $X_i, i = 1,\dots,n$ observed at (random) time points $T_{i1},\dots,T_{iM_i}$, where M_i is the (random) number of time points for the ith curve. In this Chapter, M_i is allowed to be as small as 1. We call the collection of points $\{X_i(T_{i1}),\dots,X_i(T_{iM_i})\}$ a *fragment* of the ith function and we call a collection of points $\{T_{i1},\dots,T_{iM_i}\}$ *time points for the ith fragment*. We consider a task of extending curve fragments to the (random) interval $T = [T_1; T_m]$, where

$$T_1 = \min_i \{T_{i1},\dots,T_{iM_i}\}, \tag{4.1}$$

$$T_m = \max_i \{T_{i1},\dots,T_{iM_i}\}. \tag{4.2}$$

The spinal bone mineral density in individuals in [1] is the example of a practical situation, where such data type occurs (see Figure 4.1). As we can see from Figure 4.1, the measurements were taken irregularly across individuals. Moreover, not only the number of repeated measurements for an individual differs across individuals but is also very small for all individuals.

Suppose we put all time points for all fragments into the ordered vector $[T_1, T_2, \dots, T_m]^T$, where $T_j \in \{T_{i1},\dots,T_{iM_i}\}$, and $T_1 < T_2 < \cdots < T_m$, keeping only unique time points, and convert each fragment into an m-vector $X_i = [X_{i1},\dots,X_{im}]$:

$$X_{ij} = \begin{cases} X_i(T_{ij}), & \text{if } X_i(T_{ij}) \text{ exists,} \\ \varnothing, & \text{otherwise,} \end{cases}$$

where '$\varnothing$' denotes a missing value. We can then see that the resulting design matrix $X = [X_1,\dots,X_n]^T$ is extremely sparse and that the problem of extending curves to the interval T can then be also understood as a matrix completion problem. The sparsity of the design matrix X creates problems when we want to use the observed data for further statistical analysis, such as classification to different groups. Therefore, there is a need to predict, in some way, the missing values for each fragment.

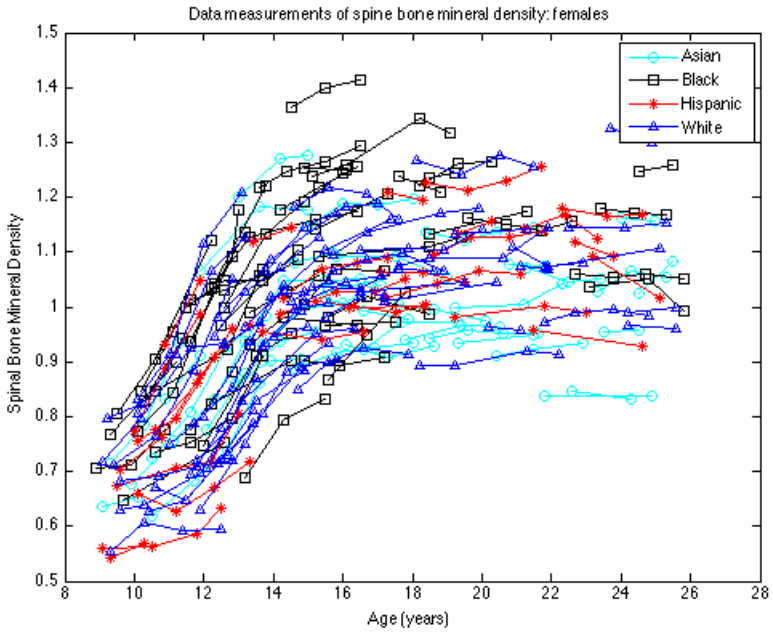

Figure 4.1: Data measurements of spinal bone mineral density for 153 females. Measurements taken for the same individual are joined by a curve. The data are described in [1] and provided by prof. James Gareth.

The popular methods for predicting missing values could be linear interpolation/extrapolation or filtering (smoothing splines). However, as was discussed in [35], these methods fail for such type of data. Even though many reasons were discussed in [35], the main reason seems to be the fact that we do not have any reference measurements for extrapolation or filtering to work at the both ends of the observed fragment.

To make the idea clearer, let us make an example. Suppose the ith fragment contributes to the design matrix as $[\varnothing, \varnothing, \varnothing, X_i(T_4), \varnothing, X_i(T_6), \varnothing, \varnothing]$. Let us call $L_i = [\varnothing, \varnothing, \varnothing]$ the *region to the left*, by $R_i = [\varnothing, \varnothing]$ the *region to the right* and $C_i = [X_i(T_4), \varnothing, X_i(T_6)]$ the *central region* of the contribution of the ith fragment to the design matrix X. Then, we can clearly see that both methods, interpolation and filtering, would work only in the central region C_i, where predicting the missing value is relatively easy, given the two endpoints. However, regions L_i and R_i do not have any references except the neighboring $X_i(T_4)$ and $X_i(T_6)$, respectively. Therefore, if we applied extrapolation to regions L_i and R_i, our prediction would

be something like a straight line $\hat{L}_i = [X_i(T_4), X_i(T_4), X_i(T_4)]$, $\hat{R}_i = [X_i(T_6), X_i(T_6)]$. If we applied filtering, we would have infinitely many possible solutions to predict entries for both L_i and R_i.

Therefore, more sophisticated methods to extend sparse functional data have been proposed. If we assume that a random curve $X_i \in L^2[0,1]$, we can express it via selected basis functions and the coefficients next to basis functions will be random. Naturally, we can always select such basis system so that the coefficients are uncorrelated. We will call such data model a *multiplicative model* to relate to the fact that the selected basis functions are multiplied by coefficients. This model was used, for example, in [36] and [37], where the curves were modeled by B-splines with random coefficients. These coefficients were assumed to follow a multivariate normal distribution and were then estimated using the EM algorithm with the constraints that reduced the number of parameters that had to be estimated. However, no asymptotic results on such extension were established. As discussed in [11] this is probably due to the fact that it is impossible to think of a consistent extension method in such a case, unless the distribution of M_i is assumed to depend on n in a way that $M_i \to \infty$ in probability or almost surely. We will call such an assumption by the *dense fragment setting* and we will refer to the situations, where such assumption is not made, as to the *sparse fragment setting*. Dense fragment setting was considered in the work of [11] and later in the works of [38], [39] and [40]. In the former, the curves were expressed through their Karhunen-Loève expansions, where the means and covariances were estimated through borrowing strength from all data points by kernel smoothers. As a result, consistency of the estimated mean, covariance functions as well as principal component scores was proved. However, as discussed in [3], the mean and covariance smoothers do not perform well in the context of sparse fragment setting, that is, when the distribution of M_i does not depend on n. A fully non-parametric approach was proposed by [3], where extensions of fragments were achieved by adjoining, to each fragment, shifted versions of other observed fragments. However, their approach forces each of the reconstructions to have exactly the shape of an observed fragment. The most

recent approach in [6] proposes the extension of sparse functional data based on the combination of Markov chains and nonparametric smoothing techniques which is specially designed for extending short fragments. For monotone functions, their approach had similar performance compared to the one of [3].

The main idea in this Chapter is to borrow the information from all observed fragments to predict the missing values for the ith curve. This is achieved by constructing consistent maximum and minimum reference functions borrowing strength from all the fragments (similarly as was done in [11] for the estimation of the mean curve) and then for each curve predicting the missing values by interpolation shifted vertically. We therefore call such an approach to inputing missing values the collaborative prediction. The key point is that we concentrate on the sparse fragment setting as in [3], even though our results are applicable also in the dense fragment setting. In the sparse fragment setting we cannot use the multiplicative model, where all coefficients are random, as in such a case the consistent extension method does not seem to exist. Therefore, in this Chapter we consider the simpler case where only one of the coefficients in the multiplicative model is random, and we prove the consistency of our proposed collaborative prediction.

There are some theoretical contributions of this Chapter as compared to other works. First, in the work of [6], the consistency is established for model parameters (transition probabilities in Markov chains) which is a traditional way of thinking about consistency. However, as discussed in [11], a more interesting and useful way is to establish consistency for extended functions $\hat{X}_{ni}$ rather than for model parameters. In their work, consistency was established for each extended curve separately. However, they worked in the dense fragment setting and their proof is not valid in the sparse fragment setting which is considered in this Chapter. In fact, in sparse fragment setting it is impossible to prove the consistency of extension for a fixed curve i because we cannot use the fact that the number of time points M_i diverges in probability or almost surely together with the sample size. Therefore, consistency can be proved only for an average $\hat{X}_{ni}$, using the fact that the sum of all M_i diverges together with the sample size. Second, in the work of [11], several

assumptions on the rate of growth of data in the local window of bandwidth h_n were used, including that of $nh_n^4 \to \infty$ which was used to establish the consistency of the estimated mean curve. As mentioned in [11], they expected that in future such assumption could be possibly reduced to the optimal one of $nh_n/\log n \to \infty$ which is exactly the assumption in this Chapter.

This Chapter is organized as follows. In Section 4.2 we describe in details the proposed method. In Section 4.3 we establish assumptions under which we prove the consistency of the proposed curve extension method. We provide a reader with a simulation study in Section 4.4, where we compare the proposed method with the method in [3] as well as study the limitations of our method. Finally, in Section 4.5 we illustrate our proposed method on the spine mineral density data mentioned earlier. The discussion is left for Section 4.6.

4.2 Proposed methods

From the theoretical point of view, the task of proposing the consistent extension of sparsely and irregularly sampled functional data in the sparse fragment setting can be shortly summarized as the following:

- $X_1, \ldots, X_n$ are independent random functions from $C[0,1]$, where $C[0,1]$ is the space of all continuous functions on $[0,1]$.
- We observe the values of the function X_i at some random time points $0 < T_{i1}, \ldots, T_{iM_i} < 1$, where M_i is the number of observed time points for the ith function.
- Based on the data, we have to construct new functions $\hat{X}_{n1}, \ldots, \hat{X}_{nn}$ such that

$$\frac{1}{n}\sum_{i=1}^{n} \|\hat{X}_{ni} - X_i\| \xrightarrow[n\to\infty]{} 0,$$

  in probability or almost surely, where $\|\cdot\|$ is the sup-norm:

$$\|X\| = \sup_{t\in[0,1]} |X(t)|. \tag{4.3}$$

Note that we used $\hat{X}_{ni}$ to underline the fact that the new functions are the estimates of the true functions that depend on (random) data. The main challenge of such a task is that as n approaches infinity, the dataset expands only vertically, that is, the number of observations increases together with n and not horizontally, that is the distribution of M_i does not depend on n in a way that $M_i \to \infty$, as n diverges.

The proposed curve extension method using interpolation shifted vertically (we will call it $\mathrm{CE}_{\mathrm{int}}$) can be described by the following 5-step procedure:

Step 1. Divide the interval $[0,1]$ into l_n equal parts and denote $h_n = 1/l_n$.

Step 2. Define the two (random) reference functions:

$$\hat{X}_n^*(t) = \max_{\substack{1\le i\le n,\ 1\le j\le M_i \\ sh_n\le T_{ij}<(s+1)h_n}} X_i(T_{ij}), \quad \text{for } sh_n \le t < (s+1)h_n, \tag{4.4}$$

$$\hat{X}_{*n}(t) = \min_{\substack{1\le i\le n,\ 1\le j\le M_i \\ sh_n\le T_{ij}<(s+1)h_n}} X_i(T_{ij}), \quad \text{for } sh_n \le t < (s+1)h_n. \tag{4.5}$$

If $t \in [sh_n;(s+1)h_n)$ but there is no T_{ij} in that interval such that $1 \le i \le n$ and $1 \le j \le M_i$, we will suppose that $\hat{X}_n^*(t) = \hat{X}_{*n}(t) = 0$.

Step 3. For every $i = 1,\dots,n$, find $T_{i,(1)} = \min_{1\le j\le M_i} T_{ij}$ and then $\hat{\alpha}_{ni} \in [0,1]$ such that

$$X_i(T_{i,(1)}) = (1-\hat{\alpha}_{ni})\hat{X}_{*n}(T_{i,(1)}) + \hat{\alpha}_{ni}\hat{X}_n^*(T_{i,(1)}).$$

Remark 4.1. In principle, any time point $\bar{T}_{ni}$ from the set $\{T_{i1},\dots,T_{iM_i}\}$ could be used instead of $T_{i,(1)}$. If $\hat{X}_n^*(\bar{T}_{ni}) - \hat{X}_{*n}(\bar{T}_{ni}) = 0$, $\hat{\alpha}_{ni}$ could be any number: in this case the value of $\hat{X}_{ni}$ does not depend on $\hat{\alpha}_{ni}$ and is equal to $\hat{X}_n^*(\bar{T}_{ni}) = \hat{X}_{*n}(\bar{T}_{ni}) = X_i(\bar{T}_{ni})$. For the sake of completeness of the argument, we will suppose that in such a case $\hat{\alpha}_{ni} = 0$.

Step 4. Finally, define

$$\hat{X}_{ni}(t) = (1-\hat{\alpha}_{ni})\hat{X}_{*n}(t) + \hat{\alpha}_{ni}\hat{X}_n^*(t). \tag{4.6}$$

Step 5. The reference functions $\hat{X}_n^*(t)$ and $\hat{X}_{*n}(t)$ as well as the resulting $\hat{X}_{ni}$ are piecewise constant. Even though this does not make much difference in theory,

for practical applications of the proposed methods smooth reference functions may be preferred. We can obtain smooth functions $\tilde{X}_{ni}(t)$ from the piecewise constant functions $\hat{X}_{ni}(t)$ in two different ways. The first way involves using the usual basis expansion, where the final (smooth) estimated functions are defined by

$$\tilde{X}_{ni}(t) = \sum_{k=1}^{K} \hat{c}_{ik} e_k(t),$$

where $\{e_k(t)\}_k$ is the chosen basis system and $\hat{c}_{ik}$ are such that they minimize the following functional:

$$\frac{1}{m} \sum_{j=1}^{m} \left(\hat{X}_{ni}(T_{ij}) - \sum_{k=1}^{K} \hat{c}_{ik} e_k(T_{ij}) \right)^2.$$

However, obtaining smooth functions in such a way requires defining the value of parameter K which complicates the theoretical analysis. Therefore, in this Chapter we use another non-parametric approach, where we define the final (smooth) estimated functions by

$$\tilde{X}_{ni}(t) = \begin{cases} (1-\alpha)\hat{X}_{ni}(sh_n) + \alpha \hat{X}_{ni}((s+1)h_n), & \text{for } sh_n \le t < (s+1)h_n, s < l_n - 1, \\ \hat{X}_{ni}, & \text{for } sh_n \le t < (s+1)h_n, s = l_n - 1, \end{cases} \tag{4.7}$$

where α is obtained by solving the following equation:

$$t = (1-\alpha)sh_n + \alpha(s+1)h_n.$$

Also note that for certain data models $\mathrm{CE}_{\mathrm{Int}}$ method slightly changes the function values at observed time points (to be more precise, $\mathrm{CE}_{\mathrm{Int}}$ changes function values at all observed time points, except which was used to calculate $\hat{\alpha}_{ni}$). Therefore, in some practical applications, where we believe the data were observed without noise or where we are not sure how to model the data, one can prefer using a modification of $\mathrm{CE}_{\mathrm{Int}}$ approach which would keep the observed function values

unchanged. To this end, instead of only one $\hat{\alpha}_{ni}$ value for the ith curve, we can calculate M_i values of $\hat{\alpha}_{ni}$, one for each $T_{ij}, j = 1,\ldots,M_i$, and each time use exactly that $\hat{\alpha}_{ni}$ for the extension of the ith curve in the interval 'to the left' with respect to that T_{ij}. For example, if T_{ij} is the minimal time point at which we observed the ith curve, that is, $T_{ij} = T_{i,(1)}$, then the interval 'to the left' w.r.t. T_{ij} is $[0; T_{i,(1)}]$. If T_{ij} is not the minimal time point, then the interval 'to the left' is $(T_{ij-1}, T_{ij}]$. If T_{ij} is the maximum time point at which we observed the ith curve, then the extension is done also in the 'interval to the right' $[T_{ij}, T_m]$. In other words, for each $T_{ij}, j = 1,\ldots,M_i$, we calculate $\hat{\alpha}_{nij}$ such that

$$X_i(T_{ij}) = (1-\hat{\alpha}_{nij})\hat{X}_{*n}(T_{ij}) + \hat{\alpha}_{nij}\hat{X}_n^*(T_{ij}).$$

Then, for each $t \in T_{\text{left}}$, where $T_{\text{left}} \subset T$ is the respective interval 'to the left',

$$\hat{X}_{ni}(t) = (1-\hat{\alpha}_{nij})\hat{X}_{*n}(t) + \hat{\alpha}_{nij}\hat{X}_n^*(t).$$

Again, if smooth function estimates are preferred, they can be obtained from $\hat{X}_{ni}(t)$ by (4.7). We will call the latter approach the local CE_{Int} approach.

4.3 Consistency

Recall that c_* is called *essential infimum* and c^* is called *essential supremum* of a random variable C, if for any $\varepsilon > 0$

$$\mathsf{P}(C \geq c_*) = 1,\ \mathsf{P}(C > c_* + \varepsilon) < 1,$$
$$\mathsf{P}(C \leq c^*) = 1,\ \mathsf{P}(C < c^* - \varepsilon) < 1.$$

In this Section we will establish the consistency of our proposed CE_{int} approach under certain assumptions:

(A) $X_i = C_i a + b$. Here a and b are unknown but fixed nonrandom functions from $C[0,1]$ and C_i are independent copies of a random variable $C \in \mathbb{R}$.

(B) The function a gains only positive values.

(C) If $c_* = \operatorname{ess\,inf} C$ and $c^* = \operatorname{ess\,sup} C$, then $0 < c_* < c^* < \infty$.

(D) M_i are independent copies of a random variable M. The random variable M gains values that are only positive integers.

(E) $(T_{ij} \mid i, j \geq 1)$ are independent copies of a random variable T. For simplicity, we assume that T is distributed uniformly in the interval $[0,1]$. However, in general, it would be enough to require that T has a density that is bounded away from zero.

(F) The families (C_i), (M_i) and (T_{ij}) are mutually independent.

Remark 4.2. Note that if we consider the multiplicative model discussed in Introduction of this Chapter, where only one of the coefficients is random, the ith curve can be expressed via selected basis functions $\{e_j\}$ in the following way:

$$X_i(t) = C_i e_1(t) + \sum_{j=2}^{\infty} c_j e_j(t),\ c_j \in \mathbb{R}.$$

In fact, assumption (A) refers to this situation, where we have a instead of e_1 and b instead of $\sum_{j=2}^{\infty} c_j e_j$.

Denote by

$$x_*(t) = c_* a(t) + b(t), \quad x^*(t) = c^* a(t) + b(t)$$

the true reference functions. We will first prove that the estimated reference functions $\hat{X}_n^*$ and $\hat{X}_{*n}$ are consistent, that is they tend to the true reference functions x^* and x_*, as $n \to \infty$. Denote by $\|\cdot\|_{L_2}$ the usual L_2-norm and recall that $\|\cdot\|$ denotes the sup-norm, that is

$$\|X\|_{L_2}^2 = \int_{t\in[0,1]} X^2(t)\mathrm{d}t,$$
$$\|X\| = \sup_{t\in[0,1]} |X(t)|.$$

Theorem 4.1: Consistency of reference functions

Let $\hat{X}_n^*$ and $\hat{X}_{*n}$ be the reference functions obtained by (4.4)-(4.5). Let assumptions (A)-(F) hold and let $h_n \to 0$. If furthermore

$$\frac{nh_n}{\log n} \to \infty, \tag{4.8}$$

then almost surely

$$\|\hat{X}_n^* - x^*\| \to 0 \quad \text{and} \quad \|\hat{X}_{*n} - x_*\| \to 0.$$

The main result is the following Theorem.

Theorem 4.2: Consistency of piecewise constant extended functions

Let $\hat{X}_{ni}$ be the extended functions obtained by (4.6). Let assumptions (A)-(F) hold and let $h_n \to 0$. If furthermore

$$\frac{nh_n}{\log n} \to \infty,$$

then almost surely

$$\frac{1}{n}\sum_{i=1}^{n} \|\hat{X}_{ni} - X_i\| \to 0.$$

Because L_2-norm is weaker ($\|\cdot\|_{L_2} \leq \|\cdot\|$) and all the values of all the functions do not exceed $c^*\|a\| + \|b\|$,

$$\|\hat{X}_{ni} - X_i\|_{L_2}^2 \leq \|\hat{X}_{ni} - X_i\|^2 \leq 2\left[c^*\|a\| + \|b\|\right] \|\hat{X}_{ni} - X_i\|.$$

Therefore, Theorem 4.2 implies consistency also w.r.t. the norm in L_2. Moreover, Theorem 4.2 implies also analogous results, if piecewise linear function estimates are used. Therefore, we can add the following Theorem.

Theorem 4.3: Consistency of piecewise linear extended functions

Let $\tilde{X}_{ni}$ be the piecewise linear extended functions obtained by (4.7). Let assumptions (A)-(F) hold and let $h_n \to 0$. If furthermore

$$\frac{nh_n}{\log n} \to \infty,$$

then almost surely

$$\frac{1}{n}\sum_{i=1}^{n} \|\tilde{X}_{ni} - X_i\| \to 0.$$

4.4 Simulation study

4.4.1 All assumptions hold

Here we implemented our proposed method in the case where all the assumptions needed for its consistency hold. To this end, we considered the following multiplicative model:

$$X_i(t) = C_i a(t),$$

where $a(t) = \exp(-t)$ and $C_i \sim U(0.5, 1)$. Note that for simplicity, we took $b(t) = 0$. We took $h_n = 1/\sqrt{n}$. For each curve $i = 1, \ldots, n$, we generated the $M_i \sim 2 + \text{Poiss}(\lambda)$ time points $T_{i1}, \ldots, T_{iM_i}$ at which we calculated the function values. We then passed those values to algorithms for curve extension to a common interval $[T_1; T_m]$, where T_1, T_m are defined by (4.1)-(4.2). We considered various simulation settings with $\lambda \in \{5, 10, 30\}$ and $n \in \{10, 50, 100, 300, 500\}$. As the method in [3] does not provide extensions of fragments for those time points at which there are no observations, for each (λ, n) setting, we approximated the distance

$$d(\hat{X}, X) = \exp\left(-\left[\frac{1}{n}\sum_{i=1}^{n} \|\hat{X}_{ni} - X_i\|\right]\right) \tag{4.9}$$

by

$$\hat{d}(\hat{X},X) = \exp\left(-\left[\frac{1}{n}\sum_{i=1}^{n}\max_{t\in\{T_1,\ldots,T_m\}}\left|\hat{X}_{ni}(t)-X_i(t)\right|\right]\right)*100\ (\%). \tag{4.10}$$

Note that we took exponential here so that numerical results would nicely lie between 0 and 100, where 0 indicates that curves were extended extremely poorly and 100 means that they were extended perfectly. We compared the proposed method with the curve extension approach proposed by Delaigle and Hall in [3] for which we will use letters DH. For the latter, we used MATLAB code provided by prof. Delaigle. This included extending curves with the nearest-neighbor method described in [3], where gaps were filled by copying the mean curve estimated by the method of [11]. The results are in Table 4.1.

Remark 4.3. Note that our proposed method works even if some realization of M_i is equal to 1. Therefore we could have taken $M_i = 1 + \text{Poiss}(\lambda)$ instead. However, here we took $M_i = 2 + \text{Poiss}(\lambda)$ because the method of [3] did not work when some realization of M_i was equal to 1.

Table 4.1: The values of $\hat{d}(\hat{X},X)$ ($\pm$sd) for different methods, averaged over 1000 independent runs. Here $\hat{d}_{\text{Int}}(\hat{X},X)$ and $\hat{d}_{\text{DH}}(\hat{X},X)$ denote the distance (4.10) where $\hat{X}_n$ are obtained by using proposed method or the method of [3], respectively.

n	10	50	100	300	500
			$\lambda = 5$		
$\hat{d}_{\text{Int}}(\hat{X},X)$ ($\pm$sd)	85.34 ($\pm$1.66)	92.18($\pm$0.74)	94.16 ($\pm$0.49)	96.28 ($\pm$0.28)	97.06($\pm$0.15)
$\hat{d}_{\text{DH}}(\hat{X},X)$ ($\pm$sd)	92.82($\pm$0.89)	95.58($\pm$0.24)	96.61($\pm$0.13)	97.72($\pm$0.03)	98.1($\pm$0.02)
			$\lambda = 10$		
$\hat{d}_{\text{Int}}(\hat{X},X)$ ($\pm$sd)	84.34 ($\pm$1.9)	92.24 ($\pm$0.62)	94.36($\pm$0.4)	96.5($\pm$0.21)	97.19($\pm$0.16)
$\hat{d}_{\text{DH}}(\hat{X},X)$ ($\pm$sd)	92.8 ($\pm$0.74)	95.56 ($\pm$0.19)	96.67($\pm$0.1)	97.88($\pm$0.03)	98.31($\pm$0.02)
			$\lambda = 30$		
$\hat{d}_{\text{Int}}(\hat{X},X)$ ($\pm$sd)	82.52($\pm$1.47)	92.05($\pm$0.67)	94.38($\pm$0.4)	96.61($\pm$0.15)	97.41($\pm$0.07)
$\hat{d}_{\text{DH}}(\hat{X},X)$ ($\pm$sd)	92.85($\pm$0.47)	95.65($\pm$0.13)	96.72($\pm$0.07)	97.93($\pm$0.02)	98.36($\pm$0.01)

As we can see from Table 4.1, for all (λ,n) settings $\hat{d}_{\text{Int}}$ tends to 100 as n increases which suggests that the proposed method is consistent for all (λ,n) settings. Moreover, for all settings the proposed method is comparable to that of [3], even

though the method of [3] works a bit better for these data. This might be because of the small variance of coefficients C_i in which case all C_i are close to each other across observations $i = 1, \ldots, n$ which in turn causes the potential errors from using nearest neighbor method in [3] being small.

4.4.2 Other than strictly positive functions

To study the limitations of the proposed method, we tested them in the case where function a gains not necessary positive values. To this end, we generated the data from the following model:

$$X_i(t) = C_i a(t),$$

where $a(t) = 2t - 1$ and $C_i \sim U(0.5, 1)$. Again, we took $b(t) = 0$, for simplicity. Here we used distance (4.9). The other settings were left as before. The results are presented in Table 4.2. As we can see from Table 4.2, $\mathrm{CE}_{\mathrm{Int}}$ method seems to have a poor convergence rate for all the settings. More optimistic results are seen for $\mathrm{CE}_{\mathrm{Int}}$ approach for settings with $\lambda = 30$. This suggests that the assumption of a gaining positive values may be indeed needed.

Table 4.2: The values of $\hat{d}_{\mathrm{Int}}(\hat{X}, X)$ ($\pm$sd), averaged over 1000 independent runs. Here $\hat{d}_{\mathrm{Int}}(\hat{X}, X)$ denotes the distance (4.10) where $\hat{X}_n$ are obtained by using the proposed method.

λ/n	10	50	100	300	500
5	60.04($\pm$3.34)	66.34($\pm$2.13)	67.42($\pm$1.61)	68.54($\pm$0.96)	68.83($\pm$0.78)
10	61.29($\pm$2.5)	69.82($\pm$1.52)	71.31($\pm$1.15)	72.7($\pm$0.73)	73.11($\pm$0.58)
30	60.18($\pm$2.19)	71.15($\pm$1.31)	73.24($\pm$1)	75.17($\pm$0.6)	75.74($\pm$0.49)

4.4.3 Other multiplicative models

To study the extensions beyond the multiplicative model, where only one coefficient is random, we generated the data from the following model:

$$X_i(t) = C_i a(t) + D_i b(t),$$

where $C_i, D_i \sim U(0.5, 1), a(t) = \exp(-t), b(t) = \cos(t)$. Here we used distance (4.9). The other settings were left as before. The results are presented in Table

4.3. Again, as we can see from Table 4.3, CE_{Int} method does not seem to perform consistently also in this case, except for the settings with $\lambda = 30$. This reflects the limitations of the proposed approach.

Table 4.3: The values of $\hat{d}_{\text{Int}}(\hat{X}, X)$ ($\pm$sd), averaged over 1000 independent runs. Here $\hat{d}_{\text{Int}}(\hat{X}, X)$ denotes the distance (4.10) where $\hat{X}_n$ are obtained by using the proposed method.

λ/n	10	50	100	300	500
5	76.9($\pm$2.81)	83.19($\pm$2.05)	84.31($\pm$1.8)	85.34($\pm$1.34)	85.6($\pm$1.26)
10	78.34($\pm$2.05)	87.31($\pm$1.2)	88.8($\pm$1.07)	90.12($\pm$0.87)	90.48($\pm$0.85)
30	77.75($\pm$1.47)	89.44($\pm$0.79)	91.73($\pm$0.58)	93.65($\pm$0.41)	94.21($\pm$0.38)

4.5 Real data example

We tested CE_{int} approach on spinal bone mineral density data mentioned in Introduction and compared with the approach of [3] which we will label as DH. There are four groups in the dataset: Asian females, Black females, Hispanic females and White females. We considered each group as a separate dataset and performed extension to each dataset independently. This involved constructing four different grids, each of which contained unique sorted time points $t_{ij}, i = 1,\ldots,n, j = 1,\ldots,m_i$, separately for each dataset. Here, unlike in the simulation study, we were interested to measure the distance between the true and predicted function values not only at the ends of each interval but also at each grid point.

Since the full ground true data is not available, we performed leave-one-out cross-validation (LOOCV) to access the performance of the methods. To this end, for each fragment i, each time we left out its value at one of the time points (say, T_{ij}) and performed the extension of that fragment without using its value at time point T_{ij}. We then calculated the distance between the predicted function value at that point $\tilde{X}_i(T_{ij})$ with the true value $X_i(T_{ij})$:

$$\exp(-|\tilde{X}_{ni}(T_{ij}) - X_i(T_{ij})|).$$

We repeated this procedure for all time points T_{ij} for the ith fragment and calculated

Table 4.4: Average distance (4.11) ($\pm$ std) calculated using local CE_{int} approach and DH approach. For each of the four datasets, the average is taken over all fragments that were considered for extension in that dataset.

Method	Asian	Black	Hispanic	White
Local CE_{int}	96.21 ± 2.86	95.61 ± 4.98	96.37 ± 2.24	97.53 ± 1.72
DH	97.62 ± 1.65	92.44 ± 18.45	91.88 ± 22.98	94.64 ± 16.55

the final distance measure for that fragment

$$\hat{d}(\tilde{X}, X) = \max_j \exp(-|\tilde{X}_{ni}(T_{ij}) - X_i(T_{ij})|). \tag{4.11}$$

Note that $\tilde{X}_{ni}$ notation is used here because we used local CE_{Int} approach with piecewise linear estimated functions.

Remark 4.4. Since the DH approach did not work for fragments that were observed at only 1 time point, we performed LOOCV only for those fragments which were observed at more than 2 time points.

For our approach, for each point to be predicted we used the best-case-scenario value for h_n. To this end, for each point to be predicted we generated 100 values for h_n, equally spaced in the interval $[0.05, 0.5]$, and selected that h_n value which resulted in the highest distance value $\hat{d}(\tilde{X}, X)$, where $\hat{d}(\tilde{X}, X)$ was calculated considering the training set of all observed points except for which the extension was done. This resulted in the following average h_n values: $h_n = 0.0933$ for Asian females, $h_n = 0.1111$ for Black females, $h_n = 0.1364$ for Hispanic females and $h_n = 0.1608$ for White females. Numerical results are presented in Table 4.4, while the visual performance is displayed in Figure 4.2, where extension for all datasets was performed using all data points and reported average h_n values for each dataset.

As we can see from Table 4.4, the accuracy of the extensions of Asian female fragments is comparable for both local CE_{Int} method and the DH method. However, the extensions of Black, Hispanic and White female fragments have a better accuracy when using local CE_{Int} approach. DH method for these data had very wide standard deviation which suggests that the prediction at some time points in LOOCV process deviated highly from the the true values at those time points. In

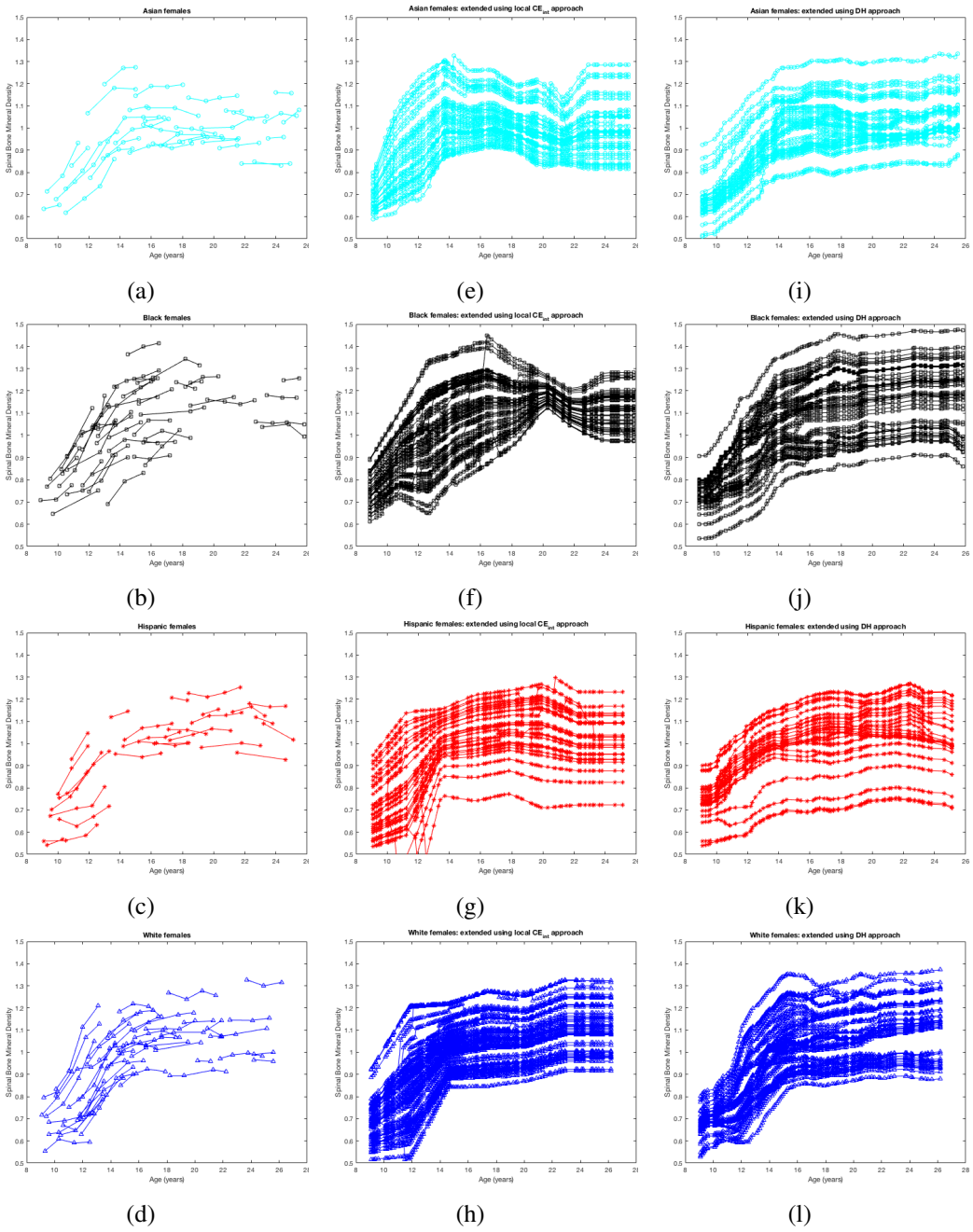

Figure 4.2: Data measurements of spinal bone mineral density for 153 females: (a)-(d) observed data measurements for Asian, Black, Hispanic and White females, respectively; (e)-(h) extended data measurements using local CE_{int} approach for Asian, Black, Hispanic and White females, respectively; (i)-(l) extended data measurements using DH approach for Asian, Black, Hispanic and White females, respectively.

fact, those points were the first or the last time points in the respective grids. For some reason in those cases the DH method makes very poor predictions which we think may be a programming error. On the other hand, even though local $\mathrm{CE}_{\mathrm{Int}}$ approach works better, we can also see its limitations. For example, plot (f) in Figure 4.2 reflects the strong dependence of local $\mathrm{CE}_{\mathrm{Int}}$ approach on the data, where there is somewhat unnatural pattern around age 19-22 for Black females caused mainly due to the lack of data in that time interval. It seems that some conditions required for consistency is not satisfied for Black female data such as, perhaps, the distribution of time points T_{ij} at which we observe the curves are not distributed uniformly in the time grid T. In fact, there seem to be much less time points between age 19 and 22 when compared to other intervals which suggests that the underlying distribution of T_{ij} for these data might have different weights for different time intervals.

4.6 Discussion

We have proposed a consistent method for the estimation of the reference functions using which curve extension can be performed by using interpolation shifted vertically. As simulation study revealed, despite being quite simple, the proposed method seems to work well in certain situations. We believe that the reason for the good performance of the proposed method is the quite narrow multiplicative model that we considered, where only one random coefficient is involved. This is also the main limitation of this Chapter. An interesting task would be to investigate how consistent extension can be achieved also in more general cases such as multiplicative models with at least two random coefficients. Possibly, the expansion of curves via principal components could be used, similarly as was done in [11]. However, the theoretical contributions in [11] are valid only in dense fragment setting. It would be therefore interesting to extend the ideas in [11] to fit the sparse fragment setting. We leave this for future investigations.

Chapter 5

Conclusions and Future Work

Chapter 2 considered the problem of point separation in logistic regression. Even though the Chapter explicitly considers functional observations, it is easy to see that the results in this Chapter are valid also for data from $\mathbb{R}^k$. One of the two contributions of this Chapter involved proving that a separating hyperplane of data points can be found from a finite set of candidate hyperplanes, where candidate hyperplanes were those passing through $k-1$ projected sample point. Therefore, this result can be interpreted as a new general procedure of determining whether or not sample points are separable (and thus whether or not maximum-likelihood estimate of the parameter vector exists and is unique) in practice.

Future work in this research direction could therefore be comparing the proposed procedure with other state-of-the-art procedures of detecting sample point separation for both functional and vectorial observations. For this matter, the theoretical results should be first implemented in practice. It is truth that Theorem 2.1 can be easily implemented by using *brute force*. However, such implementation is not optimal as the computational time increases dramatically only with an incremental increase of the sample size because the calculations of the binomial coefficient are involved. For example, if $n = 100$ and $k = 10$, there are approximately $17 * 10^{12}$ potential hyperplanes to consider. Therefore, some more theoretical work has to be done first that further reduces the set of potential hyperplanes. It is likely, that such theoretical work would override the work done in this thesis.

Chapter 3 involved determining assumptions on the distribution of function X

as well as the dimension for projection k_n so that the resulting logistic classifier in functional space is consistent. The main contribution was in achieving consistency for logistic regression with Hilbert space-valued random variables. This is not the only achievement of this Chapter. As we can see from the proofs of Chapter 3, the main difference between this Chapter and the work in [10] in a special case of logistic regression is in the strategy of proving the consistency. In [10], by assuming that the distribution of Y depends not on θ_0 itself but rather on its projection on some subspace E_k, the authors approximated infinite-dimensional model by the finite-dimensional one by assuming under the assumption that the error of such approximation tends to 0 as the sample size diverges. However, this rather complicated statement was not proved in [10] and left an open question whether or not it holds true. On the contrary, no such assumption was made in Chapter 3, where a different strategy was employed based on the ideas in [2], which included several tricks based on inverse function as well as Brouwer's fixed point theorems. Another conclusion of this Chapter is that the assumption $n\tau_{k_n}^4 \to \infty$ appeared not to be necessary for consistent classification as suggested by the simulation study. Moreover, neither appeared to be necessary the weaker assumption $n\tau_k^2 \to \infty$ which was used in [10] for other models. Therefore, the true asymptotic lower bound for τ_{k_n} is still not known.

The future work in this research direction could be therefore investigating what the true asymptotic bound could really be. One could proceed with further simulation studies that consider weaker than $n\tau_k^2 \to \infty$ assumptions. However, the real question is how to prove the consistency using asymptotic results obtained in this thesis with weaker requirement than that of $n\tau_k^4 \to \infty$. A good starting point could be investigating the proof of Theorem 3.1, namely inequality (B.5). There, more precise estimates could be given to the three probabilities which may influence the ultimate result. Also, τ_k depends only on X and not the label information Y. A more precise estimate could be therefore achieved, if Y was incorporated into the definition of τ_k. For example, τ_k could be defined as $\min_{d\theta \in E_k, \|d\theta\|=1} M''(\theta_0)(d\theta, d\theta)$.

Finally, Chapter 4 considered a common situation in practice, where we ob-

serve functional data sparsely and irregularly. As discussed in this Chapter, the main difficulty and the difference from the work of [11] is that sparse fragment setting was considered, that is, the distribution of observed time points for the ith curve M_i does not depend on the number of observations n. The main limitation of this Chapter is that the consistency is proved under a somewhat narrow multiplicative data model, where only one random coefficient is involved in the basis expansion.

One way for the future work in this research direction could be therefore investigating possible ways to extend the results of this Chapter allowing more complicated models. Another way for future work could be providing the link between the results in this Chapter and the results in Chapter 3 by proving that the logistic classifier with the extended full curves is consistent. Establishing this link is more difficult than it may seem at first look. The main challenge is that extended curves based on the data imply that the subspaces E_k in Chapter 3 become random. This in turn imply that results of Chapter 3 must be extended to fit the scenario where E_k are random and some advanced probabilistic techniques must be applied in dealing with such randomness. For this reason this task was left out of the scope of this thesis.

Secondly, consistency for the proposed curve extension method $\mathrm{CE}_{\mathrm{Int}}$ was established under somewhat restrictive model $X_i = C_i a + b$ and we proposed a modification of $\mathrm{CE}_{\mathrm{Int}}$ method that recalculates α values which we named local $\mathrm{CE}_{\mathrm{Int}}$ approach. Naturally, local $\mathrm{CE}_{\mathrm{Int}}$ approach should not perform worse than the original $\mathrm{CE}_{\mathrm{Int}}$ method. On the contrary, we believe that local $\mathrm{CE}_{\mathrm{Int}}$ approach has a potential to be consistent even for less restrictive models. Therefore, future work could be extending the theory in Chapter 4 for local $\mathrm{CE}_{\mathrm{Int}}$ method under less restrictive models.

Moreover, consistency of the proposed method was established requiring that the time points T_{ij} follow uniform distribution (or any other distribution with the density that is bounded away from 0). As results on Black female data suggests, the situation, where such assumption is violated, may be quite common in practice. Therefore, another way of future work could be towards obtaining consis-

tency for the scenarios, where the distribution of T depends on n. For example, the situation of Black female data might be well represented by $T = T(n)$ such that $\mathsf{P}(T(n) \in [19,22]) \asymp 1/n$. Then, the number of time points in the interval $[19,22]$ does not diverge, as $n \to \infty$. To overcome such data arrangement scenario, one could investigate a method based on partition of interval $[0,1]$ in the intervals of different sizes. Moreover, partition could be done in a data-driven fashion. For example, if the number of time points T_{ij} is 100 and we decided to divide interval $[0,1]$ into 10 parts, the data-driven fashion would be to choose the partition so that in each interval there are exactly 10 time points T_{ij}.

Lastly, extending sparsely and irregularly sampled functional data to full curves was chosen purely for reasons to provide (even if in the future) the link to the results in Chapter 3. One could, of course, question whether such extension is at all needed for classification of functional data in practice. Some arguments favoring extension were given in [3] and [6], where they also considered extension for practical classification of functional data. However, no theoretical proofs were given to support such arguments. Therefore, extending functional data to full curves officially has not been proved to enhance the classification performance yet. For this reason it is natural that good classification performance can be expected also when it is performed based purely on the observed data, using techniques similar to [35].

Appendix A

Proofs for Chapter 2

A.1 Proof of Theorem 2.1

Proof. Fix ω and denote $x_i = X_i^{(k)}(\omega), y_i = Y_i(\omega)$ and $b_{i_1,\ldots,i_{k-1}} = Z_{i_1,\ldots,i_{k-1}}(\omega)$. Let a be some vector separating the sample points. First we will prove that there exists a vector a' that also separates the sample points and is perpendicular to at least $k-1$ vector x_i. Let $i_1,\ldots,i_l$ be all the indices that $\langle a, x_i\rangle = 0$. If $l \geq k-1$, we can simply take $a' = a$. If $l < k-1$, then $k \geq 2$ and it is enough to construct a vector a' that separates the sample points and is perpendicular to vectors $x_{i_1},\ldots,x_{i_l}$, as well as to one other vector x_i. This is because we can always repeat the same procedure until we reach $l = k-1$.

Because $l < k-1 \leq n-1 < n$, there exists at least one i that differs from $i_1,\ldots,i_l$. For any such i, $\langle a, x_i\rangle \neq 0$. Let

$$x_i = c_1 x_{i_1} + \cdots + c_l x_{i_l} + z_i,$$

with some z_i perpendicular to $x_{i_1},\ldots,x_{i_l}$. If $i \neq i_1,\ldots,i_l$, then $z_i \neq 0$ because otherwise $\langle a, x_i\rangle = 0$ and we would get a contradiction. Moreover, $\langle a, x_i\rangle = \langle a, z_i\rangle$, for all i.

Find $i \neq i_1,\ldots,i_l$ such that for all $j \neq i_1,\ldots,i_l$,

$$\frac{|\langle a, z_i\rangle|}{\|z_i\|} \leq \frac{|\langle a, z_j\rangle|}{\|z_j\|}.$$

Denote

$$\varepsilon = \frac{\langle a, z_i \rangle}{\|z_i\|^2} \quad \text{and} \quad a' = a - \varepsilon z_i.$$

Obviously, a' is perpendicular to all $x_{i_1}, \ldots, x_{i_l}$. Moreover, it is perpendicular also to x_i because

$$\langle a', x_i \rangle = \langle a, x_i \rangle - \varepsilon \langle z_i, x_i \rangle = \langle a, z_i \rangle - \varepsilon \langle z_i, z_i \rangle = 0.$$

Let $j \neq i_1, \ldots, i_l, i$. Then

$$\left| \frac{\langle a', z_j \rangle}{\|z_j\|} - \frac{\langle a, z_j \rangle}{\|z_j\|} \right| = \left| \frac{\varepsilon \langle z_i, z_j \rangle}{\|z_j\|} \right| = \frac{|\langle a, z_i \rangle|}{\|z_i\|} \frac{|\langle z_i, z_j \rangle|}{\|z_i\| \|z_j\|} \leq \frac{|\langle a, z_i \rangle|}{\|z_i\|} \leq \frac{|\langle a, z_j \rangle|}{\|z_j\|}$$

and therefore

$$\frac{\langle a, z_j \rangle - |\langle a, z_j \rangle|}{\|z_j\|} \leq \frac{\langle a', z_j \rangle}{\|z_j\|} \leq \frac{\langle a, z_j \rangle + |\langle a, z_j \rangle|}{\|z_j\|},$$

$$\langle a, z_j \rangle - |\langle a, z_j \rangle| \leq \langle a', z_j \rangle \leq \langle a, z_j \rangle + |\langle a, z_j \rangle|.$$

This implies that $\langle a', z_j \rangle$ is equal to 0, or it is of the same sign as $\langle a, z_j \rangle$. Because $\langle a, z_j \rangle = \langle a, x_j \rangle$ and $\langle a', z_j \rangle = \langle a', x_j \rangle$, all $\langle a', x_j \rangle$ are equal to 0, or are of the same sign as $\langle a, x_j \rangle$. This means that a' separates the sample points.

We proved that there exists a vector a' that separates sample points and that is perpendicular to some $x_{i_1}, \ldots, x_{i_{k-1}}$. Vector $b_{i_1 \ldots i_{k-1}}$ is also perpendicular to $x_{i_1}, \ldots, x_{i_{k-1}}$, therefore $a' = \varepsilon b_{i_1 \ldots i_{k-1}}$ with some $\varepsilon \neq 0$. Obviously, vector $a'/|\varepsilon|$ belongs to the set S and separates the sample points. □

A.2 Proof of Theorem 2.2

We begin with the following Lemma.

Lemma A.1. *Suppose (FR) holds. Then almost surely*

$$\det[X_1^{(k)}, \ldots, X_k^{(k)}] \neq 0.$$

Proof. From linear algebra we know that $\det[X_1^{(k)},\dots,X_k^{(k)}] = 0$ if and only if $X_1^{(k)},\dots,X_k^{(k)}$ are linearly dependent. We will prove that the probability of such event is 0. Let P^{k-1} denote the conditional probability w.r.t. $(X_1,\dots,X_{k-1})$, and let Z be a random vector from E_k such that

$$\det[X_1^{(k)},\dots,X_{k-1}^{(k)},x] = \langle Z,x\rangle.$$

Obviously, Z is some function of $(X_1,\dots,X_{k-1})$. Therefore,

$$\begin{aligned}\mathsf{P}^{k-1}(\det[X_1^{(k)},\dots,X_k^{(k)}] = 0) &= \mathsf{P}^{k-1}(\langle Z,X_k^{(k)}\rangle = 0)\\ &= \mathsf{P}^{k-1}(\langle Z,X_k\rangle = 0) = 1_{\{Z=0\}}.\end{aligned}$$

and

$$\begin{aligned}&\mathsf{P}(X_1^{(k)},\dots,X_k^{(k)}\text{are linearly dependent})\\ &\quad= \mathsf{P}(\det[X_1^{(k)},\dots,X_k^{(k)}] = 0) = \mathsf{E}\mathsf{P}^{k-1}(\det[X_1^{(k)},\dots,X_k^{(k)}] = 0)\\ &\quad= \mathsf{E}1_{\{Z=0\}} = \mathsf{P}(Z=0) = \mathsf{P}(X_1^{(k)},\dots,X_{k-1}^{(k)}\text{are linearly dependent})\\ &\quad\le \mathsf{P}(X_1^{(k-1)},\dots,X_{k-1}^{(k-1)}\text{are linearly dependent}).\end{aligned}$$

Now, by induction and assumption (FR),

$$\begin{aligned}\mathsf{P}(X_1^{(k)},\dots,X_k^{(k)}\text{are linearly dependent}) &\le \mathsf{P}(X_1^{(k-1)},\dots,X_{k-1}^{(k-1)}\text{are linearly dependent})\\ &\vdots\\ &\le \mathsf{P}(X_1^{(1)}\text{is linearly dependent})\\ &= \mathsf{P}(X_1^{(1)} = 0) = \mathsf{P}(\langle e_1,X_1\rangle = 0) = 0,\end{aligned}$$

where e_1 is the basis of E_1. □

We are now ready to prove Theorem 2.2.

Proof. By letters I we will denote the subsets of $\{1,\dots,n\}$ that contain $(k-1)$ elements. If $I = \{i_1,\dots,i_{k-1}\}$ with $i_1 < \dots < i_{k-1}$, let W_I and W_I' denote events,

defined by equations (2.4) and (2.5) in Chapter 2, respectively. Then

$$q_{kn} = \mathsf{P}\left(\bigcup_I W_I \cup \bigcup_I W_I'\right) \le \sum_I \mathsf{P}(W_I) + \sum_I \mathsf{P}(W_I'). \tag{A.1}$$

Let $I = \{i_1, \ldots, i_{k-1}\}$ with $i_1 < \cdots < i_{k-1}$ and let P^I denote the conditional probability w.r.t. $X_{i_1}, \ldots, X_{i_{k-1}}$. Then

$$\mathsf{P}^I(W_I) = \left[\alpha(X_{i_1}, \ldots, X_{i_{k-1}})\right]^{n-k+1} \text{ and } \mathsf{P}^I(W_I') = \left[\beta(X_{i_1}, \ldots, X_{i_{k-1}})\right]^{n-k+1},$$

where

$$\alpha(x_1, \ldots, x_{k-1}) = \mathsf{E}1_{\{\det[x_1, \ldots, x_{k-1}, X] \ge 0\}} p_{\theta_0}(X) + \mathsf{E}1_{\{\det[x_1, \ldots, x_{k-1}, X] \le 0\}} (1 - p_{\theta_0}(X)),$$
$$\beta(x_1, \ldots, x_{k-1}) = \mathsf{E}1_{\{\det[x_1, \ldots, x_{k-1}, X] \le 0\}} p_{\theta_0}(X) + \mathsf{E}1_{\{\det[x_1, \ldots, x_{k-1}, X] \ge 0\}} (1 - p_{\theta_0}(X)).$$

By Lemma A.1 we get that almost surely

$$\alpha(X_{i_1}, \ldots, X_{i_{k-1}}) \le q \quad \text{and} \quad \beta(X_{i_1}, \ldots, X_{i_{k-1}}) \le q,$$

where

$$q = \mathsf{E} \max\left(p_{\theta_0}(X), 1 - p_{\theta_0}(X)\right) < 1.$$

In other words, almost surely,

$$\mathsf{P}^I(W_I) \le q^{n-k+1}.$$

Therefore, for all I,

$$\mathsf{P}(W_I) = \mathsf{E}\mathsf{P}^I(W_I) \le q^{n-k+1}$$

and, analogously, $\mathsf{P}(W_I') \le q^{n-k+1}$. The statement of the theorem is now implied by the fact that there are exactly $\binom{n}{k-1}$ distinct sets I. □

A.3 Proof of Corollary 2.1

Proof. It is enough to prove that

$$\log \binom{n}{k_n - 1} = o(n).$$

If $k < (n-1)/2$, then

$$\frac{\binom{n}{k+1}}{\binom{n}{k}} = \frac{n!k!(n-k)!}{(k+1)!(n-k-1)!n!} = \frac{n-k}{k+1} > 1.$$

Therefore, the sequence $\binom{n}{k}$ is increasing until $k < (n-1)/2$. Fix $\varepsilon < 1/2$ and denote $l_n = \lfloor \varepsilon n \rfloor$. Then for n large enough

$$\binom{n}{k_n - 1} \leq \binom{n}{l_n}.$$

By Stirling's formula

$$\binom{n}{l_n} = \frac{\sqrt{2\pi n}\mathrm{e}^{-n}n^n}{\sqrt{2\pi l_n}\mathrm{e}^{-l_n}l_n^{l_n}\sqrt{2\pi(n-l_n)}\mathrm{e}^{-(n-l_n)}(n-l_n)^{n-l_n}}(1+o(1)).$$

Therefore,

$$\begin{aligned}
\log \binom{n}{l_n} &= O(\log n) + n\log n - l_n \log l_n - (n-l_n)\log(n-l_n) \\
&= O(\log n) - l_n \log \frac{l_n}{n} - (n-l_n)\log \frac{n-l_n}{n} \\
&= O(\log n) - l_n \log(\varepsilon + O(1/n)) - (n-l_n)\log(1-\varepsilon+O(1/n)) \\
&= O(\log n) - l_n \log \varepsilon - (n-l_n)\log(1-\varepsilon) \\
&= O(\log n) - \varepsilon n \log \varepsilon - (1-\varepsilon)n\log(1-\varepsilon)
\end{aligned}$$

and

$$\overline{\lim_{n\to\infty}} \frac{1}{n} \log \binom{n}{k_n - 1} \leq \overline{\lim_{n\to\infty}} \frac{1}{n} \log \binom{n}{l_n} = -\varepsilon \log \varepsilon - (1-\varepsilon)\log(1-\varepsilon).$$

It is enough to note that

$$-\varepsilon \log \varepsilon - (1-\varepsilon)\log(1-\varepsilon) \to 0,$$

as $\varepsilon \to 0$. □

Appendix B

Proofs for Chapter 3

B.1 Facts from probability theory

Further in this Section, $\to_p$ and $\to_d$ denote convergence in probability and convergence in distribution, respectively, while $\to$ is used for the usual convergence in $\mathbb{R}$, or convergence in norm in E. For convenience of reference we recall some well-known facts about convergence and uniform integrability of random variables.

Proposition B.1.1 (Continuous mapping theorem, see [41], Theorem 3.7)**.** *Let U_n and U be random elements of some metric space S, $\mathsf{P}(U \in C) = 1$, T another metric space, and f_n, f measurable functions from S to T. If $u_n \to u \in C$ implies $f_n(u_n) \to f(u)$, then $U_n \to_d U$ implies $f_n(U_n) \to_d f(U)$.*

Proposition B.1.2 (Subsequence criterion, see [41], Lemma 3.2)**.** *Let U_n and U be random elements of some metric space S. Then $U_n \to_p U$ if and only if each subsequence of (U_n) has a further subsequence which converges in probability to U.*

Proposition B.1.3 (see [41], Lemma 3.10)**.** *If (Z_n) is a uniformly integrable sequence of random variables, then $\sup_n \mathsf{E}|Z_n| < \infty$ and $\mathsf{P}(W_n) \to 0$ implies $\mathsf{E}Z_n 1_{W_n} \to 0$.*

Proposition B.1.4 (see [41], Lemma 3.11)**.** *If (Z_n) is a uniformly integrable sequence of random variables, then $Z_n \to_d Z$ implies $\mathsf{E}Z_n \to \mathsf{E}Z$.*

Proposition B.1.5 (Weak convergence version of Fatou's lemma, see [41], Lemma 3.11)**.** *If (Z_n) is a sequence of positive random variables, then $Z_n \to_d Z$ implies $\underline{\lim}_{n\to\infty} \mathsf{E}Z_n \geq \mathsf{E}Z$.*

B.2 The function $M(\theta)$

We begin by establishing some properties of the function $M(\theta)$. Recall that θ_0 denotes the "true" value of parameter θ.

Proposition B.2.1. *1. If $\mathsf{E}\|X\| < \infty$, then, for all θ,*

$$0 < M(\theta_0) \leq M(\theta) < \infty.$$

2. If $\mathsf{E}\|X\| < \infty$, then $\theta_n \to \theta$ implies $M(\theta_n) \to M(\theta)$.

3. If $M(\theta_n) \to M(\theta_0)$, then $\langle \theta_n, X\rangle \to_p \langle \theta_0, X\rangle$.

Proof. 1. Inequality $M(\theta) > 0$ is implied by the fact that $m_\theta(x,y) > 0$ for all x and y. Because log function is increasing,

$$\begin{aligned} M(\theta) = \mathsf{E}\log(1+\mathrm{e}^{-Y\langle\theta,X\rangle}) &\leq \mathsf{E}\log(1+\mathrm{e}^{\|\theta\|\|X\|}) \\ &\leq \mathsf{E}\log(2\mathrm{e}^{\|\theta\|\|X\|}) = \log 2 + \|\theta\|\mathsf{E}\|X\| < \infty. \end{aligned}$$

Finally, convexity of the function $-\log$ yields

$$\begin{aligned} M(\theta) - M(\theta_0) &= -\mathsf{E}\log\frac{1+\mathrm{e}^{-Y\langle\theta_0,X\rangle}}{1+\mathrm{e}^{-Y\langle\theta,X\rangle}} \geq -\log\mathsf{E}\frac{1+\mathrm{e}^{-Y\langle\theta_0,X\rangle}}{1+\mathrm{e}^{-Y\langle\theta,X\rangle}} \\ &= -\log\mathsf{E}\left(\frac{1+\mathrm{e}^{\langle\theta_0,X\rangle}}{1+\mathrm{e}^{\langle\theta,X\rangle}}(1-p_{\theta_0}(X)) + \frac{1+\mathrm{e}^{-\langle\theta_0,X\rangle}}{1+\mathrm{e}^{-\langle\theta,X\rangle}}p_{\theta_0}(X)\right) \\ &= -\log\mathsf{E}\left(\frac{1}{1+\mathrm{e}^{\langle\theta,X\rangle}} + \frac{1}{1+\mathrm{e}^{-\langle\theta,X\rangle}}\right) = -\log 1 = 0. \end{aligned}$$

2. The statement follows from the dominated convergence theorem, because $\theta_n \to \theta$ implies that

$$m_{\theta_n}(X,Y) \to m_\theta(X,Y)$$

and

$$m_{\theta_n}(X,Y) \le \log(1+\mathrm{e}^{\|\theta_n\|\|X\|}) \le \log 2 + \|\theta_n\|\|X\| \le \log 2 + c\|X\|$$

with $c = \sup_n \|\theta_n\| < \infty$.

3. Let $M(\theta_n) \to M(\theta_0)$. By Proposition B.1.2, we have to prove that any subsequence $(\langle\theta_{n_k}, X\rangle)$ contains a further subsequence that tends in probability to $\langle\theta_0, X\rangle$. Note that $M(\theta_{n_k}) \to M(\theta_0)$, therefore, for ease of notation, we omit the index k.

The sequence of random vectors $(\langle\theta_n, X\rangle, \langle\theta_0, X\rangle)$ is tight in the space $\bar{\mathbb{R}} \times \mathbb{R}$. Indeed, if $K \subset \mathbb{R}$ is a compact interval such that $\mathsf{P}(\langle\theta_0, X\rangle \in K) \ge 1-\varepsilon$ (and we can always find such K), then the set $\bar{\mathbb{R}} \times K$ is also compact and for all n

$$\mathsf{P}((\langle\theta_n, X\rangle, \langle\theta_0, X\rangle) \in \bar{\mathbb{R}} \times K) = \mathsf{P}(\langle\theta_0, X\rangle \in K) \ge 1-\varepsilon.$$

By the Prokhorov's theorem (see [41], Theorem 14.3), there exists a subsequence $(\langle\theta_{n_k}, X\rangle, \langle\theta_0, X\rangle)$, which converges in distribution in the space $\bar{\mathbb{R}} \times \mathbb{R}$ to some random vector (U_1, U_2).

By Proposition B.1.5,

$$\mathsf{E}\left(\frac{\log(1+\mathrm{e}^{U_1})}{1+\mathrm{e}^{U_2}} + \frac{\log(1+\mathrm{e}^{-U_1})}{1+\mathrm{e}^{-U_2}}\right)$$
$$\le \varliminf_{k\to\infty} \mathsf{E}\left(\frac{\log(1+\mathrm{e}^{\langle\theta_{n_k},X\rangle})}{1+\mathrm{e}^{\langle\theta_0,X\rangle}} + \frac{\log(1+\mathrm{e}^{-\langle\theta_{n_k},X\rangle})}{1+\mathrm{e}^{-\langle\theta_0,X\rangle}}\right) = \varliminf_{k\to\infty} M(\theta_{n_k}) = M(\theta_0).$$

Obviously, U_2 is distributed identically to $\langle\theta_0, X\rangle$. Hence

$$M(\theta_0) = \mathsf{E}\left(\frac{\log(1+\mathrm{e}^{\langle\theta_0,X\rangle})}{1+\mathrm{e}^{\langle\theta_0,X\rangle}} + \frac{\log(1+\mathrm{e}^{-\langle\theta_0,X\rangle})}{1+\mathrm{e}^{-\langle\theta_0,X\rangle}}\right)$$
$$= \mathsf{E}\left(\frac{\log(1+\mathrm{e}^{U_2})}{1+\mathrm{e}^{U_2}} + \frac{\log(1+\mathrm{e}^{-U_2})}{1+\mathrm{e}^{-U_2}}\right)$$

and therefore

$$\mathsf{E}\left(\frac{\log(1+\mathrm{e}^{U_1})}{1+\mathrm{e}^{U_2}} + \frac{\log(1+\mathrm{e}^{-U_1})}{1+\mathrm{e}^{-U_2}}\right) \le \mathsf{E}\left(\frac{\log(1+\mathrm{e}^{U_2})}{1+\mathrm{e}^{U_2}} + \frac{\log(1+\mathrm{e}^{-U_2})}{1+\mathrm{e}^{-U_2}}\right).$$

Let V be a random variable gaining values -1 and 1 with (conditional w.r.t. (U_1,U_2)) probabilities $\frac{1}{1+e^{U_2}}$ and $\frac{1}{1+e^{-U_2}}$. Then the above inequality can be re-written as

$$\mathsf{E}\log(1+e^{-VU_1}) \le \mathsf{E}\log(1+e^{-VU_2}).$$

This yields

$$0 \le \mathsf{E}\log\frac{1+e^{-VU_2}}{1+e^{-VU_1}} \le \log\mathsf{E}\frac{1+e^{-VU_2}}{1+e^{-VU_1}}$$
$$= \log\mathsf{E}\left(\frac{1}{1+e^{U_1}}+\frac{1}{1+e^{-U_1}}\right) = \log 1 = 0.$$

Therefore, both inequality signs can be replaced by equalities. However, Jensen's inequality becomes equality if and only if the variable that is being integrated almost surely is a constant. In this case that constant is 0, that is, almost surely

$$\log\frac{1+e^{-VU_2}}{1+e^{-VU_1}} = 0$$

and $U_1 = U_2$.

Hence $(\langle\theta_{n_k},X\rangle,\langle\theta_0,X\rangle) \to_d (U_2,U_2)$ and therefore $\langle\theta_{n_k},X\rangle - \langle\theta_0,X\rangle \to_d U_2 - U_2 = 0$. When the limit random variable is 0 (or a constant), convergence in distribution is equivalent to convergence in probability ([41], Lemma 3.7). Therefore, $\langle\theta_{n_k},X\rangle - \langle\theta_0,X\rangle \to_p 0$ and $\langle\theta_{n_k},X\rangle \to_p \langle\theta_0,X\rangle$. □

For any $f \in C^r(E_k)$ we assume that its rth derivative at the point $\theta \in E_k$ is a symmetric r-linear form on E_k defined by

$$f^{(r)}(\theta)(d\theta_1,\ldots,d\theta_r) = D_{d\theta_r}\cdots D_{d\theta_1}f(\theta),$$

where $D_{d\theta}$ stands for the directional derivative along $d\theta \in E_k$. Its norm is defined by

$$\|f^{(r)}(\theta)\| = \sup_{\|d\theta_1\|\le 1,\ldots,\|d\theta_r\|\le 1} |f^{(r)}(\theta)(d\theta_1,\ldots,d\theta_r)|.$$

The function $d\theta \mapsto f^{(r)}(\theta)(d\theta,\ldots,d\theta)$ is called the rth differential of f and is

denoted by $d^r f(\theta)$. For example, $d^2 f(\theta)$ is a quadratic form associated with the bilinear form $f''(\theta)$.

For any $x \in E$ and $y \in \{-1, 1\}$, function $\theta \mapsto m_\theta(x, y)$ is infinitely differentiable on E_k and

$$\begin{aligned}
m'_\theta(x,y)d\theta &= \frac{\mathrm{e}^{-y\langle\theta,x\rangle}}{1+\mathrm{e}^{-y\langle\theta,x\rangle}}(-y\langle d\theta,x\rangle),\\
m''_\theta(x,y)(d\theta_1,d\theta_2) &= \frac{\mathrm{e}^{-y\langle\theta,x\rangle}}{(1+\mathrm{e}^{-y\langle\theta,x\rangle})^2}\langle d\theta_1,x\rangle\langle d\theta_2,x\rangle,\\
m'''_\theta(x,y)(d\theta_1,d\theta_2,d\theta_3) &= \frac{\mathrm{e}^{-y\langle\theta,x\rangle}-\mathrm{e}^{-2y\langle\theta,x\rangle}}{(1+\mathrm{e}^{-y\langle\theta,x\rangle})^3}\langle d\theta_1,x\rangle\langle d\theta_2,x\rangle(-y\langle d\theta_3,x\rangle).
\end{aligned}$$

It is obvious that

$$\begin{aligned}
|m'_\theta(X,Y)d\theta| &\le \|d\theta\|\|X\|,\\
|m''_\theta(X,Y)(d\theta_1,d\theta_2)| &\le |\langle d\theta_1,X\rangle||\langle d\theta_2,X\rangle| \le \|d\theta_1\|\|d\theta_2\|\|X\|^2,\\
|m'''_\theta(X,Y)(d\theta_1,d\theta_2,d\theta_3)| &\le \|d\theta_1\|\|d\theta_2\|\|d\theta_3\|\|X\|^3.
\end{aligned}$$

Therefore,

$$\|m'_\theta(X,Y)\| \le \|X\|, \quad \|m''_\theta(X,Y)\| \le \|X\|^2, \quad \|m'''_\theta(X,Y)\| \le \|X\|^3,$$

moreover, $\|X\|, \|X\|^2, \|X\|^3$ are integrable, if $\mathsf{E}\|X\|^3 < \infty$. Hence $M(\theta)$, as a function on E_k, belongs to $C^3(E_k)$, and

$$\begin{aligned}
\mathrm{d}M(\theta) &= -\mathsf{E}\frac{\mathrm{e}^{-Y\langle\theta,X\rangle}}{1+\mathrm{e}^{-Y\langle\theta,X\rangle}}Y\langle d\theta,X\rangle,\\
\mathrm{d}^2M(\theta) &= \mathsf{E}\frac{\mathrm{e}^{-Y\langle\theta,X\rangle}}{(1+\mathrm{e}^{-Y\langle\theta,X\rangle})^2}\langle d\theta,X\rangle^2,\\
\mathrm{d}^3M(\theta) &= -\mathsf{E}\frac{\mathrm{e}^{-Y\langle\theta,X\rangle}-\mathrm{e}^{-2Y\langle\theta,X\rangle}}{(1+\mathrm{e}^{-Y\langle\theta,X\rangle})^3}Y\langle d\theta,X\rangle^3.
\end{aligned}$$

If the distribution of X is of full rank, then, for any $d\theta \neq 0$, almost surely $\langle d\theta, X\rangle^2 > 0$ and therefore $\mathrm{d}^2M(\theta) > 0$. Hence, for all θ, $\mathrm{d}^2M(\theta)$ is a positive definite quadratic form. According to [42], $M(\theta)$ is strictly convex on E_k.

Proposition B.2.2. *If assumptions (FR) and (M) hold, then, for any $k \geq 1$, the function $M(\theta)$ has a unique minimum point in the space E_k. Furthermore, if θ_k is that point, then $M(\theta_k) \to M(\theta_0)$, as $k \to \infty$.*

Proof. Step 1: we will prove that sets $A_q = \{\theta \in E_k \mid M(\theta) \leq q\}$ are bounded.

Suppose the contrary. Then there exists some set A_q that is not bounded. Find a sequence $(\theta_m) \subset E_k$ such that $M(\theta_m) \leq q$ for all m, and $\|\theta_m\| \to \infty, \theta_m/\|\theta_m\| \to a$, as $m \to \infty$. Because $\|a\| = 1$ and the distribution of X is of full rank, either $\langle a, X\rangle < 0$ or $\langle a, X\rangle > 0$ with a positive probability. Since $0 < p_{\theta_0} < 1$,

$$0 < \mathsf{P}(Y\langle a, X\rangle < 0) \leq \mathsf{P}(\lim_{m\to\infty} m_{\theta_m}(X,Y) = \infty)$$

and so $\mathsf{E} \varliminf_{m\to\infty} m_{\theta_m}(X,Y) = \infty$. On the other hand, by Fatou's lemma,

$$\mathsf{E} \varliminf_{m\to\infty} m_{\theta_m}(X,Y) \leq \varliminf_{m\to\infty} M(\theta_m) \leq q.$$

A contradiction.

Step 2: the end of the proof.

The existence of θ_k follows from Proposition 2.1.1 of [42]. Since $M(\theta)$ is strictly convex, the minimum point is unique.

If $\theta_0^{(k)}$ is the projection of θ_0 in the space E_k, then $M(\theta_0) \leq M(\theta_k) \leq M(\theta_0^{(k)})$. From $\theta_0^{(k)} \to \theta_0$ we get that $M(\theta_0^{(k)}) \to M(\theta_0)$. Therefore, also $M(\theta_k) \to M(\theta_0)$. □

We are now ready to establish the consistency criterion. The following Proposition provides the consistency conditions for the estimate of the type $\hat{p} = p_{\hat{\theta}_n}$, where $\hat{\theta}_n$ is any estimate of θ. If $\hat{\theta}_n$ is defined by (2.1)-(2.2), we get the consistency criterion for the logistic estimate.

Proposition B.2.3. *1. If $M(\hat{\theta}_n) \to_p M(\theta_0)$, then the estimate $p_{\hat{\theta}_n}$ is consistent.*

2. Suppose assumptions (FR) and (M) hold, and θ_k is the minimum of the function M in the space E_k. If $k_n \to \infty$ and $M(\hat{\theta}_n) - M(\theta_{k_n}) \to_p 0$, then the estimate $p_{\hat{\theta}_n}$ is consistent.

Proof. 1. By Proposition B.2.1, $M(\theta_n) \to M(\theta_0)$ implies $\langle \theta_n, X\rangle \to_p \langle \theta_0, X\rangle$. Then $p_{\theta_n}(X) \to_p p_{\theta_0}(X)$ and, by Proposition B.1.4, $\mathsf{E}|p_{\theta_n}(X) - p_{\theta_0}(X)| \to 0$.

Let now $M(\hat{\theta}_n) \to_p M(\theta_0)$. We have to prove that $\mathsf{E}|p_{\hat{\theta}_n}(X) - p_{\theta_0}(X)| \to 0$. It is enough to prove that any subsequence $\mathsf{E}|p_{\hat{\theta}_{n_s}}(X) - p_{\theta_0}(X)|$ has a further subsequence that tends to 0. Moreover, it is well-known that any sequence that converges in probability has a subsequence that converges almost everywhere. Therefore, it is enough to prove that, if almost surely $M(\hat{\theta}_{n_s}) \to M(\theta_0)$, then $\mathsf{E}|p_{\hat{\theta}_{n_s}}(X) - p_{\theta_0}(X)| \to 0$.

However, if almost surely $M(\hat{\theta}_{n_s}) \to M(\theta_0)$, then from the first paragraph of this proof we get that almost surely

$$\mathsf{E}^*|p_{\hat{\theta}_{n_s}}(X) - p_{\theta_0}(X)| \to 0,$$

where E^* denotes the conditional mean w.r.t. sequence $((X_i, Y_i) \mid i \geq 1)$. It is enough to use the dominated convergence theorem.

2. The second statement follows from the first one and from Proposition B.2.2.

□

B.3 The function $M_n(\theta)$

Now suppose that k and n are fixed and consider $M_n(\theta)$, as a function on E_k. For all $\theta, d\theta \in E_k, x \in E$ and $y \in \{-1, 1\}$,

$$m''_\theta(x,y)(d\theta, d\theta) = \frac{\mathrm{e}^{-y\langle\theta,x\rangle}}{(1+\mathrm{e}^{-y\langle\theta,x\rangle})^2}\langle d\theta, x\rangle^2 \geq 0.$$

Therefore, the function $\theta \mapsto m_\theta(x,y)$ is convex in E_k. Then also the function $M_n(\theta)$ is convex. We first give conditions for its strict convexity.

Note that if $\theta \in E_k$, then $\langle \theta, X_i\rangle = \langle \theta, X_i^{(k)}\rangle$, where $X_i^{(k)}$ denotes the projection of vector X_i in the space E_k.

Proposition B.3.1. *If $n \geq k$ and $X_1^{(k)}, \ldots, X_k^{(k)}$ are linearly independent, then function $M_n(\theta)$ is strictly convex on E_k. If assumption (FR) holds, the probability of such event is* 1.

Proof. The function $M_n(\theta)$ is strictly convex if its second differential $\mathrm{d}^2 M_n(\theta)$ is a positive definite quadratic form. Since

$$\mathrm{d}^2 M_n(\theta) = \frac{1}{n}\sum_{i=1}^{n} \frac{\mathrm{e}^{-Y_i\langle\theta,X_i\rangle}}{(1+\mathrm{e}^{-Y_i\langle\theta,X_i\rangle})^2}\langle d\theta, X_i^{(k)}\rangle^2,$$

and all summands in the right-hand side are nonnegative, $\mathrm{d}^2 M_n(\theta) = 0$ implies that $d\theta$ is perpendicular to all $X_i^{(k)}$. If $n \geq k$ and $X_1^{(k)}, \ldots, X_k^{(k)}$ are linearly independent, then $d\theta = 0$.

The second statement follows from Lemma A.1. □

Recall some notions from Chapter 2. Let $(x_1, y_1), \ldots, (x_n, y_n)$ be n vectors from $E_k \times \{-1, 1\}$, called *sample points*, and $a \neq 0$ be another vector from E_k. We say that the vector *a separates sample points* if, for all i,

$$y_i\langle a, x_i\rangle \geq 0.$$

We say that sample points are *separable*, if there exists some $a \neq 0$ that separates them. Note that this definition is equivalent to the definition of quasi-complete separation, given by [12]. Next, the statement "the sample $(X_1, Y_1), \ldots, (X_n, Y_n)$ is k-separable" defines some event, the set of all elementary events ω such that sample points

$$(X_1^{(k)}(\omega), Y_1(\omega)), \ldots, (X_n^{(k)}(\omega), Y_n(\omega)) \tag{B.1}$$

are separable.

Proposition B.3.2. *If the sample $(X_1, Y_1), \ldots, (X_n, Y_n)$ is not k-separable then, for any $q > 0$, the (random) set $A_q = \{\theta \in E_k \mid M_n(\theta) \leq q\}$ is bounded.*

Proof. Fix any ω such that the set $A_q(\omega)$ is not bounded and denote $x_i = X_i^{(k)}(\omega)$, $y_i = Y_i(\omega)$. Find a sequence $(\theta_m) \subset A_q$ such that $\|\theta_m\| \to \infty$ and $\theta_m/\|\theta_m\| \to a$. Then, for all m and all $i = 1, \ldots, n$,

$$\log(1+\mathrm{e}^{-y_i\langle\theta_m,x_i\rangle}) \leq \sum_{i=1}^{n} \log(1+\mathrm{e}^{-y_i\langle\theta_m,x_i\rangle}) \leq nq.$$

But

$$-y_i\langle\theta_m,x_i\rangle = -\|\theta_m\|y_i\Big\langle\frac{\theta_m}{\|\theta_m\|},x_i\Big\rangle \to \infty$$

if $y_i\langle a,x_i\rangle < 0$. Hence $y_i\langle a,x_i\rangle \geq 0$ for all i, that is, a separates sample points (B.1).

□

Now suppose $n \geq k$ and let W_{kn} denote the following event: $X_1^{(k)},\ldots,X_k^{(k)}$ vectors are linearly independent and the sample is not k-separable. If $\omega \in W_{kn}$ then, by Propositions B.3.1 and B.3.2, the function $M_n(\theta)$ is strictly convex and all its sub-level sets A_q are bounded. As is seen from the proof of Proposition B.2.2, then $M_n(\theta)$ has the unique minimum point, which is, of course $\hat{\theta}_{kn}(\omega)$. If $\omega \notin W_{kn}$, we suppose that $\hat{\theta}_{kn}(\omega) = 0$.

Denote $q_{kn} = \mathsf{P}(W_{kn}^c)$. Then, by Proposition B.3.1 and by Corollary 2.1, $q_{k_n n} \to 0$, provided that assumption (FR) holds and $k_n/n \to 0$.

B.4 Proof of Theorem 3.1

We follow the proof of Theorem 5.42 from [2].

For $k \geq 1$ and $\theta \in E_k, x \in E, y \in \{-1,1\}$ let us define

$$\psi_{k,\theta}(x,y) = -\frac{\mathrm{e}^{-y\langle\theta,x\rangle}}{1+\mathrm{e}^{-y\langle\theta,x\rangle}}yx^{(k)},$$

where $x^{(k)}$ denotes the orthogonal projection of x in the space E_k. It is obvious that the function $\theta \mapsto \psi_{k,\theta}(x,y)$ is the gradient of the restriction of the function $m_\theta(x,y)$ on E_k. Also let us define

$$\Psi_{k,n}(\theta) = \overline{\psi_{k,\theta}(X,Y)}, \quad \text{and} \quad \Psi_k(\theta) = \mathsf{E}\psi_{k,\theta}(X,Y).$$

These functions are the gradients of the functions $M_n(\theta)$ and $M(\theta)$, as functions on E_k, respectively. Therefore, both $\Psi_{k,n}$ and Ψ_k are C^2-smooth functions from E_k to E_k. The derivative $\Psi_k'(\theta)$ is the linear operator from E_k to E_k which maps $d\theta_1 \in E_k$

to a vector $\Psi_k'(\theta)d\theta_1 \in E_k$ such that, for all $d\theta_2 \in E_k$,

$$\langle \Psi_k'(\theta)d\theta_1, d\theta_2 \rangle = M''(\theta)(d\theta_1, d\theta_2).$$

Proposition B.4.1. *The function Ψ_k is a diffeomorphism.*

Proof. Suppose $\Psi_k(\theta_1) = \Psi_k(\theta_2)$ and denote $d\theta = \theta_2 - \theta_1$. Then, for some $t \in (0,1)$,

$$0 = \langle \Psi_k(\theta_2), d\theta \rangle - \langle \Psi_k(\theta_1), d\theta \rangle = M''(\theta_1 + td\theta)(d\theta, d\theta).$$

This yields $d\theta = 0$, that is, $\theta_1 = \theta_2$. Therefore, the function Ψ_k is injective.

Analogously, from $\Psi_k'(\theta)d\theta = 0$ we get that

$$0 = \langle \Psi_k'(\theta)d\theta, d\theta \rangle = M''(\theta)(d\theta, d\theta)$$

and $d\theta = 0$. Therefore, the operator $\Psi_k'(\theta)$ is invertible for all θ.

The statement of the theorem now follows from the inverse function theorem. □

Proposition B.4.1 implies that the set $V = \Psi_k(E_k)$ is open. Moreover, $0 \in V$ because $\Psi_k(\theta_k) = 0$. Let us take some δ_k such that $\bar{U}(0,\delta_k) \subset V$ and denote $U_k = \Psi_k^{-1}(U(0,\delta_k))$. Then U_k is the neighborhood of the point θ_k. Moreover, because Ψ_k is a homeomorphism between E_k and V,

$$\Psi_k(\bar{U}_k) = \overline{\Psi_k(U_k)} = \overline{U(0,\delta_k)} = \bar{U}(0,\delta_k).$$

Denote

$$W_{kn}' = \{ \sup_{\theta \in \bar{U}_k} \|\Psi_{k,n}(\theta) - \Psi_k(\theta)\| \leq \delta_k \}.$$

The following reasoning is under the assumption that event $W_{kn} \cap W_{kn}'$ occurred.

If $z \in \bar{U}(0,\delta_k)$, then $\Psi_k^{-1}(z) \in \bar{U}_k$ and then

$$\|z - \Psi_{k,n}(\Psi_k^{-1}(z))\| = \|\Psi_k(\Psi_k^{-1}(z)) - \Psi_{k,n}(\Psi_k^{-1}(z))\| \leq \delta_k.$$

Therefore $z \mapsto z - \Psi_{k,n}(\Psi_k^{-1}(z))$ is a continuous function from $\bar{U}(0,\delta_k)$ to $\bar{U}(0,\delta_k)$. From the Brouwer's Fixed Point Theorem we get that, for some $z \in \bar{U}(0,\delta_k)$,

$$z = z - \Psi_{k,n}(\Psi_k^{-1}(z)),$$

that is, $\Psi_{k,n}(\Psi_k^{-1}(z)) = 0$. Because the function $M_n(\theta)$ is strictly convex, $\hat{\theta}_{kn}$ is the unique zero of the function $\Psi_{k,n}$. Therefore, $\hat{\theta}_{kn} = \Psi_k^{-1}(z) \in \bar{U}_k$.

Let $d_k = \text{diam}\bar{U}_k$. Then $\|\hat{\theta}_{kn} - \theta_k\| \leq d_k$ and

$$|M(\hat{\theta}_{kn}) - M(\theta_k)| \leq \sup_{\theta} \|\Psi_k(\theta)\| d_k \leq \mathsf{E}\|X\| d_k.$$

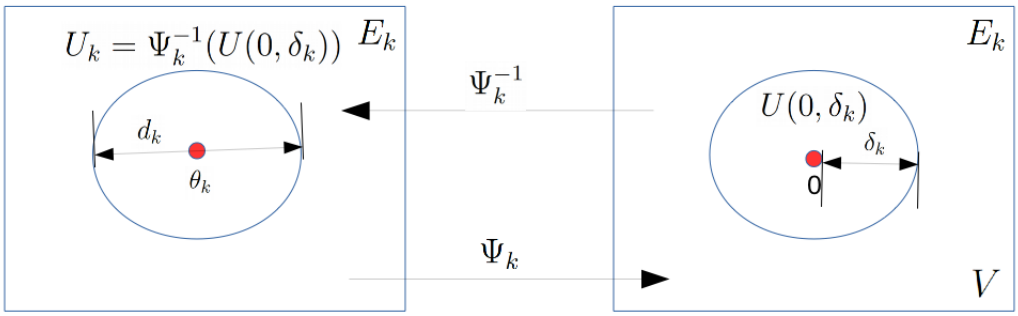

Figure B.1: Conceptual illustration of ideas from Theorem 5.42 in [2] that solves the well-known problem in statistics: by Law of Large Numbers, empirical expectation tends to true expectation. How to prove that the $\hat{\theta}_{kn}$ that minimizes the empirical expectation tends to θ_k that minimizes the true expectation? As van der Vaart suggests, if the distance between gradients of empirical and true expectations are bounded by δ_k, then the distance between $\hat{\theta}_{kn}$ and θ_k is bounded by d_k.

Therefore, in order to prove Theorem 3.1 it is enough to choose δ_k in such a way that $d_{k_n} \to 0$ and $\mathsf{P}(W_{k_n,n}^{\prime c}) \to 0$.

We now need to evaluate the diameter d_k. The following Proposition gives the necessary result.

Proposition B.4.2. *Suppose assumptions (FR), (M) and (UI) are satisfied and* $\delta_k = o(\sqrt{\tau_k})$, *as* $k \to \infty$. *Then* $d_k = O(\delta_k/\tau_k)$.

The proof of Proposition B.4.2 is preceded with three lemmas.

Lemma B.1. *Let (Z_n) be a sequence of positive integrable variables such that the sequence $(Z_n/\mathsf{E}Z_n)$ is uniformly integrable. Then, for all $q < 1$,*

$$\underline{\lim}_{n\to\infty} \mathsf{P}(Z_n \geq q\mathsf{E}Z_n) > 0.$$

Proof. Suppose the contrary. Without loss of generality, we can assume that

$$\mathsf{P}(Z_n \geq q\mathsf{E}Z_n) \to 0.$$

From uniform integrability we get that

$$\mathsf{E}\frac{Z_n}{\mathsf{E}Z_n}1_{\{Z_n \geq q\mathsf{E}Z_n\}} \to 0.$$

Therefore, there exists n such that

$$\mathsf{E}Z_n 1_{\{Z_n \geq q\mathsf{E}Z_n\}} < (1-q)\mathsf{E}Z_n.$$

But then

$$\mathsf{E}Z_n = \mathsf{E}Z_n 1_{\{Z_n \geq q\mathsf{E}Z_n\}} + \mathsf{E}Z_n 1_{\{Z_n < q\mathsf{E}Z_n\}} < (1-q)\mathsf{E}Z_n + q\mathsf{E}Z_n = \mathsf{E}Z_n.$$

A contradiction. □

Lemma B.2. *Suppose the assumptions (FR), (M) and (UI) hold and $\delta_k = o(\sqrt{\tau_k})$, as $k \to \infty$. Then there exists k_0 such that, for all $k \geq k_0$ and all $d\theta \in E_k$ with $\|d\theta\| = 1$,*

$$\exists t > 0 \ \langle \Psi_k(\theta_k + t d\theta), d\theta \rangle > \delta_k. \tag{B.2}$$

Proof. Step 1: we prove that if (B.2) fails, for some $k \geq 1$ and $d\theta \in E_k$ with $\|d\theta\| = 1$, then

$$\mathsf{E}(Y\langle d\theta, X\rangle)^- \leq \delta_k. \tag{B.3}$$

If (B.2) fails then, for some $t_m \to \infty$,

$$\delta_k \geq \langle \Psi_k(\theta_k + t_m d\theta), d\theta \rangle = -\mathsf{E} \frac{\mathrm{e}^{-Y\langle \theta_k, X\rangle - t_m Y\langle d\theta, X\rangle}}{1+\mathrm{e}^{-Y\langle \theta_k, X\rangle - t_m Y\langle d\theta, X\rangle}} Y\langle d\theta, X\rangle.$$

Note that

$$\frac{\mathrm{e}^{-Y\langle \theta_k, X\rangle - t_m Y\langle d\theta, X\rangle}}{1+\mathrm{e}^{-Y\langle \theta_k, X\rangle - t_m Y\langle d\theta, X\rangle}} \xrightarrow[m\to\infty]{} \begin{cases} 0, & \text{if } Y\langle d\theta, X\rangle > 0, \\ 1, & \text{if } Y\langle d\theta, X\rangle < 0, \end{cases}$$

Therefore (B.3) follows by dominated convergence.

Step 2: the end of the proof.

Suppose $\delta_k = o(\sqrt{\tau_k})$, as $k \to \infty$, but the assertion of the Lemma is false. Then there exists a sequence $k_m \to \infty$ and a sequence $(d\theta_m)$ such that, for all $m \geq 1$, $d\theta_m \in E_{k_m}$, $\|d\theta_m\| = 1$ and, by the result of Step 1, $\mathsf{E}(Y\langle d\theta_m, X\rangle)^- \leq \delta_{k_m}$. Hence

$$\frac{\mathsf{E}(Y\langle d\theta_m, X\rangle)^-}{\sqrt{\tau_{k_m}}} \xrightarrow[m\to\infty]{} 0.$$

Then also

$$\frac{\mathsf{E}(Y\langle d\theta_m, X\rangle)^-}{\sqrt{C(d\theta_m, d\theta_m)}} \xrightarrow[m\to\infty]{} 0.$$

But

$$\begin{aligned}
\mathsf{E}(Y\langle d\theta_m, X\rangle)^- &= -\mathsf{E}\langle d\theta_m, X\rangle 1_{\{\langle d\theta_m, X\rangle < 0, Y=1\}} + \mathsf{E}\langle d\theta_m, X\rangle 1_{\{\langle d\theta_m, X\rangle > 0, Y=-1\}} \\
&= \mathsf{E}|\langle d\theta_m, X\rangle| \left(\frac{1_{\{\langle d\theta_m, X\rangle < 0\}}}{1+\mathrm{e}^{-\langle \theta_0, X\rangle}} + \frac{1_{\{\langle d\theta_m, X\rangle > 0\}}}{1+\mathrm{e}^{\langle \theta_0, X\rangle}} \right) \\
&\geq \mathsf{E} \frac{|\langle d\theta_m, X\rangle|}{1+\mathrm{e}^{|\langle \theta_0, X\rangle|}} \\
&\geq \frac{\sqrt{C(d\theta_m, d\theta_m)}}{2} \mathsf{E} \frac{1_{\{|\langle d\theta_m, X\rangle| \geq \sqrt{C(d\theta_m, d\theta_m)}/2\}}}{1+\mathrm{e}^{|\langle \theta_0, X\rangle|}},
\end{aligned}$$

therefore

$$\mathsf{E} \frac{1_{\{|\langle d\theta_m, X\rangle| \geq \sqrt{C(d\theta_m, d\theta_m)}/2\}}}{1+\mathrm{e}^{|\langle \theta_0, X\rangle|}} \to 0.$$

This yields

$$\frac{1_{\{|\langle d\theta_m, X\rangle| \geq \sqrt{C(d\theta_m, d\theta_m)}/2\}}}{1+\mathrm{e}^{|\langle \theta_0, X\rangle|}} \to_p 0$$

and then

$$1_{\{|\langle d\theta_m, X\rangle| \geq \sqrt{C(d\theta_m, d\theta_m)}/2\}} \to_p 0,$$

that is,

$$\mathsf{P}(\langle d\theta_m, X\rangle^2 \geq C(d\theta_m, d\theta_m)/4) \to 0.$$

This contradicts Lemma B.1. □

If Z is a positive random variable and $\mathsf{E}Z = 1$, we can consider Z as a density, that is, with any random vector U there exists a random vector $\tilde{U}$ such that with any nonnegative or any bounded Borel function f

$$\mathsf{E}f(\tilde{U}) = \mathsf{E}f(U)Z.$$

We need the following property of the transformation $U \mapsto \tilde{U}$.

Lemma B.3. *Let (Z_n) be a sequence of positive random variables, $\mathsf{E}Z_n = 1$ for all n, (U_n) be another sequence of random variables and let $\tilde{U}_n$ be a random variable such that with any nonnegative or any bounded Borel function f*

$$\mathsf{E}f(\tilde{U}_n) = \mathsf{E}f(U_n)Z_n.$$

If the sequence (Z_n) is uniformly integrable, then $U_n = O_p(1)$ implies $\tilde{U}_n = O_p(1)$.

Proof. Fix ε and find c_1 such that

$$\sup_n \mathsf{E}Z_n 1_{\{Z_n > c_1\}} < \varepsilon.$$

Then find c such that

$$\sup_n \mathsf{P}(|U_n| > c) < \varepsilon/c_1.$$

Then for all n,

$$\begin{aligned}\mathsf{P}(|\tilde{U}_n| > c) &= \mathsf{E}1_{\{|\tilde{U}_n|>c\}} = \mathsf{E}1_{\{|U_n|>c\}}Z_n \\ &= \mathsf{E}1_{\{|U_n|>c, Z_n\le c_1\}}Z_n + \mathsf{E}1_{\{|U_n|>c, Z_n>c_1\}}Z_n \\ &\le c_1\mathsf{P}(|U_n|>c) + \mathsf{E}1_{\{Z_n>c_1\}}Z_n < 2\varepsilon.\end{aligned}$$

Therefore, $\tilde{U}_n = O_p(1)$. □

Now we are ready to prove Proposition B.4.2.

Proof. Lemma B.2 implies that if k is large enough then, for any $d\theta \in E_k$ with $\|d\theta\| = 1$, at least one of the values of the function $f(t) = \langle \Psi_k(\theta_k + td\theta), d\theta\rangle$ is greater than δ_k. The function is continuous, strictly increasing and equal to 0, when $t = 0$. Therefore, there exists unique $t = t_k(d\theta) > 0$ such that $\langle \Psi_k(\theta_k + td\theta), d\theta\rangle = \delta_k$.

Step 1: we will prove that $d_k \le 2\alpha_k$, where

$$\alpha_k = \sup_{\substack{d\theta\in E_k \\ \|d\theta\|=1}} t_k(d\theta).$$

It is enough to prove that $\Psi_k^{-1}(\bar{U}(0,\delta_k)) \subset \bar{U}(\theta_k, \alpha_k)$. Let $\theta \in \Psi_k^{-1}(\bar{U}(0,\delta_k))$, that is $\|\Psi_k(\theta)\| \le \delta_k$. Denote $d\theta = (\theta - \theta_k)/\|\theta - \theta_k\|$. Then

$$\langle \Psi_k(\theta_k + \|\theta - \theta_k\|d\theta), d\theta\rangle = \langle \Psi_k(\theta), d\theta\rangle \le \|\Psi_k(\theta)\|\|d\theta\| \le \delta_k.$$

Therefore, $\|\theta - \theta_k\| \le t_k(d\theta) \le \alpha_k$.

Step 2: transforming the task to a simpler one.

From the result in Step 1 we get that it is enough to prove that $\alpha_k = O(\delta_k/\tau_k)$, that is that $\alpha_k\tau_k/\delta_k = O(1)$. Suppose the contrary, that there exists some subsequence that is unbounded. Then, without loss of generality, we can assume

$$\alpha_k\tau_k/\delta_k \to \infty$$

and we need to get a contradiction.

Let $d\theta_k$ be unit-length vectors from E_k such that $t_k(d\theta_k)/\alpha_k \to 1$. Then

$$\tau_k t_k(d\theta_k)/\delta_k \to \infty$$

and so

$$C(d\theta_k, d\theta_k) t_k(d\theta_k)/\delta_k \to \infty. \tag{B.4}$$

For short, denote

$$t_k = t_k(d\theta_k), \quad u_k = t_k\sqrt{C(d\theta_k, d\theta_k)}, \quad \beta_k = \frac{\delta_k}{\sqrt{C(d\theta_k, d\theta_k)}}$$

and

$$Z_{1k} = \langle \theta_k, X\rangle, \quad Z_{2k} = \frac{\langle d\theta_k, X\rangle}{\sqrt{C(d\theta_k, d\theta_k)}}.$$

It is obvious that $\beta_k \le \delta_k/\sqrt{\tau_k} \to 0$ and from (B.4) we get that $u_k/\beta_k \to \infty$. Moreover,

$$\delta_k = \langle \Psi_k(\theta_k + t_k d\theta_k), d\theta_k\rangle = f_k(1) - f_k(0) = \int_0^1 f_k'(t)\mathrm{d}t,$$

where

$$f_k(t) = \langle \Psi_k(\theta_k + t t_k d\theta_k), d\theta_k\rangle$$

and

$$\begin{aligned} f_k'(t) &= t_k M''(\theta_k + t t_k d\theta_k)(d\theta_k, d\theta_k) = t_k \mathsf{E}\frac{\mathrm{e}^{-Y\langle \theta_k + t t_k d\theta_k, X\rangle}}{(1+\mathrm{e}^{-Y\langle \theta_k + t t_k d\theta_k, X\rangle})^2}\langle d\theta_k, X\rangle^2 \\ &= t_k C(d\theta_k, d\theta_k)\mathsf{E}\frac{\mathrm{e}^{-Y(Z_{1k} + t u_k Z_{2k})}}{(1+\mathrm{e}^{-Y(Z_{1k} + t u_k Z_{2k})})^2} Z_{2k}^2. \end{aligned}$$

Therefore,

$$\beta_k \to 0, \quad \beta_k/u_k \to 0, \quad \beta_k = u_k \mathsf{E}\int_0^1 \frac{\mathrm{e}^{-Y(Z_{1k} + t u_k Z_{2k})}}{(1+\mathrm{e}^{-Y(Z_{1k} + t u_k Z_{2k})})^2}\mathrm{d}t Z_{2k}^2$$

and we have to obtain a contradiction.

Step 3: selecting one more subsequence.

Since $\mathsf{E}Z_{2k}^2 = 1$, we can consider Z_{2k}^2 as a density. Then there exist random variables $\tilde{Y}_k, \tilde{Z}_{1k}$ and $\tilde{Z}_{2k}$ such that with any Borel function f

$$\mathsf{E}f(\tilde{Y}_k, \tilde{Z}_{1k}, \tilde{Z}_{2k}) = \mathsf{E}f(Y, Z_{1k}, Z_{2k})Z_{2k}^2.$$

As a separate case,

$$\mathsf{P}(|\tilde{Y}_k| = 1) = \mathsf{E}1_{\{|\tilde{Y}_k|=1\}} = \mathsf{E}1_{\{|Y|=1\}}Z_k^2 = \mathsf{E}Z_k^2 = 1,$$

that is, almost surely $\tilde{Y}_k \in \{-1, 1\}$. Moreover,

$$\beta_k = u_k \mathsf{E} \int_0^1 \frac{\mathrm{e}^{-\tilde{Y}_k(\tilde{Z}_{1k} + t u_k \tilde{Z}_{2k})}}{(1 + \mathrm{e}^{-\tilde{Y}_k(\tilde{Z}_{1k} + t u_k \tilde{Z}_{2k})})^2} \mathrm{d}t.$$

Since $Z_{1k} = \langle \theta_k, X \rangle \to_p \langle \theta_0, X \rangle$, we get $Z_{1k} = O_p(1)$. Since the sequence (Z_{2k}^2) is uniformly integrable, $Z_{2k}^2 = O_p(1)$ and then also $Z_{2k} = O_p(1)$. Then from Lemma B.3 we get that $\tilde{Y}_k = O_p(1), \tilde{Z}_{1k} = O_p(1)$ and $\tilde{Z}_{2k} = O_p(1)$. This means that also $(\tilde{Y}_k, \tilde{Z}_{1k}, \tilde{Z}_{2k}) = O_p(1)$. From Prochorov's theorem we get that some subsequence of that sequence converges in distribution. Therefore we can suppose that $u_k \to u$ (where u can be infinite), and $(\tilde{Y}_k, \tilde{Z}_{1k}, \tilde{Z}_{2k}) \to_d (\tilde{Y}, \tilde{Z}_1, \tilde{Z}_2)$.

Step 4: the case, where $u_k \to u < \infty$.

Denote

$$g_u(y, z_1, z_2) = \int_0^1 \frac{\mathrm{e}^{-y(z_1 + t u z_2)}}{(1 + \mathrm{e}^{-y(z_1 + t u z_2)^2})} \mathrm{d}t.$$

If $(y_k, z_{1k}, z_{2k}) \to (y, z_1, z_2)$, then for all t,

$$\frac{\mathrm{e}^{-y_k(z_{1k} + t u_k z_{2k})}}{(1 + \mathrm{e}^{-y_k(z_{1k} + t u_k z_{2k})})^2} \to \frac{\mathrm{e}^{-y(z_1 + t u z_2)}}{(1 + \mathrm{e}^{-y(z_1 + t u z_2)})^2}.$$

The sequence on the left is not greater than 1 for all t. Therefore, by the dominated convergence theorem $g_{u_k}(y_k, z_{1k}, z_{2k}) \to g_u(y, z_1, z_2)$. Then, by Proposition B.1.1,

$$g_{u_k}(\tilde{Y}_k, \tilde{Z}_{1k}, \tilde{Z}_{2k}) \to_d g_u(\tilde{Y}, \tilde{Z}_1, \tilde{Z}_2).$$

The sequence of random variables on the left hand side is not greater than 1. Therefore, by the Proposition B.1.4

$$\mathsf{E}g_u(\tilde{Y},\tilde{Z}_1,\tilde{Z}_2)=\lim_{k\to\infty}\mathsf{E}g_{u_k}(\tilde{Y}_k,\tilde{Z}_{1k},\tilde{Z}_{2k})=\lim_{k\to\infty}\frac{\beta_k}{u_k}=0.$$

We got a contradiction because g_u function is everywhere positive.

Step 5: the case, where $u_k\to\infty$.

From

$$\mathsf{E}\frac{1}{\tilde{Z}_{2k}^2}=\mathsf{E}\frac{Z_{2k}^2}{Z_{2k}^2}=1$$

we get that the sequence of random variables $(1/|\tilde{Z}_{2k}|)$ is uniformly integrable. Then by Proposition B.1.3

$$\mathsf{E}\frac{1}{|\tilde{Z}_2|}=\lim_{k\to\infty}\mathsf{E}\frac{1}{|\tilde{Z}_{2k}|}\le\sup_k\mathsf{E}\frac{1}{|\tilde{Z}_{2k}|}<\infty.$$

Therefore almost surely $\tilde{Z}_2\neq 0$.

For all $u>0, y\in\{-1,1\}, z_1\in\mathbb{R}$ and $z_2\neq 0$,

$$\begin{aligned}ug_u(y,z_1,z_2)&=u\int_0^1\frac{\mathrm{e}^{-y(z_1+tuz_2)}}{(1+\mathrm{e}^{-y(z_1+tuz_2)})^2}\mathrm{d}t=\frac{1}{yz_2}\frac{1}{1+\mathrm{e}^{-y(z_1+tuz_2)}}\Big|_0^1\\&=\frac{1}{yz_2}\left(\frac{1}{1+\mathrm{e}^{-y(z_1+uz_2)}}-\frac{1}{1+\mathrm{e}^{-yz_1}}\right)=\frac{\mathrm{e}^{-yz_1}-\mathrm{e}^{-y(z_1+uz_2)}}{yz_2(1+\mathrm{e}^{-y(z_1+uz_2)})(1+\mathrm{e}^{-yz_1})}.\end{aligned}$$

Let $u_k\to\infty$ and $(y_k,z_{1k},z_{2k})\to(y,z_1,z_2)$ with $z_2\neq 0$. Then if $yz_2<0$, then

$$u_kg_{u_k}(y_k,z_{1k},z_{2k})\to-\frac{1}{yz_2(1+\mathrm{e}^{-yz_1})},$$

and if $yz_2>0$, then

$$u_kg_{u_k}(y_k,z_{1k},z_{2k})\to\frac{\mathrm{e}^{-yz_1}}{yz_2(1+\mathrm{e}^{-yz_1})}.$$

In other words,

$$u_kg_{u_k}(y_k,z_{1k},z_{2k})\to\frac{1}{|yz_2|(1+\mathrm{e}^{-yz_1})}h(y,z_1,z_2)=\frac{1}{|z_2|(1+\mathrm{e}^{-yz_1})}h(y,z_1,z_2),$$

where

$$h(y,z_1,z_2)=\begin{cases}1, & \text{if } yz_2<0,\\ \mathrm{e}^{-yz_1}, & \text{if } yz_2>0.\end{cases}$$

By Proposition B.1.1,

$$u_k g_{u_k}(\tilde{Y}_k,\tilde{Z}_{1k},\tilde{Z}_{2k}) \to_d \frac{1}{|\tilde{Z}_2|(1+\mathrm{e}^{-\tilde{Y}\tilde{Z}_1})}h(\tilde{Y},\tilde{Z}_1,\tilde{Z}_2).$$

The sequence of random variables on the left hand side is dominated by the sequence $(1/|\tilde{Z}_{2k}|)$ which is uniformly integrable. Therefore by Proposition B.1.4

$$\mathsf{E}\frac{1}{|\tilde{Z}_2|}h(\tilde{Y},\tilde{Z}_1,\tilde{Z}_2)=\lim_{k\to\infty}u_k\mathsf{E}g_{u_k}(\tilde{Y}_k,\tilde{Z}_{1k},\tilde{Z}_{2k})=\lim_{k\to\infty}\beta_k=0.$$

Again, we got a contradiction because almost surely $\frac{1}{|\tilde{Z}_2|}h(\tilde{Y},\tilde{Z}_1,\tilde{Z}_2)>0$. □

It remains to estimate the probability $\mathsf{P}(W_{kn}^{\prime c})$. In order to do this, we have to estimate

$$\sup_{\theta\in\bar{U}_k}\|\Psi_{k,n}(\theta)-\Psi_k(\theta)\|.$$

Fix $\theta\in\bar{U}_k$ and denote $d\theta=\theta-\theta_k$. By using Taylor's expansion we get

$$\begin{aligned}\Psi_{k,n}(\theta)&=\Psi_{k,n}(\theta_k)+\Psi'_{k,n}(\theta_k)d\theta+r_{k,n}(\theta,d\theta),\\ \Psi_k(\theta)&=\Psi'_k(\theta_k)d\theta+r_k(\theta,d\theta),\end{aligned}$$

where

$$\begin{aligned}\|r_{k,n}(\theta,d\theta)\|&\le\sup_{0<t<1}\|\Psi''_{k,n}(\theta_k+td\theta)\|\|d\theta\|^2\le\overline{\|X\|^3}d_k^2,\\ \|r_k(\theta,d\theta)\|&\le\sup_{0<t<1}\|\Psi''_k(\theta_k+td\theta)\|\|d\theta\|^2\le\mathsf{E}\|X\|^3d_k^2.\end{aligned}$$

Therefore,

$$\sup_{\theta\in\bar{U}_k}\|\Psi_{k,n}(\theta)-\Psi_k(\theta)\|\le\|\Psi_{k,n}(\theta_k)\|+d_k\|\Psi'_{k,n}(\theta_k)-\Psi'_k(\theta_k)\|+d_k^2(\overline{\|X\|^3}+\mathsf{E}\|X\|^3)$$

and

$$\begin{aligned}\mathsf{P}(W_{nk}^{'c}) \leq \mathsf{P}(\|\Psi_{k,n}(\theta_k)\| > \delta_k/3) + \mathsf{P}(d_k\|\Psi'_{k,n}(\theta_k) - \Psi'_k(\theta_k)\| > \delta_k/3) \\ + \mathsf{P}(d_k^2(\overline{\|X\|^3} + \mathsf{E}\|X\|^3) > \delta_k/3).\end{aligned} \tag{B.5}$$

The first term on the right hand of (B.5) is estimated as follows. Let $(e_1, \ldots, e_k)$ be an orthonormal basis of E_k. Then

$$\begin{aligned}\mathsf{E}\|\Psi_{k,n}(\theta_k)\|^2 &= \sum_{j=1}^{k} \mathsf{E}\langle \Psi_{k,n}(\theta_k), e_j\rangle^2 = \sum_{j=1}^{k} \mathrm{Var}\langle \Psi_{k,n}(\theta_k), e_j\rangle \\ &= \frac{1}{n}\sum_{j=1}^{k} \mathrm{Var}\langle \psi_{k,\theta_k}(X,Y), e_j\rangle = \frac{1}{n}\sum_{j=1}^{k} \mathsf{E}\langle \psi_{k,\theta_k}(X,Y), e_j\rangle^2 \\ &= \frac{1}{n}\mathsf{E}\|\psi_{k,\theta_k}(X,Y)\|^2 \leq \frac{1}{n}\mathsf{E}\|X\|^2.\end{aligned}$$

Therefore, the probability that we are interested does not exceed

$$\frac{9\mathsf{E}\|X\|^2}{n\delta_{k_n}^2}.$$

Similarly, we can evaluate the second term of (B.5). Again, we would like to apply Chebyshev's inequality and get that

$$\mathsf{P}(Z > \delta_k/3d_k) \leq \frac{9d_k^2}{\delta_k^2}\mathsf{E}Z^2,$$

where $Z = \|\Psi'_{k,n}(\theta_k) - \Psi'_k(\theta_k)\|$. However, since $\Psi_{k,n}$ is a vector-valued function, its derivative is a linear operator which makes the exact computation of its norm very complex. To make things simpler, here we can use the Hilbert-Schmidt norm

instead, which is known to be greater than usual norm. Therefore,

$$
\begin{aligned}
&\mathsf{E}\|\Psi'_{k,n}(\theta_k) - \Psi'_k(\theta_k)\|^2 \\
&\le \sum_{j,j'=1}^{k} \mathsf{E}(\langle \Psi'_{k,n}(\theta_k)e_{j'}, e_j\rangle - \langle \Psi'_k(\theta_k)e_{j'}, e_j\rangle)^2 \\
&= \sum_{j,j'=1}^{k} \mathrm{Var}\langle \Psi'_{k,n}(\theta_k)e_{j'}, e_j\rangle = \frac{1}{n}\sum_{j,j'=1}^{k} \mathrm{Var}\langle \psi'_{k,\theta_k}(X,Y)e_{j'}, e_j\rangle \\
&\le \frac{1}{n}\sum_{j,j'=1}^{k} \mathsf{E}\langle \psi'_{k,\theta_k}(X,Y)e_{j'}, e_j\rangle^2 = \frac{1}{n}\sum_{j,j'=1}^{k} \mathsf{E}(m''_{\theta_k}(X,Y)(e_{j'}, e_j))^2 \\
&\le \frac{1}{n}\sum_{j,j'=1}^{k} \mathsf{E}\langle X, e_j\rangle^2\langle X, e_{j'}\rangle^2 = \frac{1}{n}\mathsf{E}\left(\sum_{j=1}^{k}\langle X, e_j\rangle^2\right)^2 = \frac{1}{n}\mathsf{E}\|X^{(k)}\|^4 \\
&\le \frac{1}{n}\mathsf{E}\|X\|^4
\end{aligned}
$$

and the second term on the right hand side of (B.5) does not exceed

$$\frac{9\mathsf{E}\|X\|^4 d_{k_n}^2}{n\delta_{k_n}^2}.$$

The third term of (B.5) tends to 0, if $d_{k_n}^2/\delta_{k_n} \to 0$.

Therefore, Theorem 3.1 will be proved, if we can select δ_k such that

$$d_{k_n} \to 0, \quad n\delta_{k_n}^2 \to \infty, \quad \frac{d_{k_n}^2}{n\delta_{k_n}^2} \to 0, \quad \frac{d_{k_n}^2}{\delta_{k_n}} \to 0.$$

Note that the third condition is implied by the first and the second ones. If we take $\delta_k = o(\tau_k^2)$, then the first and the fourth conditions are met because then $d_k = O(\delta_k/\tau_k) = o(1)$ and $d_k^2/\delta_k = O(\delta_k/\tau_k^2) = o(1)$. Therefore, it is enough to select $\delta_k = o(\tau_k^2)$ such that $n\delta_{k_n}^2 \to \infty$, that is, in such a way that asymptotically

$$n^{-1/2} \prec \delta_{k_n} \prec \tau_{k_n}^2,$$

where $a \prec b$ means that $a = o(b)$. Clearly, we can achieve this, if

$$n^{-1/2} \prec \tau_{k_n}^2,$$

that is, if $n\tau_{k_n}^4 \to \infty$ which is exactly the assumption of Theorem 3.1.

B.5 Proof of Theorem 3.2

Proof. Define a new Hilbert space $\bar{E} = \mathbb{R} \times E$ with the inner product

$$\langle (\alpha, \theta), (a, x) \rangle = \alpha a + \langle \theta, x \rangle,$$

where $\alpha, a \in \mathbb{R}$ and $\theta, x \in E$, and set $\bar{X} = (1, X) \in \bar{E}$. Take any $\bar{\theta} = (\alpha, \theta) \neq 0$. If $\theta \neq 0$, then $\mathsf{P}(\langle \bar{\theta}, \bar{X} \rangle = 0) = 0$ because of (FR').

If $\theta = 0$, then $\alpha \neq 0$ and therefore

$$\mathsf{P}(\langle \bar{\theta}, \bar{X} \rangle = 0) = \mathsf{P}(\alpha = 0) = 0.$$

Hence $\bar{X}$ satisfies condition (FR). Moreover, if X satisfies (M), then

$$\mathsf{E}\|\bar{X}\|^4 = \mathsf{E}\langle \bar{X}, \bar{X} \rangle^2 = \mathsf{E}\,(1 + \langle X, X \rangle)^2 = 1 + 2\mathsf{E}\|X\|^2 + \mathsf{E}\|X\|^4 < \infty,$$

that is, $\bar{X}$ also satisfies (M). Finally, suppose X satisfies (UI). Fix ε and find c_0 such that for all $c > c_0$ and all θ

$$\mathsf{E}\langle \theta, X \rangle^2 1_{\{\langle \theta, X \rangle^2 > (c\mathsf{E}\langle \theta, X \rangle^2)/2\}} \leq \varepsilon \mathsf{E}\langle \theta, X \rangle^2.$$

Denote $\bar{c}_0 = \max(c_0, 2, 1/\varepsilon)$. Take $c > \bar{c}_0$ and any $\bar{\theta} = (\alpha, \theta)$ with norm equal to 1. Then by Chebyshev's inequality

$$\alpha^2 \mathsf{P}((\alpha + \langle \theta, X \rangle)^2 > c(\alpha^2 + \mathsf{E}\langle \theta, X \rangle^2)) \leq \alpha^2 / c \leq \alpha^2 \varepsilon$$

and

$$\begin{aligned}
\mathsf{E}\langle\theta,X\rangle^2 1_{\{(\alpha+\langle\theta,X\rangle)^2>c(\alpha^2+\mathsf{E}\langle\theta,X\rangle^2)\}} &\le \mathsf{E}\langle\theta,X\rangle^2 1_{\{2\alpha^2+2\langle\theta,X\rangle^2>c(\alpha^2+\mathsf{E}\langle\theta,X\rangle^2)\}} \\
&= \mathsf{E}\langle\theta,X\rangle^2 1_{\{2\langle\theta,X\rangle^2>c\mathsf{E}\langle\theta,X\rangle^2+(c-2)\alpha^2\}} \\
&\le \mathsf{E}\langle\theta,X\rangle^2 1_{\{\langle\theta,X\rangle^2>c/2\mathsf{E}\langle\theta,X\rangle^2\}} \\
&< \varepsilon\mathsf{E}\langle\theta,X\rangle^2.
\end{aligned}$$

Therefore,

$$\begin{aligned}
&\mathsf{E}\langle\bar\theta,\bar X\rangle^2 1_{\{\langle\bar\theta,\bar X\rangle^2>c(\mathsf{E}\langle\bar\theta,\bar X\rangle^2)\}} \\
&= \mathsf{E}(\alpha+\langle\theta,X\rangle)^2 1_{\{(\alpha+\langle\theta,X\rangle)^2>c(\alpha^2+\mathsf{E}\langle\theta,X\rangle^2)\}} \\
&\le 2\alpha^2\mathsf{E}1_{\{(\alpha+\langle\theta,X\rangle)^2>c(\alpha^2+\mathsf{E}\langle\theta,X\rangle^2)\}} + 2\mathsf{E}\langle\theta,X\rangle^2 1_{\{(\alpha+\langle\theta,X\rangle)^2>c(\alpha^2+\mathsf{E}\langle\theta,X\rangle^2)\}} \\
&\le 2\varepsilon(\alpha^2+\mathsf{E}\langle\theta,X\rangle^2),
\end{aligned}$$

that is, $\bar X$ satisfies condition (UI).

Define

$$\bar C(\bar\theta_1,\bar\theta_2) = \mathsf{E}\langle\bar\theta_1,\bar X\rangle\langle\bar\theta_2,\bar X\rangle, \quad \bar\tau_k = \min_{\substack{\bar\theta\in\mathbb{R}\times E_k \\ \|\bar\theta\|=1}} \bar C(\bar\theta,\bar\theta).$$

Note that

$$\bar C(\bar\theta,\bar\theta) = \mathsf{E}\langle\bar\theta,\bar X\rangle^2 = \mathsf{E}(\alpha+\langle\theta,X\rangle)^2 = \alpha^2+2\alpha\mathsf{E}\langle\theta,X\rangle+\mathsf{E}\langle\theta,X\rangle^2 = \alpha^2+\mathsf{E}\langle\theta,X\rangle^2.$$

Since C is a bilinear form, for all $\theta\in E_k$

$$\alpha^2+C(\theta,\theta) = \alpha^2+\|\theta\|^2 C(\theta/\|\theta\|,\theta/\|\theta\|) \ge \alpha^2+\|\theta\|^2\tau_k.$$

Therefore,

$$\bar\tau_k \ge \min_{|\alpha|\le 1}\left(\alpha^2+(1-\alpha^2)\tau_k\right) = \min(1,\tau_k)$$

and

$$n\bar{\tau}_{k_n}^4 = n\min(1,\tau_{k_n}^4) = \min(n, n\tau_{k_n}^4) \to \infty.$$

Then, by Theorem 1, the corresponding logistic estimate

$$\tilde{\theta}_{kn} = \arg\min_{\bar{\theta}\in\mathbb{R}\times E_k} \bar{M}_n(\bar{\theta}), \tag{B.6}$$

where

$$\bar{M}_n(\bar{\theta}) = \overline{m_{\bar{\theta}}(\bar{X},Y)}, \quad m_{\bar{\theta}}(\bar{x},y) = \log(1+\mathrm{e}^{-y\langle\bar{\theta},\bar{x}\rangle})$$

is consistent on $\bar{E} = \mathbb{R}\times E$. It remains to note that the logistic estimate (B.6) is the same as the estimate (3.4). □

Appendix C

Proofs for Chapter 4

C.1 Proof of Theorem 4.1

With $t \in [0,1]$ define

$$
\begin{aligned}
a_{*n}(t) &= \inf_{sh_n \le t' < (s+1)h_n} a(t'), \quad \text{for } sh_n \le t < (s+1)h_n, \\
a_n^*(t) &= \sup_{sh_n \le t' < (s+1)h_n} a(t'), \quad \text{for } sh_n \le t < (s+1)h_n, \\
b_{*n}(t) &= \inf_{sh_n \le t' < (s+1)h_n} b(t'), \quad \text{for } sh_n \le t < (s+1)h_n, \\
b_n^*(t) &= \sup_{sh_n \le t' < (s+1)h_n} b(t'), \quad \text{for } sh_n \le t < (s+1)h_n,
\end{aligned}
$$

We will proceed with the following Lemma.

Lemma C.1. *Suppose assumptions (A)-(F) hold and $h_n \to 0$. Then*

$$
\begin{aligned}
\sup_{t \in [0,1]} \left| a_n^*(t) - a_{*n}(t) \right| &\to 0, \\
\sup_{t \in [0,1]} \left| b_n^*(t) - b_{*n}(t) \right| &\to 0.
\end{aligned}
$$

Proof. We will prove the first statement and the second can be proved analogously.

Since function a is continuous on $[0,1]$, it is uniformly continuous on $[0,1]$, that is

$$\forall \varepsilon \; \exists \delta \; \forall t,t' \in [0,1] \; (|t-t'| < \delta \implies |a(t)-a(t')| < \varepsilon).$$

Also, for all $t \in [0,1]$,

$$a_{*n}(t) \leq a(t) \leq a_n^*(t). \tag{C.1}$$

Fix ε and find δ such that $|a(t)-a(t')| < \varepsilon$ for all $|t-t'| < \delta$. Find n_0 such that $h_n < \delta$ for all $n \geq n_0$. Take any $n \geq n_0$ and any $t \in [0,1]$. Let us suppose that $sh_n \leq t < (s+1)h_n$. Since with any t' from that interval $|t-t'| < \delta$,

$$a(t') > a(t) - \varepsilon \quad \text{and} \quad a(t') < a(t) + \varepsilon.$$

Then also

$$a_{*n}(t) \geq a(t) - \varepsilon \quad \text{and} \quad a_n^*(t) \leq a(t) + \varepsilon.$$

These inequalities hold for any t. Keeping in mind also (C.1) we get that

$$\sup_t |a_{*n}(t) - a(t)| \leq \varepsilon \quad \text{and} \quad \sup_t |a_n^*(t) - a(t)| \leq \varepsilon.$$

Therefore we proved that

$$\sup_t |a_{*n}(t) - a(t)| \to 0 \quad \text{and} \quad \sup_t |a_n^*(t) - a(t)| \to 0,$$

as $n \to \infty$. From these we get that

$$\sup_t |a_n^*(t) - a_{*n}(t)| \to 0.$$

□

We are now ready to prove Theorem 4.1.

Proof. Recall that the true reference functions are

$$x_*(t) = c_* a(t) + b(t), \quad x^*(t) = c^* a(t) + b(t).$$

Denote

$$a_* = \inf_{t\in[0,1]} a(t) > 0, \quad a^* = \sup_{t\in[0,1]} a(t) < \infty,$$

and with $n, i, j \geq 1$ and $t \in [0,1]$ denote

$$I_{nij}(t) = 1_{\{sh_n \leq T_{ij} < (s+1)h_n\}}, \quad \text{for } sh_n \leq t < (s+1)h_n,$$
$$I^*_{ni}(t) = \max_{1\leq j\leq M_i} I_{nij}(t).$$

Obviously, with any $t \in [0,1]$,

$$\mathsf{P}(I^*_{ni}(t) = 0) = \mathsf{E}(1-h_n)^M \leq (1-h_n).$$

We will suppose that i gains values from 1 to n and for a fixed i, j gains values from 1 to M_i.

1. First we will investigate functions $\hat{X}^*_n$. Fix $\varepsilon > 0$.

Step 1. Find n_{11} such that for all $n \geq n_{11}$ and all $t \in [0,1]$

$$a^*_n(t) - a_{*n}(t) \leq \frac{\varepsilon a_*}{3a^*c^*}.$$

Then for all $n \geq n_{11}$ and all $t \in [0,1]$

$$\frac{a^*_n(t)}{a_{*n}(t)} = \frac{a^*_n(t) - a_{*n}(t)}{a_{*n}(t)} + 1 \leq \frac{a^*_n(t) - a_{*n}(t)}{a_*} + 1 \leq 1 + \frac{\varepsilon}{3a^*c^*}$$

and

$$\frac{c^* a^*_n(t) - \varepsilon}{a_{*n}(t)} \leq c^*\left(1 + \frac{\varepsilon}{3a^*c^*}\right) - \frac{\varepsilon}{a_{*n}(t)} \leq c^*\left(1 + \frac{\varepsilon}{3a^*c^*}\right) - \frac{\varepsilon}{a^*} = c^* - 2\varepsilon',$$

where $\varepsilon' = \varepsilon/(3a^*)$. Denote $q = \mathsf{P}(C < c^* - \varepsilon')$. Then $q < 1$.

Analogously, find n_{12} such that for all $n \geq n_{12}$ and all $t \in [0,1]$

$$b_n^*(t) - b_{*n}(t) \leq \frac{\varepsilon a_*}{3a^*}.$$

Then for all $n \geq n_{12}$ and all $t \in [0,1]$,

$$\frac{b_n^*(t) - b_{*n}(t)}{a_{*n(t)}} \leq \frac{b_n^*(t) - b_{*n}(t)}{a_*} \leq \varepsilon'.$$

Take $n_1 = \max(n_{11}, n_{12})$. Then for all $n \geq n_1$

$$\begin{aligned}
&\mathsf{P}(\sup_{t\in[0,1]} (x^*(t) - \hat{X}_n^*(t)) > \varepsilon) \\
&= \mathsf{P}(\exists t \in [0,1]\ \hat{X}_n^*(t) < x^*(t) - \varepsilon) \\
&= \mathsf{P}(\exists t \in [0,1]\ (\forall i\ \forall j\ (I_{nij}(t) = 1 \implies X_i(T_{ij}) < x_*(t) - \varepsilon))) \\
&= \mathsf{P}(\exists t \in [0,1]\ (\forall i\ \forall j\ (I_{nij}(t) = 1 \implies C_i < (c^* a(t) + b(t) - b(T_{ij}) - \varepsilon)/a(T_{ij})))) \\
&\leq \mathsf{P}(\exists t \in [0,1]\ (\forall i\ \forall j\ (I_{nij}(t) = 1 \implies C_i < (c^* a_n^*(t) - \varepsilon)/a_{*n}(t) + (b_n^*(t) - b_{*n}(t))/a_{*n}(t)))) \\
&= \mathsf{P}(\exists t \in [0,1]\ (\forall i\ \forall j\ (I_{nij}(t) = 1 \implies C_i < c^* - \varepsilon'))) \\
&= \mathsf{P}(\exists s\ (\forall i\ \forall j\ (I_{nij}(sh_n) = 1 \implies C_i < c^* - \varepsilon'))) \\
&\leq \sum_{s=0}^{l_n - 1} \mathsf{P}(\forall i\ \forall j\ (I_{nij}(sh_n) = 1 \implies C_i < c^* - \varepsilon')) \\
&= \sum_{s=0}^{l_n - 1} \prod_{i=1}^{n} \mathsf{P}(\forall j\ (I_{nij}(sh_n) = 1 \implies C_i < c^* - \varepsilon')).
\end{aligned}$$

Let P^{TM} denote conditional probability w.r.t. families (T_{ij}) and (M_i). Then

$$\mathsf{P}^{TM}(\forall j\ (I_{nij}(sh_n) = 1 \implies C_i < c^* - \varepsilon')) = \begin{cases} q, & \text{if } \exists j\ I_{nij}(sh_n) = 1, \\ 1, & \text{otherwise.} \end{cases}$$

Then

$$\begin{aligned}\mathsf{P}(\forall j\ (I_{nij}(sh_n)=1 \implies C_i < c^* - \varepsilon')) &= \mathsf{EP}^{TM}(\forall j\ (I_{nij}(sh_n)=1 \implies C_i < c^* - \varepsilon')) \\ &= \mathsf{P}(I^*_{ni}(sh_n)=0) + q\mathsf{P}(I^*_{ni}(sh_n)=1) \\ &= q + (1-q)\mathsf{P}(I^*_{ni}(sh_n)=0) \\ &\le q + (1-q)(1-h_n) \\ &= 1 - h_n(1-q).\end{aligned}$$

Thus, for all $n \ge n_1$

$$\mathsf{P}\left(\sup_{t\in[0,1]}(x_*(t) - \hat{X}_{*n}(t)) > \varepsilon\right) \le \sum_{s=0}^{l_n-1}[1-h_n(1-q)]^n \le n[1-(1-q)h_n]^n.$$

Step 2. Find n_{21} such that for all $n \ge n_{21}$ and all $t \in [0,1]$

$$a^*_n(t) - a_{*n}(t) \le \frac{\varepsilon a_*}{3a^*c^*}.$$

Then for all $n \ge n_{21}$ and $t \in [0,1]$

$$\frac{a_{*n}(t)}{a^*_n(t)} = 1 - \frac{a^*_n(t) - a_{*n}(t)}{a^*_n(t)} \ge 1 - \frac{a^*_n(t) - a_{*n}(t)}{a_*} \ge 1 - \frac{\varepsilon}{3a^*c^*}$$

and

$$\frac{c^* a_{*n}(t) + \varepsilon}{a^*_n(t)} \ge c^*\left(1 - \frac{\varepsilon}{3a^*c^*}\right) + \frac{\varepsilon}{a^*_n(t)} \ge c^*\left(1 - \frac{\varepsilon}{3a^*c^*}\right) + \frac{\varepsilon}{a^*} = c^* + 2\varepsilon',$$

where $\varepsilon' = \varepsilon/3a^*$. Analogously, find n_{22} such that, for all $n \ge n_{22}$ and all $t \in [0,1]$,

$$b^*_n(t) - b_{*n}(t) \le \frac{\varepsilon a_*}{3a^*}.$$

Then for all $n \ge n_{22}$ and all $t \in [0,1]$

$$-\frac{b^*_n(t) - b_{*n}(t)}{a^*_n(t)} \ge -\frac{b^*_n(t) - b_{*n}(t)}{a_*} \ge -\varepsilon'.$$

Take $n_2 = \max(n_{21}, n_{22})$. Then for all $n \geq n_2$

$$\begin{aligned}
\mathsf{P}\left(\sup_{t\in[0,1]}(\hat{X}_n^*(t) - x^*(t)) > \varepsilon\right) &= \mathsf{P}(\exists t \in [0,1]\ \hat{X}_n^*(t) > x^*(t) + \varepsilon) \\
&= \mathsf{P}(\exists t \in [0,1]\ \exists i\ \exists j\ (I_{nij}(t) = 1, X_i(T_{ij}) > x^*(t) + \varepsilon)) \\
&\leq \mathsf{P}(\exists t \in [0,1]\ \exists i\ C_i > (c^* a_{*n}(t) + \varepsilon)/a_n^*(t) + (b_{*n}(t) - b_n^*(t))/a_n^*(t)) \\
&\leq \mathsf{P}(\exists i\ C_i > c^* + \varepsilon') = 0.
\end{aligned}$$

Step 3. From the results in Step 1 and Step 2 we get that for all $n \geq \max(n_1, n_2)$

$$\mathsf{P}\left(\sup_{t\in[0,1]}|\hat{X}_n^*(t) - x^*(t)| > \varepsilon\right) \leq n(1 - h_n(1-q))^n \leq \mathrm{e}^{\log n - nh_n(1-q)}. \tag{C.2}$$

If (4.8) holds, then, for n sufficiently large,

$$n^2 \mathrm{e}^{\log n - nh_h(1-q)} = \mathrm{e}^{3\log n - nh_n(1-q)} = \mathrm{e}^{-nh_n(1-q)(1+o(1))} \leq \mathrm{e}^{-nh_n(1-q)/2} \leq 1,$$

that is,

$$\mathrm{e}^{\log n - nh_h(1-q)} = O(n^{-2}).$$

Note that the term on the right hand side of (C.2) is summable. Then also

$$\sum_{n=1}^{\infty} \mathsf{P}\left(\sup_{t\in[0,1]}|\hat{X}_n^*(t) - x^*(t)| > \varepsilon\right) < \infty.$$

Therefore, almost surely $\sup_{t\in[0,1]}|\hat{X}_n^*(t) - x^*(t)| \to 0$.

2. Now we will investigate functions $\hat{X}_{*n}$. Fix $\varepsilon > 0$.

Step 1. Find n_{11} such that for all $n \geq n_{11}$ and all $t \in [0,1]$

$$a_n^*(t) - a_{*n}(t) \leq \frac{\varepsilon a_*}{3a^* c^*}.$$

Then for all $n \geq n_{11}$ and $t \in [0,1]$

$$\frac{a_{*n}(t)}{a_n^*(t)} = 1 - \frac{a_n^*(t) - a_{*n}(t)}{a_n^*(t)} \geq 1 - \frac{a_n^*(t) - a_{*n}(t)}{a_*} \geq 1 - \frac{\varepsilon}{3a^*c^*}$$

and

$$\frac{c_* a_{*n}(t) + \varepsilon}{a_n^*(t)} \geq c_* \left(1 - \frac{\varepsilon}{3a^*c^*}\right) + \frac{\varepsilon}{a^*} = c_* + 2\varepsilon',$$

where $\varepsilon' = \varepsilon/3a^*$. Analogously, find n_{12} such that for all $n \geq n_{12}$ and all $t \in [0,1]$

$$b_n^*(t) - b_{*n}(t) \leq \frac{\varepsilon a_*}{3a^*}.$$

Then for all $n \geq n_{12}$ and all $t \in [0,1]$,

$$-\frac{b_n^*(t) - b_{*n}(t)}{a_n^*(t)} \geq -\frac{b_n^*(t) - b_{*n}(t)}{a_*} \geq -\varepsilon'.$$

Denote $q = \mathsf{P}(C > c_* + \varepsilon')$. Then $q < 1$. Denote $n_1 = \max(n_{11}, n_{12})$.

Then for all $n \geq n_1$

$$\begin{aligned}
&\mathsf{P}(\sup_{t\in[0,1]} (\hat{X}_{*n}(t) - x_*(t)) > \varepsilon) \\
&= \mathsf{P}(\exists t \in [0,1]\ \hat{X}_{*n}(t) > x_*(t) + \varepsilon) \\
&= \mathsf{P}(\exists t \in [0,1]\ (\forall i\ \forall j\ (I_{nij}(t) = 1 \implies X_i(T_{ij}) > x_*(t) + \varepsilon))) \\
&= \mathsf{P}(\exists t \in [0,1]\ (\forall i\ \forall j\ (I_{nij}(t) = 1 \implies C_i a(T_{ij}) + b(T_{ij}) > c_* a(t) + b(t) + \varepsilon))) \\
&\leq \mathsf{P}(\exists t \in [0,1]\ (\forall i\ \forall j\ (I_{nij}(t) = 1 \implies C_i > (c_* a_{*n}(t) + \varepsilon)/a_n^*(t) + (b_{*n}(t) - b_n^*(t))/a_n^*(t)))) \\
&\leq \mathsf{P}(\exists t \in [0,1]\ (\forall i\ \forall j\ (I_{nij}(t) = 1 \implies C_i > c_* + \varepsilon'))) \\
&= \mathsf{P}(\exists s\ (\forall i\ \forall j\ (I_{nij}(sh_n) = 1 \implies C_i > c_* + \varepsilon'))) \\
&\leq \sum_{s=0}^{l_n - 1} \mathsf{P}(\forall i\ \forall j\ (I_{nij}(sh_n) = 1 \implies C_i > c_* + \varepsilon')) \\
&= \sum_{s=0}^{l_n - 1} \prod_{i=1}^{n} \mathsf{P}(\forall j\ (I_{nij}(sh_n) = 1 \implies C_i > c_* + \varepsilon')).
\end{aligned}$$

Let P^{TM} denote conditional probability w.r.t. families (T_{ij}) and (M_i). Then

$$\mathsf{P}^{TM}(\forall j\ (I_{nij}(sh_n)=1 \implies C_i > c_* + \varepsilon')) = \begin{cases} q, & \text{if } \exists j\ I_{nij}(sh_n)=1, \\ 1, & \text{otherwise.} \end{cases}$$

Then

$$\begin{aligned} \mathsf{P}(\forall j\ (I_{nij}(sh_n)=1 \implies C_i > c_* + \varepsilon')) &= \mathsf{E}\mathsf{P}^{TM}(\forall j\ (I_{nij}(sh_n)=1 \implies C_i > c_* + \varepsilon')) \\ &= \mathsf{P}(I^*_{ni}(sh_n)=0) + q\mathsf{P}(I^*_{ni}(sh_n)=1) \\ &= q + (1-q)\mathsf{P}(I^*_{ni}(sh_n)=0) \\ &\le q + (1-q)(1-h_n) \\ &= 1 - h_n(1-q). \end{aligned}$$

Thus, for all $n \ge n_1$

$$\mathsf{P}\left(\sup_{t\in[0,1]} (\hat{X}_{*n}(t) - x_*(t)) > \varepsilon\right) \le \sum_{s=0}^{l_n-1} [1-h_n(1-q)]^n \le n[1-(1-q)h_n]^n.$$

Step 2 Find n_{21} such that, for all $n \ge n_{21}$ and all $t \in [0,1]$

$$a^*_n(t) - a_{*n}(t) \le \frac{\varepsilon a_*}{3a^*c^*}.$$

Then for all $n \ge n_{21}$ and $t \in [0,1]$

$$\frac{a^*_n(t)}{a_{*n}(t)} = \frac{a^*_n(t) - a_{*n}(t)}{a_{*n}(t)} + 1 \le \frac{a^*_n(t) - a_{*n}(t)}{a_*} + 1 \le 1 + \frac{\varepsilon}{3a^*c^*}$$

and

$$\frac{c_* a^*_n(t) - \varepsilon}{a_{*n}(t)} \le c_*\left(1 + \frac{\varepsilon}{3a^*c^*}\right) - \frac{\varepsilon}{a_{*n}(t)} \le c_*\left(1 + \frac{\varepsilon}{3a^*c^*}\right) - \frac{\varepsilon}{a^*} = c_* - 2\varepsilon',$$

where $\varepsilon' = \varepsilon/3a^*$. Analogously, find n_{22} such that for all $n \ge n_{22}$ and all $t \in [0,1]$

$$b^*_n(t) - b_{*n}(t) \le \frac{\varepsilon a_*}{3a^*}.$$

Then for all $n \geq n_{22}$ and all $t \in [0,1]$,

$$\frac{b_n^*(t) - b_{*n}(t)}{a_{*n}(t)} \leq \frac{b_n^*(t) - b_{*n}(t)}{a_*} \leq \varepsilon'.$$

Take $n_2 = \max(n_{21}, n_{22})$. Then for all $n \geq n_2$

$$\begin{aligned}
&\mathsf{P}\left(\sup_{t\in[0,1]} (x_*(t) - \hat{X}_{*n}(t)) > \varepsilon \right) \\
&\qquad = \mathsf{P}(\exists t \in [0,1]\ \hat{X}_{*n}(t) < x_*(t) - \varepsilon) \leq \mathsf{P}(\exists t \in [0,1]\ \forall i\ I_{ni}^*(t) = 0) \\
&\qquad\qquad + \mathsf{P}\left(\exists t \in [0,1]\ \exists i\ \exists j\ (I_{nij}(t) = 1,\ X_i(T_{ij}) < x_*(t) - \varepsilon)\right).
\end{aligned}$$

Similarly as before, we get that

$$\begin{aligned}
\mathsf{P}(\exists t \in [0,1]\ \forall i\ I_{ni}^*(t) = 0) &\leq \sum_{s=0}^{l_n - 1} \mathsf{P}(\forall i\ I_{ni}^*(sh_n) = 0) = \sum_{s=0}^{l_n-1} [\mathsf{P}(I_{n1}^*(sh_n) = 0)]^n \\
&= \sum_{s=0}^{l_n-1} [\mathsf{E}(1-h_n)^M]^n \leq n(1-h_n)^n
\end{aligned}$$

and

$$\begin{aligned}
&\mathsf{P}(\exists t \in [0,1]\ \exists i\ \exists j\ (I_{nij}(t) = 1,\ X_i(T_{ij}) < x_*(t) - \varepsilon) \\
&\leq \mathsf{P}(\exists t \in [0,1]\ \exists i\, C_i < (c_* a_n^*(t) - \varepsilon)/a_{*n}(t) + (b_n^*(t) - b_{*n}(t))/a_{*n}(t)) \leq \mathsf{P}(\exists i\, C_i < c_* - \varepsilon') = 0.
\end{aligned}$$

Therefore, for all $n \geq n_2$

$$\mathsf{P}\left(\sup_{t\in[0,1]} (x_*(t) - \hat{X}_{*n}(t)) > \varepsilon \right) \leq n(1-h_n)^n.$$

Step 3. From Step 1 and Step 2 we get that for all $n \geq \max(n_1, n_2)$

$$\mathsf{P}\left(\sup_{t\in[0,1]} \left|\hat{X}_{*n}(t) - x_*(t)\right| > \varepsilon \right) \leq n[1-(1-q)h_n]^n + n(1-h_n)^n \leq 2n[1-(1-q)h_n]^n.$$

Similarly as before, from here we get that almost surely

$$\sup_{t\in[0,1]} \left|\hat{X}_{*n}(t) - x_*(t)\right| \to 0.$$

□

C.2 Proof of Theorem 4.2

Proof. Recall that

$$\hat{X}_{ni}(t) = (1-\hat{\alpha}_{ni})\hat{X}_{*n}(t) + \hat{\alpha}_{ni}\hat{X}_n^*(t),$$

where $\hat{\alpha}_{ni}$ are defined by the following equation:

$$X_i(\bar{T}_{ni}) = (1-\hat{\alpha}_{ni})\hat{X}_{*n}(\bar{T}_{ni}) + \hat{\alpha}_{ni}\hat{X}_n^*(\bar{T}_{ni}),$$

and $\bar{T}_{ni}$ is any time point from the set $\{T_{i1},\dots,T_{iM_i}\}$ (for example, the smallest element of that set $T_{i,(1)}$).

If $\hat{X}_n^*(\bar{T}_{ni}) - \hat{X}_{*n}(\bar{T}_{ni}) > 0$, that equation uniquely defines $\hat{\alpha}_{ni}$:

$$\hat{\alpha}_{ni} = \frac{X_i(\bar{T}_{ni}) - \hat{X}_{*n}(\bar{T}_{ni})}{\hat{X}_n^*(\bar{T}_{ni}) - \hat{X}_{*n}(\bar{T}_{ni})}.$$

Denote

$$Z_{ni}(t) = (1-\hat{\alpha}_{ni})x_*(t) + \hat{\alpha}_{ni}x^*(t).$$

Then almost surely

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^n \|\hat{X}_{ni} - Z_{ni}\| &\le \frac{1}{n}\sum_{i=1}^n \left((1-\hat{\alpha}_{ni})\|\hat{X}_{*n} - x_*\| + \hat{\alpha}_{ni}\|\hat{X}_n^* - x^*\|\right) \\ &\le \|\hat{X}_{*n} - x_*\| + \|\hat{X}_n^* - x^*\| \to 0\end{aligned}$$

and it is enough to prove that

$$\frac{1}{n}\sum_{i=1}^n \|Z_{ni} - X_i\| \to 0.$$

But

$$\|Z_{ni} - X_i\| = \left|(1-\hat{\alpha}_{ni})c_* + \hat{\alpha}_{ni}c^* - C_i\right| \|a\|.$$

Thus it is enough to prove that almost surely

$$\frac{1}{n}\sum_{i=1}^{n}\left|(1-\hat{\alpha}_{ni})c_* + \hat{\alpha}_{ni}c^* - C_i\right| \to 0.$$

We will prove that this holds with any ω for which

$$\|\hat{X}_n^*(\cdot,\omega) - x^*\| + \|\hat{X}_{*n}(\cdot,\omega) - x_*\| \to 0.$$

In the following, we will assume that such ω is fixed and we will omit it for the sake of convenience.

Take any $\delta < (c^* - c_*)a_*$. Because

$$|\hat{X}_n^*(\bar{T}_{ni}) - \hat{X}_{*n}(\bar{T}_{ni}) - x^*(\bar{T}_{ni}) + x_*(\bar{T}_{ni})| \le \|\hat{X}_n^* - x^*\| + \|\hat{X}_{*n} - x_*\| \to 0$$

and

$$x^*(\bar{T}_{ni}) - x_*(\bar{T}_{ni}) \ge (c^* - c_*)a_* > \delta,$$

there exists n_0 such that, for all $n \ge n_0$ and all $i = 1,\dots,n$

$$\hat{X}_n^*(\bar{T}_{ni}) - \hat{X}_{*n}(\bar{T}_{ni}) > \delta.$$

Then for all $n \ge n_0$ and all $i = 1,\dots,n$

$$\begin{aligned}
&(1-\hat{\alpha}_{ni})c_* + \hat{\alpha}_{ni}c^* - C_i \\
&= \frac{\hat{X}_n^*(\bar{T}_{ni}) - X_i(\bar{T}_{ni})}{\hat{X}_n^*(\bar{T}_{ni}) - \hat{X}_{*n}(\bar{T}_{ni})}c_* + \frac{X_i(\bar{T}_{ni}) - \hat{X}_{*n}(\bar{T}_{ni})}{\hat{X}_n^*(\bar{T}_{ni}) - \hat{X}_{*n}(\bar{T}_{ni})}c^* - C_i \\
&= \frac{\hat{X}_n^*(\bar{T}_{ni})c_* - \hat{X}_{*n}(\bar{T}_{ni})c^* + (c^* - c_*)b(\bar{T}_{ni})}{\hat{X}_n^*(\bar{T}_{ni}) - \hat{X}_{*n}(\bar{T}_{ni})} \\
&+ \frac{(c^* - c_*)(X_i(\bar{T}_{ni}) - b(\bar{T}_{ni})) - C_i(\hat{X}_n^*(\bar{T}_{ni}) - \hat{X}_{*n}(\bar{T}_{ni}))}{\hat{X}_n^*(\bar{T}_{ni}) - \hat{X}_{*n}(\bar{T}_{ni})}.
\end{aligned}$$

Moreover,

$$\frac{|\hat{X}_n^*(\bar{T}_{ni})c_* - \hat{X}_{*n}(\bar{T}_{ni})c^* + (c^* - c_*)b(\bar{T}_{ni})|}{\hat{X}_n^*(\bar{T}_{ni}) - \hat{X}_{*n}(\bar{T}_{ni})} \leq \delta^{-1}|\hat{X}_n^*(\bar{T}_{ni})c_* - \hat{X}_{*n}(\bar{T}_{ni})c^* + (c^* - c_*)b(\bar{T}_{ni})|$$
$$\leq \delta^{-1}(|\hat{X}_n^*(\bar{T}_{ni}) - c^* a(\bar{T}_{ni}) - b(\bar{T}_{ni})|c_* + |c_* a(\bar{T}_{ni}) + b(\bar{T}_{ni}) - \hat{X}_{*n}(\bar{T}_{ni})|c^*)$$
$$\leq \delta^{-1} c^* (\|\hat{X}_n^* - x^*\| + \|\hat{X}_{*n} - x_*\|)$$

and

$$\frac{|(c^* - c_*)(X_i(\bar{T}_{ni}) - b(\bar{T}_{ni})) - C_i(\hat{X}_n^*(\bar{T}_{ni}) - \hat{X}_{*n}(\bar{T}_{ni}))|}{\hat{X}_n^*(\bar{T}_{ni}) - \hat{X}_{*n}(\bar{T}_{ni})}$$
$$\leq \delta^{-1}|(c^* - c_*)(X_i(\bar{T}_{ni}) - b(\bar{T}_{ni})) - C_i(\hat{X}_n^*(\bar{T}_{ni}) - \hat{X}_{*n}(\bar{T}_{ni}))|$$
$$= \delta^{-1} C_i |x^*(\bar{T}_{ni}) - x_*(\bar{T}_{ni}) - \hat{X}_n^*(\bar{T}_{ni}) + \hat{X}_{*n}(\bar{T}_{ni})|$$
$$\leq \delta^{-1} c^* (\|\hat{X}_n^* - x^*\| + \|\hat{X}_{*n} - x_*\|).$$

Thus, for all $n \geq n_0$

$$\frac{1}{n}\sum_{i=1}^{n} \left|(1 - \hat{\alpha}_{ni})c_* + \hat{\alpha}_{ni}c^* - C_i\right| \leq 2\delta^{-1} c^* (\|\hat{X}_n^* - x^*\| + \|\hat{X}_{*n} - x_*\|) \to 0.$$

□

C.3 Proof of Theorem 4.3

Fix i. Then

$$\sup_{t\in[0,1]} |\tilde{X}_i(t) - X_i(t)| = |\tilde{X}_i(t) - X_i(t)|$$

for some t that falls into some interval $[sh_n; (s+1)h_n]$. Denote that t by T_i. Then

$$|\tilde{X}_i(T_i) - X_i(T_i)| = |\tilde{X}_i(T_i) - \tilde{X}_i(sh_n) + \tilde{X}_i(sh_n) - X_i(T_i)|$$
$$\leq |\tilde{X}_i(T_i) - \tilde{X}_i(sh_n)| + |\tilde{X}_i(sh_n) - X_i(T_i)|.$$

Note that

$$\begin{aligned}
|\tilde{X}_i(T_i) - \tilde{X}_i(sh_n)| &\le |\hat{X}_i(sh_n) - \hat{X}_i((s+1)h_n)| \\
&= |\hat{X}_i(sh_n) - X_i(sh_n) + X_i(sh_n) - X_i((s+1)h_n) + X_i((s+1)h_n) - \hat{X}_i((s+1)h_n)| \\
&\le 2\|\hat{X}_i - X_i\| + C_i|a(sh_n) - a((s+1)h_n)| + |b(sh_n) - b((s+1)h_n)| \\
&\le 2\|\hat{X}_i - X_i\| + C_i\|a_n^* - a_{*n}\| + \|b_n^* - b_{*n}\| \\
&\le 2\|\hat{X}_i - X_i\| + c^*\|a_n^* - a_{*n}\| + \|b_n^* - b_{*n}\|.
\end{aligned}$$

Therefore,

$$|\tilde{X}_i(T_i) - X_i(T_i)| \le 3\|\hat{X}_i - X_i\| + c^*\|a_n^* - a_{*n}\| + \|b_n^* - b_{*n}\|.$$

Then

$$\frac{1}{n}\sum_{i=1}^{n}\|\tilde{X}_{ni} - X_i\| \le 3\frac{1}{n}\sum_{i=1}^{n}\|\hat{X}_i - X_i\| + c^*\|a_n^* - a_{*n}\| + \|b_n^* - b_{*n}\|.$$

By Lemma 1, $\|a_n^* - a_{*n}\| \to 0$, $\|b_n^* - b_{*n}\| \to 0$, while by Theorem 4.2, almost surely $\frac{1}{n}\sum_{i=1}^{n}\|\hat{X}_i - X_i\| \to 0$. Therefore, almost surely

$$\frac{1}{n}\sum_{i=1}^{n}\|\tilde{X}_{ni} - X_i\| \to 0.$$

Bibliography

[1] L. K. Bacharack, T. J. Hastie, M. C. Wang, B. Narasimhan, and R. Marcus. Bone mineral acquisition in healthy Asian, Hispanic, Black and Caucasian youth: A longitudinal study. *J. Clin. Endocrin. Metab.*, 84(12):4702–4712, 1999. 9, 43, 44

[2] A. W. van der Vaart. *Asymptotic Statistics*. Cambridge University Press, 2000. 10, 61, 78, 80

[3] A. Delaigle and P. Hall. Classification Using Censored Functional Data. *J. Am. Statist. Ass.*, 108(504):1269–1283, 2013. 11, 45, 46, 47, 53, 54, 55, 56, 63

[4] A. Cuevas. A partial overview of the theory of statistics with functional data. *J. Statist. Plann. Inference*, 147:1–23, 2014. 13

[5] L. Devroye, L. Györfi, and G. Lugosi. *A probabilistic theory of pattern recognition*. Springer, 1996. 14

[6] A. Delaigle and P. Hall. Approximating fragmented functional data by segments of Markov chains. *Biometrika*, 103(4):779–799, 2016. 15, 46, 63

[7] J. van Ryzin. Bayes risk consistency of classification procedures using density estimation. *Sankhya: The Indian Journal of Statistics, Series A*, 28(2/3):261–270, 1966. 15, 33

[8] A. M. Alonso, D. Casado, and J. Romo. Supervised classification for functional data: A weighted distance approach. *Comput. Statist. Data Anal.*, 56(7):2334–2346, 2012. 15

[9] A. M. Aguilera, M. Escabias, and M. J. Valderrama. Using principal components for estimating logistic regression with high-dimensional multicollinear data. *Comput. Statist. Data Anal.*, 50(8):1905–1924, 2006. 16

[10] H.-G. Müller and U. Stadtmüller. Generalized Functional Linear Models. *Ann. Statist.*, 33(2):774–805, 2005. 17, 21, 32, 33, 41, 61

[11] F. Yao, H.-G. Müller, and J.-L. Wang. Functional data analysis for sparse longitudinal data. *J. Am. Statist. Assoc.*, 100(470):577–590, 2005. 17, 45, 46, 47, 54, 59, 62

[12] A. Albert and J. A. Anderson. On the existence of maximum likelihood estimates in logistic regression models. *Biometrika*, 71(1):1–10, 1984. 19, 24, 25, 77

[13] A. Christmann and P. J. Rousseeuw. Measuring overlap in binary regression. *Comput. Statist. Data Anal.*, 37(1):65–75, 2001. 20

[14] D. Firth. Bias reduction of maximum likelihood estimates. *Biometrika*, 80(1):27–38, 1993. 20

[15] S. Gao and J. Shen. Asymptotic properties of a double penalized maximum likelihood estimator in logistic regression. *Statist. Probab. Lett.*, 77(9):925–930, 2007. 20

[16] P. J. Rousseeuw and A. Christmann. Robustness against separation and outliers in logistic regression. *Comput. Statist. Data Anal.*, 43(3):315–332, 2003. 20

[17] P. Fu, A. Panneerselvam, B. Clifford, A. Dowlati, P. C. Ma, G. Zheng, B. Halmos, and R. S. Leidner. Simpsons paradoxaggregating and partitioning populations in health disparities of lung cancer patients. *Stat. Methods Med. Res.*, 24(6):937–948, 2015. 20

[18] R. Sauter and L. Held. Quasi-complete separation in random effects of binary response mixed models. *J. Stat. Comput. Simul.*, 86(14):2781–2796, 2016. 20

[19] L. Held and R. Sauter. Adaptive prior weighting in generalized regression. *Biometrics*, 2016. DOI:10.1111/biom.12541. 20

[20] J. O. Ramsay and B. W. Silverman. *Applied Functional Data Analysis: Methods and Case Studies*. Springer, 2002. 20

[21] J. O. Ramsay and B. W. Silverman. *Functional data analysis*. Springer, 2005. 20

[22] M. Escabias, A. M. Aguilera, and M. J. Valderrama. Functional PLS logit regression model. *Comput. Statist. Data Anal.*, 51(10):4891–4902, 2007. 20, 31

[23] M. Escabias, A. M. Aguilera, and M. J. Valderrama. Principal component estimation of functional logistic regression: discussion of two different approaches. *J. Nonparametr. Stat.*, 16(3-4):365–384, 2004. 20

[24] A. M. Aguilera, M. Escabias, and M. J. Valderrama. Discussion of different logistic models with functional data. Application to Systemic Lupus Erythematosus. *Comput. Statist. Data Anal.*, 53(1):151–163, 2008. 20

[25] M. Denhere and N. Billor. Robust Principal Component Functional Logistic Regression. *Comm. Statist. Simulation Comput.*, 45(1):264–281, 2016. 20

[26] G. M. James. Generalized linear models with functional predictors. *J. R. Stat. Soc. Ser. B. Stat. Methodol.*, 64(3):411–432, 2002. 20

[27] M. C. Aguilera-Morillo, A. M. Aguilera, M. Escabias, and M. J. Valderrama. Penalized spline approaches for functional logit regression. *Test*, 22(2):251–277, 2013. 21

[28] H.-G. Müller. Functional Modelling and Classification of Longitudinal Data. *Scand. J. Stat.*, 32(2):223–240, 2005. 21

[29] K. Chen, I. Hu, and Z. Ying. Strong consistency of maximum quasi-likelihood estimators in generalized linear models with fixed and adaptive designs. *Ann. Statist.*, 27(4):1155–1163, 1999. 30

[30] A. Kazakeviciute and M. Olivo. A study of logistic classifier: uniform consistency in finite-dimensional linear spaces. *Journal of Mathematics, Statistics and Operations Research*, 3(2):1–7, 2016. 30

[31] S. A. van de Geer. High-dimensional generalized linear models and the Lasso. *Ann. Statist.*, 36(2):614–645, 2008. 32

[32] J. Fan and R. Song. Sure independence screening in generalized linear models with NP-dimensionality. *Ann. Statist.*, 38(6):3567–3604, 2010. 32

[33] L. Wang. GEE analysis of clustered binary data with diverging number of covariates. *Ann. Statist.*, 39(1):389–417, 2011. 32

[34] H. Liang and P. Du. Maximum likelihood estimation in logistic regression models with a diverging number of covariates. *Electron. J. Stat.*, 6:1838–1846, 2012. 32

[35] G. M. James and T. J. Hastie. Functional linear discriminant analysis for irregularly sampled curves. *J. R. Stat. Soc. Ser. B. Stat. Methodol.*, 63(3):533–550, 2001. 44, 63

[36] G. M. James, T. J. Hastie, and C. A. Sugar. Principal component models for sparse functional data. *Biometrika*, 87(3):587–602, 2000. 45

[37] G. M. James and C. A. Sugar. Clustering for Sparsely Sampled Functional Data. *J. Am. Statist. Assoc.*, 98(462):397–408, 2003. 45

[38] D. Liebl. Modeling and forecasting electricity spot prices: A functional data perspective. *Ann. Appl. Statist.*, 7(3):1562–1592, 2013. 45

[39] Y. Goldberg, Y. Ritov, and A. Mandelbaum. Predicting the continuation of a function with applications to call center data. *J. Statist. Plan. Infer.*, 147:53–65, 2014. 45

[40] D. Kraus. Components and completion of partially observed functional data. *J. R. Stat. Soc. Ser. B. Stat. Methodol.*, 77(4):777–801, 2015. 45

[41] O. Kallenberg. *Foundations of Modern Probability*. Springer, 2nd edition, 2001. 70, 71, 72, 73

[42] D. P. Bertsekas, A. Nedic, and A. E. Ozdaglar. *Convex Analysis and Optimization*. Athena Scientific, 2003. 74, 75