Robust Semi-Parametric Inference in Semi-Supervised Settings

Robust Semi-Parametric Inference in Semi-Supervised Settings

A dissertation presented

by

Abhishek Chakrabortty

to

The Department of Biostatistics

in partial fulfillment of the requirements
for the degree of
Doctor of Philosophy
in the subject of
Biostatistics

Harvard University
Cambridge, Massachusetts

May 2016

Dissertation Advisor: Professor Tianxi Cai Abhishek Chakrabortty

Robust Semi-Parametric Inference in Semi-Supervised Settings

Abstract

In this dissertation, we consider semi-parametric estimation problems under semi-supervised (SS) settings, wherein the available data consists of a small or moderate sized labeled data (L), and a much larger unlabeled data (U). Such data arises naturally from settings where the outcome, unlike the covariates, is expensive to obtain, a frequent scenario in modern studies involving large electronic databases. It is often of interest in SS settings to investigate if and when U can be exploited to improve estimation efficiency, compared to supervised estimators based on L only.

In Chapter 1, we propose a class of Efficient and Adaptive Semi-Supervised Estimators (EASE) for linear regression. These are semi-non-parametric imputation based two-step estimators adaptive to model mis-specification, leading to improved efficiency under model mis-specification, and equal (optimal) efficiency when the linear model holds. This adaptive property is crucial for advocating safe use of U. We provide asymptotic results establishing our claims, followed by simulations and application to real data.

In Chapter 2, we provide a unified framework for SS M-estimation problems based on general estimating equations, and propose a family of EASE estimators that are always as efficient as the supervised estimator and more efficient whenever U is actually informative for the parameter of interest. For a subclass of problems, we also provide a flexible semi-non-parametric imputation strategy for constructing EASE. We provide asymptotic results establishing our claims, followed by simulations and application to real data.

In Chapter 3, we consider regressing a binary outcome (Y) on some covariates ($\mathbf{X}$) based

on a large unlabeled data with observations only for $\mathbf{X}$, and additionally, a surrogate (S) which can predict Y with high accuracy when it assumes extreme values. Assuming Y and S both follow single index models versus $\mathbf{X}$, we show that under sparsity assumptions, we can recover the regression parameter of Y versus $\mathbf{X}$ through a least squares LASSO estimator based on the subset of the data restricted to the extreme sets of S with Y imputed using the surrogacy of S. We provide sharp finite sample performance guarantees for our estimator, followed by simulations and application to real data.

Contents

Acknowledgments

I owe my deepest gratitude to my dissertation advisor, Professor Tianxi Cai, for all her invaluable support and guidance throughout. I am really thankful to Tianxi for introducing me to several challenging problems, interesting from both theoretical and practical perspectives, and giving me the encouragement and freedom to work on them, and also for being very generous and patient whenever I have got stuck and/or lazy with my work :-)

I would like to sincerely thank my dissertation committee members Professor James Robins and Professor Eric Tchetgen Tchetgen for all their helpful comments and suggestions, and their general enthusiasm regarding my research. I would especially like to thank Jamie for his valuable insights regarding the connection of the semi-supervised learning problem to missing data problems and semi-parametric theory, which turned out to be immensely helpful for me to understand the problem better. I would also like to thank Professor Nan Laird and Dr. Judith Lok for all their assistance during the early stages of my graduate life. A special thanks to Judith also for all the wonderful and enriching experiences I have had as a teaching assistant under her instructorship.

This acknowledgement would be incomplete without the mentions of Rajarshi Mukherjee and Kaustubh Adhikari, whom I have known since my ISI days. My heartfelt thanks to both of them, and also to Wan-Chen Lee (Chicky), for all their assistance in helping me settle in comfortably, as well as make my life in Boston over the last few years thoroughly enjoyable. Without going into specific names, as the list would be too long, I would also like to take this opportunity to thank all my friends, as well as the staff, at Harvard Biostatistics, whom I have got to know over the years through innumerable chats/gossips/academic discussions. Thank you all for making my time at the department truly wonderful.

This dissertation is dedicated to my parents, Dalia Chakrabortty and Jagannath Chakrabortty, the two most loving and influential people in my life. They have always been there for me, and believed in me and supported me even in the most difficult situa-

tions. They never stop learning from me about my research, perhaps without understanding much of it :-) I owe almost everything in my life to them. As a small gesture of gratitude, I therefore dedicate this dissertation to them. Last but not the least, this dissertation would not have been possible without Debarati Ghose, in whom I have been fortunate enough to find a wonderful combination of a great companion and a patient listener. I can never thank her enough for being with me through the thick and thin of my graduate life, and for being my support during the most difficult times, despite having a very difficult last year herself. Thank you so much to all of you!

Efficient and Adaptive Linear Regression in Semi-Supervised Settings

Abhishek Chakrabortty and Tianxi Cai

Department of Biostatistics

Harvard University

1.1 Summary

We consider the linear regression problem under semi-supervised settings wherein the available data typically consists of: (i) a small or moderate sized 'labeled' data, and (ii) a much larger sized 'unlabeled' data. Such data arises naturally from settings where the outcome, unlike the covariates, is expensive to obtain, a frequent scenario in modern studies involving large databases like electronic medical records (EMR). Supervised estimators like the ordinary least squares (OLS) estimator utilize only the labeled data. It is often of interest to investigate if and when the unlabeled data can be exploited to improve estimation of the regression parameter in the adopted linear model.

In this paper, we propose a class of 'Efficient and Adaptive Semi-Supervised Estimators' (EASE) to improve estimation efficiency over OLS. The proposed estimators are two-step estimators adaptive to model mis-specification, thus leading to improved efficiency under model mis-specification, and equal (optimal) efficiency when the linear model holds. This adaptive property, often unaddressed in the existing literature, is quite crucial for advocating 'safe' use of the unlabeled data. The construction of EASE primarily involves: a flexible 'semi-non-parametric' imputation step for imputing the outcomes in the unlabeled data, followed by simply fitting a linear model to the imputed unlabeled data. The imputation step involves: a smoothing step that works well even when the number of covariates is not small (through use of dimension reduction techniques, if needed), and a follow up 'refitting' step along with a cross-validation (CV) strategy that are employed to address under-smoothing and over-fitting, two issues often encountered in two-step estimators involving a first-step smoothing. We establish asymptotic results including consistency, asymptotic normality and the adaptive properties of EASE. We also provide influence function expansions and a 'double' CV strategy for consistent variance estimation. The results are further validated through extensive finite sample simulations followed by application to a real dataset from an EMR study of autoimmune diseases.

1.2 Introduction

In recent years, semi-supervised learning (SSL) has emerged as an exciting new area of research in statistics and machine learning. A detailed discussion on SSL including its practical relevance, the primary question of interest in SSL, and the existing relevant literature can be found in Chapelle et al. (2006) and Zhu (2008). A typical semi-supervised (SS) setting, as represented in Figure 1.1, is characterized by two types of available data: (i) a small or moderate sized 'labeled' data, $\mathcal{L}$, containing observations for both an outcome Y and a set of covariates $\mathbf{X}$ of interest, and (ii) an 'unlabeled' data, $\mathcal{U}$, of *much larger* size but having observations *only* for the covariates $\mathbf{X}$. By virtue of its large size, $\mathcal{U}$ essentially gives us the distribution of $\mathbf{X}$, denoted henceforth by $\mathbb{P}_{\mathbf{X}}$. Such a setting arises naturally whenever the covariates are easily available so that unlabeled data is plentiful, but the outcome is costly or difficult to obtain, thereby limiting the size of $\mathcal{L}$. This scenario is directly relevant to a variety of practical problems, especially in the modern 'big data' era, with massive unlabeled datasets (often electronically recorded) becoming increasingly available and tractable. A few familiar examples include machine learning problems like text mining, web page classification, speech recognition, natural language processing etc.

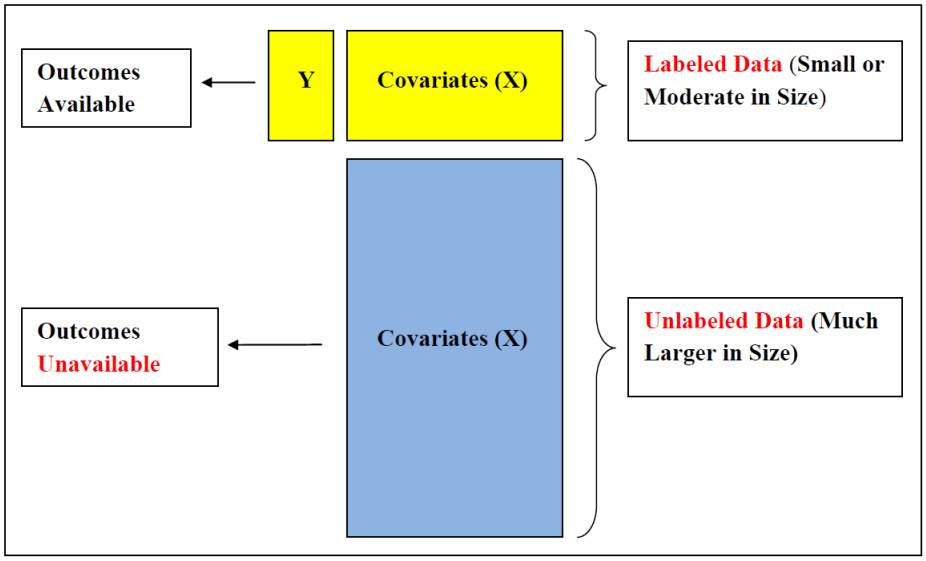


Figure 1.1: Schematic representation of a typical semi-supervised setting.

Among biomedical applications, a particularly interesting problem where SSL can be of great use is the statistical analysis of electronic medical records (EMR) data. Endowed with a wealth of de-identified clinical and phenotype data for large patient cohorts, EMR linked with bio-repositories are increasingly gaining popularity as rich resources of data for discovery research (Kohane, 2011). Such large scale datasets obtained in a cost-effective and timely manner are of great importance in modern medical research for addressing important questions such as the biological role of genetic variants in disease susceptibility and progression (Kohane, 2011). However, one major bottleneck impeding EMR driven research is the difficulty in obtaining validated phenotype information (Liao et al., 2010) since they are labor intensive or expensive to obtain. Thus, gold standard labels and genomic measurements are typically available only for a small subset nested within a large cohort. In contrast, digitally recorded data on the clinical variables are often available on all subjects, highlighting the necessity and utility of developing robust SSL methods that can leverage such rich source of auxiliary information to improve phenotype definition and estimation precision.

SSL primarily distinguishes from standard supervised methods by making use of $\mathcal{U}$, an information that is ignored by the latter. The ultimate question of interest in SSL is to investigate if and when this information can be exploited to improve the efficiency over a given supervised approach. It is important to note that while the SS set-up can be viewed as a missing data problem, it is quite different from a standard missing data setting as the probability of missingness tends to 1 in SSL (so that the 'positivity assumption' typically assumed in the classical missing data literature is violated here). Interestingly, characterization of the missingness mechanism, although quite crucial, has often stayed implicit in the SSL literature (Lafferty and Wasserman, 2007). Nevertheless, it has mostly been assumed as 'missing completely at random' which is typically the case, with the labeled data being obtained from labeling a random subset, selected by design, from a large unlabeled data. It is also worth noting that the analysis of SS settings under more general missingness mechanisms is considerably more complicated due to the violation of the positivity assumption.

The theoretical underpinnings of SSL including its scope and the consequences of using the unlabeled data have been studied to some extent by Castelli and Cover (1995, 1996) for classification problems and later, more generally by Lafferty and Wasserman (2007), where it has been noted that SSL can improve efficiency only if $\mathbb{P}_{\mathbf{X}}$ and the conditional distribution of Y given $\mathbf{X}$, denoted henceforth by $\mathbb{P}_{Y|\mathbf{X}}$, are somehow related. In recent years, several graph based non-parametric SSL approaches have been proposed (Zhu, 2005; Belkin et al., 2006) relying, implicitly or explicitly, on assumptions relating $\mathbb{P}_{\mathbf{X}}$ to $\mathbb{P}_{Y|\mathbf{X}}$. These assumptions have been characterized more formally in Lafferty and Wasserman (2007).

Simple parametric modeling, often appealing due to its interpretability, however has been a little less studied in SSL. Perhaps the most well-known method in the literature is the 'generative model' approach for classification problems (Nigam et al., 2000; Nigam, 2001) which is based on modeling the joint distribution of $(Y, \mathbf{X})$ as an identifiable mixture of parametric models, thereby implicitly relating $\mathbb{P}_{Y|\mathbf{X}}$ and $\mathbb{P}_{\mathbf{X}}$. However, these approaches depend strongly on the validity of the assumed mixture model, violation of which can actually *degrade* their performance compared to the supervised approach (Cozman and Cohen, 2001; Cozman et al., 2003). Recently, Culp (2013) proposed SS methods for regularized linear regression. However, no theoretical properties were provided for their method that could desirably guarantee that it is always at least as efficient as the supervised counterpart.

In general, if the assumed working model for $\mathbb{P}_{Y|\mathbf{X}}$ is correct and the parameter of interest is not related to $\mathbb{P}_{\mathbf{X}}$, then one *cannot* possibly gain through SSL by using the knowledge of $\mathbb{P}_{\mathbf{X}}$ (Zhang and Oles, 2000; Seeger, 2002). On the other hand, under model mis-specification, the target parameter may inherently *depend* on $\mathbb{P}_{\mathbf{X}}$, and thus imply the potential utility of $\mathcal{U}$ in improving the estimation. However, inappropriate usage of $\mathcal{U}$ may lead to degradation of the estimation precision. This therefore signifies the need for *robust* and efficient SS estimators that are *adaptive* to model mis-specification, so that they are as efficient as the supervised estimator under the correct model and more efficient under model mis-specification. To the best of our knowledge, work done along these lines is relatively scarce in the SSL literature,

one notable exception being the recent paper by Kawakita and Kanamori (2013) where they propose such adaptive SS estimators using density ratio methods that rely on basis function expansions. However, no data adaptive procedure was provided for selecting the bases, which could greatly impact the amount of the efficiency gain. Lastly, most existing methods rely on asymptotic results without accounting for the finite sample over-fitting bias, which can significantly affect the finite sample performance of the estimators, as well as also pose challenges in interval estimation and inference based on these estimators.

To address these questions, we propose here a class of Efficient and Adaptive Semi-Supervised Estimators (EASE) in the context of linear regression problems. We essentially adopt a semi-parametric perspective wherein the adopted linear 'working' model can be potentially mis-specified, and the goal is to obtain efficient and adaptive SS estimators of the regression parameter through robust usage of $\mathcal{U}$. The EASE estimators are two-step estimators with a simple and scalable construction based on a first step of 'semi-non-parametric' (SNP) imputation which includes a smoothing step and a follow-up 'refitting' step. In the second step, we regress the imputed outcomes against the covariates using the unlabeled data to obtain our SNP imputation based SS estimator, and then further combine it optimally with the supervised estimator to obtain the final EASE estimator. Dimension reduction methods are also employed in the smoothing step to accommodate higher dimensional $\mathbf{X}$, if necessary. Further, we extensively adopt cross-validation (CV) techniques in the imputation, leading to some interesting and useful theoretical properties typically not observed for smoothing based two-step estimators. We demonstrate that EASE is guaranteed to be efficient and adaptive in the sense discussed above, and further achieves semi-parametric optimality whenever the SNP imputation is 'sufficient' or the linear model holds. We also provide data adaptive methods to optimally select the directions for smoothing when dimension reduction is needed.

The rest of this paper is organized as follows. In section 1.3, we formally introduce and formulate the SS linear regression problem. In section 1.4, we construct a family of SS

estimators based on SNP imputation and establish all their properties, and further propose the EASE as a refinement of these estimators. For all our proposed estimators, we also address their associated inference procedures based on 'double' CV methods. In section 1.5, we discuss a kernel smoothing based implementation of the SNP imputation and establish all its properties. In section 1.6, we discuss SS dimension reduction techniques, useful for implementing the SNP imputation. Simulation results and an application to an EMR study are shown in section 1.7, followed by a concluding discussion in section 1.8. Proofs of all the theoretical results are given in Appendix A.

1.3 Problem Set-up

1.3.1 Preliminaries

Data Representation: Let $Y \in \mathbb{R}$ denote the outcome variable and $\mathbf{X} \in \mathbb{R}^p$ denote the covariate vector, where p is fixed, and let $\mathbf{Z} = (Y, \mathbf{X}')'$. Then the entire data available for analysis can be represented as $\mathbb{S} = (\mathcal{L} \cup \mathcal{U})$, where $\mathcal{L} = \{\mathbf{Z}_i \equiv (Y_i, \mathbf{X}_i')' : i = 1, ..., n\}$ consists of n independent and identically distributed (i.i.d.) observations from the joint distribution $\mathbb{P}_{\mathbf{Z}}$ of $\mathbf{Z}$, $\mathcal{U} = \{\mathbf{X}_j : j = 1, ..., N\}$ consists of N i.i.d. realizations of $\mathbf{X}$, and $\mathcal{L} \perp\!\!\!\perp \mathcal{U}$.

Basic Assumptions: (a) We assume that $N \gg n$ and hence as $n \to \infty$, the proportion of observed outcomes, n/N, tends to 0, which makes SSL different from a classical missing data problem where this proportion is typically assumed to be bounded away from zero. (b) The underlying Y for subjects in $\mathcal{U}$ are assumed to be 'missing completely at random', so that $\mathbf{Z} \sim \mathbb{P}_{\mathbf{Z}}$ for all subjects in $\mathcal{S}$. (c) We assume throughout that $\mathbf{Z}$ has finite 2^{nd} moments, $\text{Var}(\mathbf{X})$ is positive definite and $\mathbf{X}$ has a compact support $\mathcal{X} \subseteq \mathbb{R}^p$. Let $\mathcal{L}_2(\mathbb{P}_{\mathbf{X}})$ denote the space of all $\mathbb{R}$-valued measurable functions of $\mathbf{X}$ having finite L_2 norm with respect to (w.r.t.) $\mathbb{P}_{\mathbf{X}}$. (d) Since the moments of $\mathbf{X}$ are essentially known due to the large (potentially infinite) size of $\mathcal{U}$, we also assume without loss of generality (w.l.o.g.) that $\mathbb{E}(\mathbf{X}) = \mathbf{0}$ and $\text{Var}(\mathbf{X}) = I_p$, where I_p denotes the $(p \times p)$ identity matrix.

1.3.2 The Target Parameter and Its Supervised Estimator

We consider the linear regression *working model* given by:

$$Y = \overrightarrow{\mathbf{X}}'\boldsymbol{\theta} + \epsilon, \quad \text{with} \quad \mathbb{E}(\epsilon \mid \mathbf{X}) = 0, \tag{1.1}$$

where, for any vector $\mathbf{x}$, $\overrightarrow{\mathbf{x}} = (1, \mathbf{x}')' \in \mathbb{R}^{(p+1)}$ and $\boldsymbol{\theta} \in \mathbb{R}^{(p+1)}$ is an unknown regression parameter. Accounting for the potential mis-specification of the working model (1.1), we define the target parameter of interest as a model free parameter, as follows:

Definition 1.1. The target parameter $\boldsymbol{\theta}_0 = (\alpha_0, \boldsymbol{\beta}_0')'$ for linear regression may be defined as the solution to the normal equations: $\mathbb{E}\{\overrightarrow{\mathbf{X}}(Y - \overrightarrow{\mathbf{X}}'\boldsymbol{\theta})\} = \mathbf{0}$ in $\boldsymbol{\theta} \in \mathbb{R}^{(p+1)}$, or equivalently, $\boldsymbol{\theta}_0$ may be defined as: $\boldsymbol{\theta}_0 = \underset{\boldsymbol{\theta} \in \mathbb{R}^{(p+1)}}{\arg\min}\, \mathbb{E}(Y - \overrightarrow{\mathbf{X}}'\boldsymbol{\theta})^2$.

Existence and uniqueness of $\boldsymbol{\theta}_0$ in 1.1 is clear. Further, $\overrightarrow{\mathbf{X}}'\boldsymbol{\theta}_0$ is the L_2 projection of $\mathbb{E}(Y|\mathbf{X}) \in \mathcal{L}_2(\mathbb{P}_{\mathbf{X}})$ onto the subspace of all linear functions of $\mathbf{X}$ and hence, is the best linear predictor of Y given $\mathbf{X}$. The linear model (1.1) is *correct* (else, *mis-specified*) if and only if $\mathbb{E}(Y|\mathbf{X})$ lies in this space (in which case $\mathbb{E}(Y|\mathbf{X}) = \overrightarrow{\mathbf{X}}'\boldsymbol{\theta}_0$). When the model is correct, $\boldsymbol{\theta}_0$ depends only on $\mathbb{P}_{Y|\mathbf{X}}$, not on $\mathbb{P}_{\mathbf{X}}$. Hence, improved estimation of $\boldsymbol{\theta}_0$ through SSL is impossible in this case unless further assumptions relating $\boldsymbol{\theta}_0$ to $\mathbb{P}_{\mathbf{X}}$ are made. On the other hand, under model mis-specification, the normal equations defining $\boldsymbol{\theta}_0$ inherently depend on $\mathbb{P}_{\mathbf{X}}$, thereby implying the potential utility of SSL in improving the estimation of $\boldsymbol{\theta}_0$ in this case.

The usual supervised estimator of $\boldsymbol{\theta}_0$ is the OLS estimator $\widehat{\boldsymbol{\theta}}$, the solution in $\boldsymbol{\theta}$ to the equation: $n^{-1}\sum_{i=1}^{n} \overrightarrow{\mathbf{X}}_i(Y_i - \overrightarrow{\mathbf{X}}_i'\boldsymbol{\theta}) = \mathbf{0}$, the normal equations based on $\mathcal{L}$. Under assumptions (c)-(d), it is well known that as $n \to \infty$,

$$n^{\frac{1}{2}}(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) = n^{-\frac{1}{2}} \sum_{i=1}^{n} \boldsymbol{\psi}_0(\mathbf{Z}_i) + O_p\left(n^{-\frac{1}{2}}\right) \xrightarrow{d} \mathcal{N}_{(p+1)}[\mathbf{0}, \boldsymbol{\Sigma}(g_{\boldsymbol{\theta}_0})], \tag{1.2}$$

where $\boldsymbol{\psi}_0(\mathbf{Z}) = \overrightarrow{\mathbf{X}}(Y - \overrightarrow{\mathbf{X}}'\boldsymbol{\theta}_0)$, $\boldsymbol{\Sigma}(g) = \mathbb{E}[\overrightarrow{\mathbf{X}}\overrightarrow{\mathbf{X}}'\{Y - g(\mathbf{X})\}^2]$ for any $g(.) \in \mathcal{L}_2(\mathbb{P}_{\mathbf{X}})$, $g_{\boldsymbol{\theta}}(\mathbf{X}) = \overrightarrow{\mathbf{X}}'\boldsymbol{\theta}\ \forall\ \boldsymbol{\theta} \in \mathbb{R}^{(p+1)}$, and for any a, $\mathcal{N}_a[\mathbf{0}, \boldsymbol{\Sigma}]$ denotes the a-variate Gaussian distribution with

mean $\mathbf{0}$ and covariance matrix $\mathbf{\Sigma}$.

Our primary goal is to obtain an efficient SS estimator of $\boldsymbol{\theta}_0$ using the *entire* training data $\mathcal{S}$ and compare its efficiency to that of $\widehat{\boldsymbol{\theta}}$. It is worth noting that the estimation efficiency of $\boldsymbol{\theta}_0$ also relates to the predictive performance of the fitted linear model since its out-of-sample prediction error is directly related to the mean squared error of the parameter estimate.

1.4 A Family of Imputation Based SS Estimators

If Y in $\mathcal{U}$ were actually observed, then one would simply fit the working model to the entire data in $\mathcal{S}$ for estimating $\boldsymbol{\theta}_0$. Our general approach is precisely motivated by this intuition. We first attempt to impute the missing Y in $\mathcal{U}$ based on suitable training of $\mathcal{L}$ in step (I). Then in step (II), we fit the linear model (1.1) to $\mathcal{U}$ with the imputed outcomes. Clearly, the imputation is critical. Inaccurate imputation would lead to biased estimate of $\boldsymbol{\theta}_0$, while inadequate imputation would result in loss of efficiency. We next consider SS estimators constructed under two imputation strategies for step (I) including a fully non-parametric imputation based on kernel smoothing (KS), and a semi-non-parametric (SNP) imputation that involves a smoothing step and a follow up 'refitting' step. Although the construction of the final EASE estimator is based on the SNP imputation strategy, it is helpful to begin with a discussion of the first strategy in order to appropriately motivate and elucidate the discussion on EASE and the SNP imputation strategy.

1.4.1 A Simple SS Estimator via Fully Non-Parametric Imputation

We present here an estimator based on a fully non-parametric imputation involving KS when p is small. For simplicity, we shall assume here that $\mathbf{X}$ is continuous with a density $f(.)$. Let $m(\mathbf{x}) = \mathbb{E}(Y|\mathbf{X} = \mathbf{x})$ and $l(\mathbf{x}) = m(\mathbf{x})f(\mathbf{x})$. Consider the local constant KS estimator of $m(\mathbf{x})$,

$$\widehat{m}(\mathbf{x}) \;=\; \frac{\frac{1}{nh^p}\sum_{i=1}^{n}\{K_h(\mathbf{X}_i,\mathbf{x})\}Y_i}{\frac{1}{nh^p}\sum_{i=1}^{n}K_h(\mathbf{X}_i,\mathbf{x})} \;=\; \frac{\widehat{l}(\mathbf{x})}{\widehat{f}(\mathbf{x})}, \tag{1.3}$$

where $K_h(\mathbf{u},\mathbf{v}) = K\{(\mathbf{u}-\mathbf{v})/h\}$ with $K : \mathbb{R}^p \to \mathbb{R}$ being some suitable kernel function and $h = h(n) > 0$ being the bandwidth. With $\widehat{m}(.)$ as defined in (1.3), we now fit (1.1) to the imputed unlabeled data: $[\{\widehat{m}(\mathbf{X}_j), \mathbf{X}_j'\}' : j = 1, ..., N]$ and obtain a SS estimator $\widehat{\boldsymbol{\theta}}_{np}$ of $\boldsymbol{\theta}_0$ as the solution in $\boldsymbol{\theta}$ to:

$$\frac{1}{N}\sum_{j=1}^{N} \overrightarrow{\mathbf{X}}_j\{\widehat{m}(\mathbf{X}_j) - \overrightarrow{\mathbf{X}}_j'\boldsymbol{\theta}\} = \mathbf{0}. \tag{1.4}$$

In order to study the properties of $\widehat{\boldsymbol{\theta}}_{np}$, we would require uniform (L_∞) convergence of $\widehat{m}(.)$ to $m(.)$, a problem that has been extensively studied in the non-parametric statistics literature (Newey, 1994; Andrews, 1995; Masry, 1996; Hansen, 2008) under fairly general settings and assumptions. In particular, we would assume the following regularity conditions to hold:

Assumption 1.1. (i) $K(.)$ is a symmetric q^{th} order kernel for some integer $q \geq 2$. (ii) $K(.)$ is bounded, Lipschitz continuous and has a bounded support $\mathcal{K} \subseteq \mathbb{R}^p$. (iii) $\mathbb{E}(|Y|^s) < \infty$ for some $s > 2$. $\mathbb{E}(|Y|^s|\mathbf{X} = \mathbf{x})f(\mathbf{x})$ and $f(\mathbf{x})$ are bounded on $\mathcal{X}$. (iv) $f(\mathbf{x})$ is bounded away from 0 on $\mathcal{X}$. (v) $m(.)$ and $f(.)$ are q times continuously differentiable with bounded q^{th} derivatives on some open set $\mathcal{X}_0 \supseteq \mathcal{X}$. (vi) For any $\delta > 0$, let $A_\delta \subseteq \mathbb{R}^p$ denote the set $\{(\mathbf{x}-\mathbf{X})/\delta : \mathbf{x} \in \mathcal{X}\}$. Then, for small enough δ, $A_\delta \supseteq \mathcal{K}$ almost surely (a.s.).

Conditions (i)-(v) are fairly standard in the literature. In (v), the set $\mathcal{X}_0$ is needed mostly to make the notion of differentiability well-defined, with both $m(.)$ and $f(.)$ understood to have been analytically extended over $(\mathcal{X}_0 \backslash \mathcal{X})$. Condition (vi) implicitly controls the tail behaviour of $\mathbf{X}$, requiring that perturbations of $\mathbf{X}$ in the form of $(\mathbf{X} + \delta\boldsymbol{\phi})$ with $\boldsymbol{\phi} \in \mathcal{K}$ (bounded) and δ small enough, belong to $\mathcal{X}$ a.s. $[\mathbb{P}_\mathbf{X}]$. We now present our result on $\widehat{\boldsymbol{\theta}}_{np}$.

Theorem 1.1. *Suppose* $n^{\frac{1}{2}}h^q \to 0$ *and* $(log\ n)/(n^{\frac{1}{2}}h^p) \to 0$ *as* $n \to \infty$. *Then, under Assumption 1.1,*

$$n^{\frac{1}{2}}\left(\widehat{\boldsymbol{\theta}}_{np} - \boldsymbol{\theta}_0\right) = n^{-\frac{1}{2}}\sum_{i=1}^{n} \boldsymbol{\psi}_{\mathit{eff}}(\mathbf{Z}_i) + O_p(r_n) \quad \xrightarrow{d} \mathcal{N}_{(p+1)}[\mathbf{0}, \boldsymbol{\Sigma}(m)], \tag{1.5}$$

where $\boldsymbol{\psi}_{\mathit{eff}}(\mathbf{Z}) = \overrightarrow{\mathbf{X}}\{Y - m(\mathbf{X})\}$ *and* $r_n = n^{\frac{1}{2}}h^q + (log\ n)/(n^{\frac{1}{2}}h^p) + (n/N)^{\frac{1}{2}}$.

Remark 1.1. Theorem 1.1 establishes the efficient and adaptive nature of $\widehat{\boldsymbol{\theta}}_{np}$. The asymptotic variance $\boldsymbol{\Sigma}(m)$ of $\widehat{\boldsymbol{\theta}}_{np}$ satisfies $\boldsymbol{\Sigma}(g) - \boldsymbol{\Sigma}(m) \geqq 0 \ \forall \ g(.) \in \mathcal{L}^2(\mathbf{X})$ and the inequality is strict unless $g(.) = m(.)$ a.s. $[\mathbb{P}_{\mathbf{X}}]$. Hence, $\widehat{\boldsymbol{\theta}}_{np}$ is asymptotically *optimal* among the class of all regular and asymptotically linear (RAL) estimators of $\boldsymbol{\theta}_0$ with influence function (IF) of the form: $\overrightarrow{\mathbf{X}}\{Y - g(\mathbf{X})\}$ with $g(.) \in \mathcal{L}_2(\mathbb{P}_{\mathbf{X}})$. In particular, $\widehat{\boldsymbol{\theta}}_{np}$ is more efficient than $\widehat{\boldsymbol{\theta}}$ whenever (1.1) is mis-specified, and equally efficient when (1.1) is correct i.e. $m(.) = g_{\boldsymbol{\theta}_0}(.)$. Further, for the semi-parametric model $\mathcal{M}_{\mathbf{X}} = \{(\mathbb{P}_{Y|\mathbf{X}}, \mathbb{P}_{\mathbf{X}}) : \mathbb{P}_{\mathbf{X}}$ is known, $\mathbb{P}_{Y|\mathbf{X}}$ is unrestricted upto assumptions (a)-(d)$\}$, it can be shown that the 'efficient' IF for estimating $\boldsymbol{\theta}_0$ is given by $\boldsymbol{\psi}_{\text{eff}}(\mathbf{Z})$. Thus, $\widehat{\boldsymbol{\theta}}_{np}$ also *achieves* the semi-parametric efficiency bound under $\mathcal{M}_{\mathbf{X}}$. Lastly, note that at any parametric sub-model in $\mathcal{M}_{\mathbf{X}}$ that corresponds to (1.1) being correct, $\widehat{\boldsymbol{\theta}}$ also achieves optimality, thus showing that under $\mathcal{M}_{\mathbf{X}}$, it is not possible to improve upon $\widehat{\boldsymbol{\theta}}$ if the linear model is correct.

Remark 1.2. The asymptotic results in Theorem 1.1 require a kernel of order $q > p$ and h smaller in order than the 'optimal' bandwidth order $h_{opt} = O(n^{-1/(2q+p)})$. This *under-smoothing* requirement, often encountered in two-step estimators involving a first-step smoothing (Newey et al., 1998), generally results in sub-optimal performance of $\widehat{m}(.)$. The optimal under-smoothed bandwidth order for Theorem 1.1 is given by: $O(n^{-1/(q+p)})$.

1.4.2 SS Estimators Based on Semi-Non-Parametric (SNP) Imputation

The simple and intuitive imputation strategy in section 1.4.1 based on a fully non-parametric p-dimensional KS is however often undesirable in practice owing to the curse of dimensionality. In order to accommodate larger p, we now propose a more flexible SNP imputation method involving a dimension reduction, if needed, followed by a non-parametric calibration. An additional 'refitting' step is proposed to reduce the impact of bias from non-parametric estimation and possibly inadequate imputation due to dimension reduction. We also introduce some flexibility in terms of the smoothing methods, apart from KS, that can be used

for the non-parametric calibration.

Let $r \leq p$ be a fixed positive integer and let $\mathbf{P}_r = [\mathbf{p}_1, .., \mathbf{p}_r]_{p \times r}$ be any rank r transformation matrix. Let $\mathbf{X}_{\mathbf{P}_r} = \mathbf{P}_r' \mathbf{X}$. Given $(r, \mathbf{P}_r)$, we may now consider approximating the regression function $\mathbb{E}(Y|\mathbf{X})$ by smoothing Y over the r dimensional $\mathbf{X}_{\mathbf{P}_r}$ instead of the original $\mathbf{X} \in \mathbb{R}^p$. In general, $\mathbf{P}_r$ can be user-defined and data dependent. A few reasonable choices of $\mathbf{P}_r$ are discussed in section 1.6. If $\mathbf{P}_r$ depends only on the distribution of $\mathbf{X}$, it may be assumed to be known given the SS setting considered. If $\mathbf{P}_r$ also depends on the distribution of Y, then it needs to be estimated from $\mathcal{L}$ and the smoothing needs to be performed using the estimated $\mathbf{P}_r$.

For approximating $\mathbb{E}(Y|\mathbf{X})$, we may consider *any* reasonable smoothing technique $\mathcal{T}$. Some examples of $\mathcal{T}$ include KS, kernel machine regression and smoothing splines. Let $m(\mathbf{x}; \mathbf{P}_r)$ denote the 'target function' for smoothing Y over $\mathbf{X}_{\mathbf{P}_r}$ using $\mathcal{T}$. For notational simplicity, the dependence of $m(\mathbf{x}; \mathbf{P}_r)$ and other quantities on $\mathcal{T}$ is suppressed throughout. For $\mathcal{T} :=$ KS, the appropriate target function is given by: $m(\mathbf{x}; \mathbf{P}_r) = m_{\mathbf{P}_r}(\mathbf{P}_r' \mathbf{x})$, where $m_{\mathbf{P}_r}(\mathbf{z}) \equiv \mathbb{E}(Y|\mathbf{X}_{\mathbf{P}_r} = \mathbf{z})$. For basis function expansion based methods, $m(\mathbf{x}; \mathbf{P}_r)$ will typically correspond to the L_2 projection of $m(\mathbf{x}) \equiv \mathbb{E}(Y|\mathbf{X} = \mathbf{x}) \in \mathcal{L}_2(\mathbb{P}_{\mathbf{X}})$ onto the functional space spanned by the basis functions associated with $\mathcal{T}$. The results in this section apply to any choice of $\mathcal{T}$ that satisfies the required conditions. In section 1.5, we provide more specific results for the implementation of our methods using $\mathcal{T} :=$ KS. Note that we do *not* assume $m(\mathbf{x}; \mathbf{P}_r) = m(\mathbf{x})$ anywhere and hence the name 'semi-non-parametric' imputation. Obviously, the case with $\mathbf{P}_r = I_p$ and $\mathcal{T} :=$ KS reduces to a fully non-parametric approach.

We next describe the two sub-steps involved in step (I) of the SNP imputation: (Ia) smoothing, and (Ib) refitting.

(Ia) Smoothing Step: With $\mathbf{P}_r$ and $m(\mathbf{x}; \mathbf{P}_r)$ as defined above, let $\widehat{\mathbf{P}}_r$ and $\widehat{m}(\mathbf{x}; \widehat{\mathbf{P}}_r)$ respectively denote their estimators based on $\mathcal{L}$. In order to address potential overfitting issues in the subsequent steps, we further consider generalized versions of these estimators

based on $\mathbb{K}$-fold CV for a given fixed integer $\mathbb{K} \geq 1$. For any $\mathbb{K} \geq 2$, let $\{\mathcal{L}_k\}_{k=1}^{\mathbb{K}}$ denote a random partition of $\mathcal{L}$ into $\mathbb{K}$ disjoint subsets of equal sizes, $n_{\mathbb{K}} = n/\mathbb{K}$, with index sets $\{\mathcal{I}_k\}_{k=1}^{\mathbb{K}}$. Let $\mathcal{L}_k^-$ denote the set excluding $\mathcal{L}_k$ with size $n_{\mathbb{K}}^- = n - n_{\mathbb{K}}$ and respective index set $\mathcal{I}_k^-$. Let $\widehat{\mathbf{P}}_{r,k}$ and $\widehat{m}_k(\mathbf{x}; \widehat{\mathbf{P}}_{r,k})$ denote the corresponding estimators based on $\mathcal{L}_k^-$. Further, for notational consistency, we define for $\mathbb{K} = 1$, $\mathcal{L}_k = \mathcal{L}_k^- = \mathcal{L}$; $\mathcal{I}_k = \mathcal{I}_k^- = \{1, ..., n\}$; $n_{\mathbb{K}} = n_{\mathbb{K}}^- = n$; $\widehat{\mathbf{P}}_{r,k} = \widehat{\mathbf{P}}_r$ and $\widehat{m}_k(\mathbf{x}; \widehat{\mathbf{P}}_{r,k}) = \widehat{m}(\mathbf{x}; \widehat{\mathbf{P}}_r)$.

(Ib) Refitting Step: In this step, we fit the linear model to $\mathcal{L}$ using $\mathbf{X}$ as predictors and the estimated $m(\mathbf{X}; \mathbf{P}_r)$ as an *offset*. To motivate this, we recall that the fully non-parametric imputation given in section 1.4.1 consistently estimates $\mathbb{E}(Y|\mathbf{X})$, the L_2 projection onto a space that always contains the working model space, i.e. the linear span of $\overrightarrow{\mathbf{X}}$. This need not be true for the SNP imputation, as is shown below in Figure 1.2, since we do not assume $m(\mathbf{X}, \mathbf{P}_r) = m(\mathbf{X})$ necessarily.

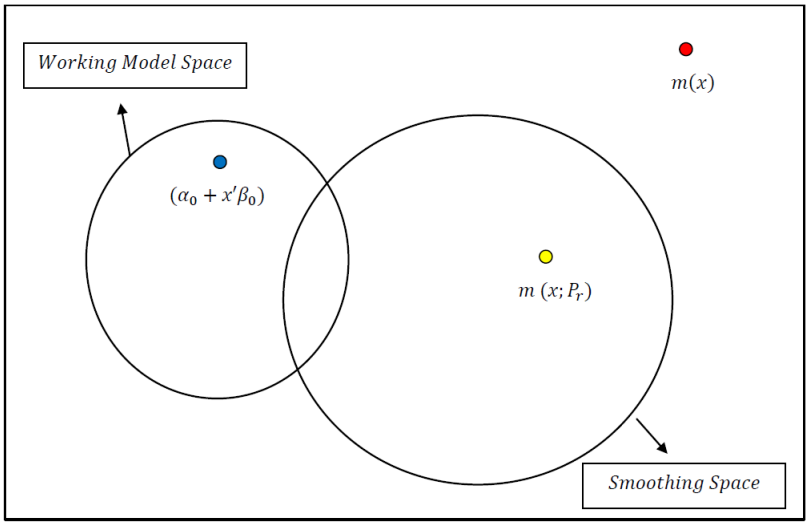


Figure 1.2: Geometric motivation behind the refitting step for an insufficient smoothing.

The refitting step essentially 'adjusts' for this so that the final imputation, combining the predictions from these two steps, targets a space that contains the working model space. In particular, for $\mathcal{T} :=$ KS with $r < p$, this step is critical to remove potential bias due to inadequate imputation. Interestingly, it turns out that the refitting step should *always* be

performed, even when $m(\mathbf{X};\mathbf{P}_r) = m(\mathbf{X})$. It plays a crucial role in reducing the bias of the resulting SS estimator due to the inherent bias from non-parametric curve estimation. In particular, for $\mathcal{T} := \text{KS}$ with *any* $r \leq p$, it ensures that a bandwidth of the optimal order can be used, thereby *eliminating the under-smoothing issue* as encountered in section 1.4.1. The target parameter for the refitting step is simply the regression coefficient obtained from regressing the residual $Y - m(\mathbf{X};\mathbf{P}_r)$ on $\mathbf{X}$ and may be defined as: $\boldsymbol{\eta}_{\mathbf{P}_r}$, the solution in $\boldsymbol{\eta} \in \mathbb{R}^{(p+1)}$ to the equation: $\mathbb{E}[\overrightarrow{\mathbf{X}}\{Y - m(\mathbf{X};\mathbf{P}_r) - \overrightarrow{\mathbf{X}}'\boldsymbol{\eta}\}] = \mathbf{0}$. For any $\mathbb{K} \geq 1$, we estimate $\boldsymbol{\eta}_{\mathbf{P}_r}$ as $\widehat{\boldsymbol{\eta}}_{(\mathbf{P}_r,\mathbb{K})}$, the solution in $\boldsymbol{\eta}$ to the equation:

$$n^{-1}\sum_{k=1}^{\mathbb{K}}\sum_{i\in\mathcal{I}_k}\overrightarrow{\mathbf{X}}_i\{Y_i - \widehat{m}_k(\mathbf{X}_i;\widehat{\mathbf{P}}_{r,k}) - \overrightarrow{\mathbf{X}}_i'\boldsymbol{\eta}\} = \mathbf{0}. \tag{1.6}$$

For $\mathbf{X}_i \in \mathcal{L}_k$, the estimate of $m(\mathbf{X}_i,\mathbf{P}_r)$ to be used as an offset is obtained from $\widehat{m}_k(.,\widehat{\mathbf{P}}_{r,k})$ that is based on data in $\mathcal{L}_k^-$. For $\mathbb{K} \geq 2$, with $\mathcal{L}_k^- \perp\!\!\!\perp \mathcal{L}_k$, the residuals are thus estimated in a cross-validated manner. For $\mathbb{K} = 1$ however, $\widehat{m}_(.,\widehat{\mathbf{P}}_r)$ is estimated using the entire $\mathcal{L}$ which can lead to considerable underestimation of the true residuals owing to over-fitting and consequently, substantial finite sample bias in the resulting SS estimator of $\boldsymbol{\theta}_0$. This bias can be effectively reduced by using the CV approach with $\mathbb{K} \geq 2$. We next estimate the *target function* for the SNP imputation given by:

$$\mu(\mathbf{x};\mathbf{P}_r) = m(\mathbf{x};\mathbf{P}_r) + \overrightarrow{\mathbf{x}}'\boldsymbol{\eta}_{\mathbf{P}_r} \text{ as:} \tag{1.7}$$

$$\widehat{\mu}(\mathbf{x};\widehat{\mathcal{P}}_{r,\mathbb{K}}) = \mathbb{K}^{-1}\sum_{k=1}^{\mathbb{K}}\widehat{m}_k(\mathbf{x},\widehat{\mathbf{P}}_{r,k}) + \overrightarrow{\mathbf{x}}'\widehat{\boldsymbol{\eta}}_{(\mathbf{P}_r,\mathbb{K})}, \tag{1.8}$$

where $\widehat{\mathcal{P}}_{r,\mathbb{K}} = \{\widehat{\mathbf{P}}_{r,k}\}_{k=1}^{\mathbb{K}}$. Using $\widehat{\mu}(.;\widehat{\mathcal{P}}_{r,\mathbb{K}})$, we now construct our final SS estimator as follows.

SS Estimator from SNP Imputation: In step (II), we fit the linear model to the SNP imputed unlabeled data: $[\{\widehat{\mu}(\mathbf{X}_j;\widehat{\mathcal{P}}_{r,\mathbb{K}}),\mathbf{X}_j'\}', j = 1,...,N]$ and obtain a SS estimator $\widehat{\boldsymbol{\theta}}_{(\mathbf{P}_r,\mathbb{K})}$

of $\boldsymbol{\theta}_0$ given by:

$$\widehat{\boldsymbol{\theta}}_{(\mathbf{P}_r,\mathbb{K})} \text{ is the solution in } \boldsymbol{\theta} \text{ to } \frac{1}{N}\sum_{j=1}^{N} \overrightarrow{\mathbf{X}}_j\{\widehat{\mu}(\mathbf{X}_j;\widehat{\mathcal{P}}_{r,\mathbb{K}}) - \overrightarrow{\mathbf{X}}_j'\boldsymbol{\theta}\} = \mathbf{0}. \tag{1.9}$$

For convenience of further discussion, let us define: $\forall\, k \in \{1,\ldots,\mathbb{K}\}$,

$$\widehat{\Delta}_k(\mathbf{x};\mathbf{P}_r,\widehat{\mathbf{P}}_{r,k}) = \widehat{m}_k(\mathbf{x};\widehat{\mathbf{P}}_{r,k}) - m(\mathbf{x};\mathbf{P}_r) \;\; \forall\, \mathbf{x} \in \mathcal{X}, \quad \text{and} \tag{1.10}$$

$$\widehat{\mathbf{G}}_k(\mathbf{x}) = \overrightarrow{\mathbf{x}}\widehat{\Delta}_k(\mathbf{x};\mathbf{P}_r,\widehat{\mathbf{P}}_{r,k}) - \mathbb{E}_{\mathbf{X}}\{\overrightarrow{\mathbf{X}}\widehat{\Delta}_k(\mathbf{X};\mathbf{P}_r,\widehat{\mathbf{P}}_{r,k})\} \;\; \forall\, \mathbf{x} \in \mathcal{X}, \tag{1.11}$$

where $\mathbb{E}_{\mathbf{X}}(.)$ denotes expectation w.r.t. $\mathbf{X} \in \mathcal{U}$. The dependence of $\widehat{\mathbf{G}}_k(.)$ on $(\mathbf{P}_r,\widehat{\mathbf{P}}_{r,k})$ and $\mathbb{P}_{\mathbf{X}}$ is suppressed here for notational simplicity. We now present our main result summarizing the properties of $\widehat{\boldsymbol{\theta}}_{(\mathbf{P}_r,\mathbb{K})}$.

Theorem 1.2. *Suppose that $\mathcal{T}$ satisfies: (i) $sup_{\mathbf{x}\in\mathcal{X}}|m(\mathbf{x};\mathbf{P}_r)| < \infty$ and (ii) $sup_{\mathbf{x}\in\mathcal{X}}|\widehat{m}(\mathbf{x};\widehat{\mathbf{P}}_r) - m(\mathbf{x};\mathbf{P}_r)| = O_p(c_n)$ for some $c_n = o(1)$. With $\widehat{\mathbf{G}}_k(.)$ as in (1.11), define $\mathbb{G}_{n,\mathbb{K}} = n^{-\frac{1}{2}}\sum_{k=1}^{\mathbb{K}}\sum_{i\in\mathcal{I}_k}\widehat{\mathbf{G}}_k(\mathbf{X}_i)$. Then, for any $\mathbb{K} \geq 1$,*

$$n^{\frac{1}{2}}\left(\widehat{\boldsymbol{\theta}}_{(\mathbf{P}_r,\mathbb{K})} - \boldsymbol{\theta}_0\right) = n^{-\frac{1}{2}}\sum_{i=1}^{n}\boldsymbol{\psi}(\mathbf{Z}_i;\mathbf{P}_r) - \mathbb{G}_{n,\mathbb{K}} + O_p(c^*_{n,\mathbb{K}}), \tag{1.12}$$

where $\boldsymbol{\psi}(\mathbf{Z};\mathbf{P}_r) = \overrightarrow{\mathbf{X}}\{Y - \mu(\mathbf{X};\mathbf{P}_r)\}$, $c^*_{n,\mathbb{K}} = c_{n^-_{\mathbb{K}}} + n^{-\frac{1}{2}} + (n/N)^{\frac{1}{2}} = o(1)$. Further, for any fixed $\mathbb{K} \geq 2$, $\mathbb{G}_{n,\mathbb{K}} = O_p(c_{n^-_{\mathbb{K}}})$, so that

$$n^{\frac{1}{2}}\left(\widehat{\boldsymbol{\theta}}_{(\mathbf{P}_r,\mathbb{K})} - \boldsymbol{\theta}_0\right) = n^{-\frac{1}{2}}\sum_{i=1}^{n}\boldsymbol{\psi}(\mathbf{Z}_i;\mathbf{P}_r) + O_p(c_{n^-_{\mathbb{K}}} + c^*_{n,\mathbb{K}}), \tag{1.13}$$

which converges in distribution to $\mathcal{N}_{(p+1)}[\mathbf{0},\boldsymbol{\Sigma}\{\mu(.\,;\mathbf{P}_r)\}]$.

Remark 1.3. If the imputation is 'sufficient' so that $\mu(\mathbf{x};\mathbf{P}_r) = m(\mathbf{x})$, then $\widehat{\boldsymbol{\theta}}_{(\mathbf{P}_r,\mathbb{K})}$, for any $\mathbb{K} \geq 2$, enjoys the same set of optimality properties as those noted in Remark 1.1 for $\widehat{\boldsymbol{\theta}}_{np}$ (while requiring less stringent assumptions about $K(.)$ and h, if KS is used). If

$\mu(\mathbf{x};\mathbf{P}_r) \neq m(\mathbf{x})$, then it is however unclear whether $\widehat{\boldsymbol{\theta}}_{(\mathbf{P}_r,\mathbb{K})}$ is always more efficient than $\widehat{\boldsymbol{\theta}}$. This will be addressed in section 1.4.3 where we develop the final EASE estimators.

Remark 1.4. Apart from the fairly mild condition (i), Theorem 1.2 *only* requires uniform consistency of $\widehat{m}(.\,;\widehat{\mathbf{P}}_r)$ w.r.t. $m(.\,;\mathbf{P}_r)$ for establishing the $n^{\frac{1}{2}}$-consistency and asymptotic normality (CAN) of $\widehat{\boldsymbol{\theta}}_{(\mathbf{P}_r,\mathbb{K})}$ for any $\mathbb{K} \geq 2$. The uniform consistency typically holds for a wide range of smoothing methods $\mathcal{T}$ under fairly general conditions. For $\mathcal{T} :=$ KS in particular, we provide explicit results in section 1.5 under mild regularity conditions that allow the use of any kernel order and the associated optimal bandwidth order. This is a notable relaxation from the stringent requirements for Theorem 1.1 that necessitate undersmoothing and the use of higher order kernels.

Remark 1.5. The CAN property of $\widehat{\boldsymbol{\theta}}_{(\mathbf{P}_r,1)}$ has *not* yet been established. The term $\mathbb{G}_{n,\mathbb{K}}$ in (1.12) behaves quite differently when $\mathbb{K} = 1$ compared to $\mathbb{K} \geq 2$. We derive the properties of $\widehat{\boldsymbol{\theta}}_{(\mathbf{P}_r,1)}$ in section 1.5 for $\mathcal{T} :=$ KS.

1.4.3 Efficient and Adaptive Semi-Supervised Estimators (EASE)

To ensure adaptivity even when $\mu(\mathbf{x};\mathbf{P}_r) \neq m(\mathbf{x})$, we now define the final EASE estimator as an optimal linear combination of $\widehat{\boldsymbol{\theta}}$ and $\widehat{\boldsymbol{\theta}}_{(\mathbf{P}_r,\mathbb{K})}$. Specifically, for any fixed $(p+1)\times(p+1)$ matrix $\boldsymbol{\Delta}$, $\widehat{\boldsymbol{\theta}}_{(\mathbf{P}_r,\mathbb{K})}(\boldsymbol{\Delta}) = \widehat{\boldsymbol{\theta}} + \boldsymbol{\Delta}(\widehat{\boldsymbol{\theta}}_{(\mathbf{P}_r,\mathbb{K})} - \widehat{\boldsymbol{\theta}})$ is a CAN estimator of $\boldsymbol{\theta}_0$ whenever $\widehat{\boldsymbol{\theta}}$ and $\widehat{\boldsymbol{\theta}}_{(\mathbf{P}_r,\mathbb{K})}$ are, and an optimal $\boldsymbol{\Delta}$ can be selected easily to minimize the asymptotic variance of the combined estimator. For simplicity, we focus here on $\boldsymbol{\Delta}$ being a diagonal matrix with $\boldsymbol{\Delta} = \text{diag}(\delta_1, ..., \delta_{p+1})$. Then the EASE is defined as $\widehat{\boldsymbol{\theta}}^{E}_{(\mathbf{P}_r,\mathbb{K})} \equiv \widehat{\boldsymbol{\theta}}_{(\mathbf{P}_r,\mathbb{K})}(\widehat{\boldsymbol{\Delta}})$ with $\widehat{\boldsymbol{\Delta}}$ being any consistent estimator (see section 1.4.4 for details) of the minimizer $\overline{\boldsymbol{\Delta}} = \text{diag}(\overline{\delta}_1, ..., \overline{\delta}_{p+1})$, where $\forall\ 1 \leq l \leq (p+1)$,

$$\overline{\delta}_l = -\lim_{\epsilon \downarrow 0} \frac{\text{Cov}\left\{\boldsymbol{\psi}_{0[l]}(\mathbf{Z}),\ \boldsymbol{\psi}_{[l]}(\mathbf{Z};\mathbf{P}_r) - \boldsymbol{\psi}_{0[l]}(\mathbf{Z})\right\}}{\text{Var}\left\{\boldsymbol{\psi}_{[l]}(\mathbf{Z};\mathbf{P}_r) - \boldsymbol{\psi}_{0[l]}(\mathbf{Z})\right\}\ +\ \epsilon}, \tag{1.14}$$

and for any vector $\mathbf{a}$, $\mathbf{a}_{[l]}$ denotes its l^{th} component. Note that in (1.14), the ϵ and the limit outside are included to formally account for the case: $\boldsymbol{\psi}_{0[l]}(\mathbf{Z}) = \boldsymbol{\psi}_{[l]}(\mathbf{Z}, \mathbf{P}_r)$ a.s. $[\mathbb{P}_{\mathbf{Z}}]$, when we define $\overline{\delta}_l = 0$ for identifiability.

It is straightforward to show that $\widehat{\boldsymbol{\theta}}^E_{(\mathbf{P}_r,\mathbb{K})}$ and $\widehat{\boldsymbol{\theta}}_{(\mathbf{P}_r,\mathbb{K})}(\overline{\boldsymbol{\Delta}})$ are asymptotically equivalent, so that $\widehat{\boldsymbol{\theta}}^E_{(\mathbf{P}_r,\mathbb{K})}$ is a RAL estimator of $\boldsymbol{\theta}_0$ satisfying:

$$n^{\frac{1}{2}}\left(\widehat{\boldsymbol{\theta}}^E_{(\mathbf{P}_r,\mathbb{K})} - \boldsymbol{\theta}_0\right) = n^{-\frac{1}{2}}\sum_{i=1}^{n}\boldsymbol{\psi}(\mathbf{Z}_i; \mathbf{P}_r, \overline{\boldsymbol{\Delta}}) + o_p(1) \quad \xrightarrow{d} \mathcal{N}_{(p+1)}[\mathbf{0}, \boldsymbol{\Sigma}_{\mathbf{P}_r}(\overline{\boldsymbol{\Delta}})],$$

as $n \to \infty$, where $\boldsymbol{\psi}(\mathbf{Z}; \mathbf{P}_r, \overline{\boldsymbol{\Delta}}) = \boldsymbol{\psi}_0(\mathbf{Z}) + \overline{\boldsymbol{\Delta}}\{\boldsymbol{\psi}(\mathbf{Z}; \mathbf{P}_r) - \boldsymbol{\psi}_0(\mathbf{Z})\}$ and $\boldsymbol{\Sigma}_{\mathbf{P}_r}(\overline{\boldsymbol{\Delta}}) = \text{Var}\{\boldsymbol{\psi}(\mathbf{Z}; \mathbf{P}_r, \overline{\boldsymbol{\Delta}})\}$. Note that when either the linear model holds or the SNP imputation is sufficient, then $\boldsymbol{\psi}(\mathbf{Z}; \mathbf{P}_r, \overline{\boldsymbol{\Delta}}) = \boldsymbol{\psi}_{\text{eff}}(\mathbf{Z})$, so that $\widehat{\boldsymbol{\theta}}^E_{(\mathbf{P}_r,\mathbb{K})}$ is asymptotically optimal (in the sense of Remark 1.1). Further, when neither cases hold, $\widehat{\boldsymbol{\theta}}^E_{(\mathbf{P}_r,\mathbb{K})}$ is no longer optimal, but is *still* efficient and adaptive compared to $\widehat{\boldsymbol{\theta}}$. Lastly, if the imputation is certain to be sufficient (e.g. if $r = p$ and $\mathcal{T} := \text{KS}$), we may simply define $\widehat{\boldsymbol{\theta}}^E_{(\mathbf{P}_r,\mathbb{K})} = \widehat{\boldsymbol{\theta}}_{(\mathbf{P}_r,\mathbb{K})}$.

1.4.4 Inference for the EASE and the SNP Imputation Based SS Estimators

We now provide procedures for making inference about $\boldsymbol{\theta}_0$ based on $\widehat{\boldsymbol{\theta}}_{(\mathbf{P}_r,\mathbb{K})}$ and $\widehat{\boldsymbol{\theta}}^E_{(\mathbf{P}_r,\mathbb{K})}$ obtained using $\mathbb{K} \geq 2$. We also employ a 'double' CV to overcome bias in variance estimation due to over-fitting. A key step involved in the variance estimation is to obtain reasonable estimates of $\{\mu(\mathbf{X}_i; \mathbf{P}_r)\}_{i=1}^n$. Although $\widehat{\boldsymbol{\eta}}_{(\mathbf{P}_r,\mathbb{K})}$ in (1.6) was constructed via CV, the corresponding estimate, $\widehat{\mu}(\mathbf{x}; \widehat{\mathcal{P}}_{r,\mathbb{K}})$ in (1.8), of $\mu(\mathbf{x}; \mathbf{P}_r)$ is likely to be over-fitted for $\mathbf{X}_i \in \mathcal{L}$. To construct bias corrected estimates of $\mu(\mathbf{X}_i; \mathbf{P}_r)$, we first obtain $\mathbb{K}$ separate *doubly cross-validated* estimates of $\boldsymbol{\eta}_{\mathbf{P}_r}$, $\{\widehat{\boldsymbol{\eta}}^k_{(\mathbf{P}_r,\mathbb{K})} : k = 1, ..., \mathbb{K}\}$, with $\widehat{\boldsymbol{\eta}}^k_{(\mathbf{P}_r,\mathbb{K})}$, for each k, being the solution in $\boldsymbol{\eta}$ to $\sum_{k' \neq k} \mathcal{S}_{k'}(\boldsymbol{\eta}) = \mathbf{0}$, where

$$\mathcal{S}_{k'}(\boldsymbol{\eta}) = \textstyle\sum_{i \in \mathcal{I}_{k'}} \overrightarrow{\mathbf{X}}_i\{Y_i - \widehat{m}_{k'}(\mathbf{X}_i; \widehat{\mathbf{P}}_{r,k'}) - \overrightarrow{\mathbf{X}}_i'\boldsymbol{\eta}\} \quad \forall\, k' \in \{1, \ldots, \mathbb{K}\}.$$

For each k and $k' \neq k$, $\mathcal{S}_{k'}(\boldsymbol{\eta})$ is constructed such that $\{\mathbf{Z}_i : i \in \mathcal{I}_{k'}\}$ used for obtaining $\widehat{\boldsymbol{\eta}}^k_{(\mathbf{P}_r,\mathbb{K})}$ is independent of $\widehat{m}_{k'}(.;\widehat{\mathbf{P}}_{r,k'})$ that is based on $\mathcal{L}^-_{k'} \perp\!\!\!\perp \mathcal{L}_{k'}$. Then, for each $\mathbf{X}_i \in \mathcal{L}_k$ and $k \in \{1,\ldots,\mathbb{K}\}$, we may estimate $\mu(\mathbf{X}_i;\mathbf{P}_r)$ as:

$$\widehat{\mu}_k(\mathbf{X}_i;\widehat{\mathcal{P}}_{r,\mathbb{K}}) = \widehat{m}_k(\mathbf{X}_i;\widehat{\mathbf{P}}_{r,k}) + \overrightarrow{\mathbf{X}}_i'\widehat{\boldsymbol{\eta}}^k_{(\mathbf{P}_r,\mathbb{K})}.$$

We exclude $\mathcal{S}_k(\boldsymbol{\eta})$ in the construction of $\widehat{\boldsymbol{\eta}}^k_{(\mathbf{P}_r,\mathbb{K})}$ to reduce over-fitting bias in the residuals $\{Y_i - \widehat{\mu}_k(\mathbf{X}_i;\widehat{\mathcal{P}}_{r,\mathbb{K}})\}$ which we now use for estimating the IFs.

For each $\mathbf{Z}_i \in \mathcal{L}_k$ and $k \in \{1,..,\mathbb{K}\}$, we estimate $\boldsymbol{\psi}_0(\mathbf{Z}_i)$ and $\boldsymbol{\psi}(\mathbf{Z}_i;\mathbf{P}_r)$, the corresponding IFs of $\widehat{\boldsymbol{\theta}}$ and $\widehat{\boldsymbol{\theta}}_{(\mathbf{P}_r,\mathbb{K})}$, respectively as:

$$\widehat{\boldsymbol{\psi}}_0(\mathbf{Z}_i) = \overrightarrow{\mathbf{X}}_i(Y_i - \overrightarrow{\mathbf{X}}_i'\widehat{\boldsymbol{\theta}}), \text{ and } \widehat{\boldsymbol{\psi}}_k(\mathbf{Z}_i;\mathbf{P}_r) = \overrightarrow{\mathbf{X}}_i\{Y_i - \widehat{\mu}_k(\mathbf{X}_i;\widehat{\mathcal{P}}_{r,\mathbb{K}})\}.$$

Then, $\boldsymbol{\Sigma}\{\mu(.;\mathbf{P}_r)\}$ in (1.13) may be consistently estimated as:

$$\widehat{\boldsymbol{\Sigma}}\{\mu(.;\mathbf{P}_r)\} = n^{-1}\sum_{k=1}^{\mathbb{K}}\sum_{i\in\mathcal{I}_k}\widehat{\boldsymbol{\psi}}_k(\mathbf{Z}_i;\mathbf{P}_r)\widehat{\boldsymbol{\psi}}_k'(\mathbf{Z}_i;\mathbf{P}_r).$$

To estimate the combination matrix $\overline{\boldsymbol{\Delta}}$ in (1.14) and the asymptotic variance, $\boldsymbol{\Sigma}_{\mathbf{P}_r}(\overline{\boldsymbol{\Delta}})$, of EASE consistently, let us define, $\forall\ 1 \leq l \leq (p+1)$,

$$\begin{aligned}
\widehat{\sigma}_{l,12} &= -\ n^{-1}\textstyle\sum_{k=1}^{\mathbb{K}}\sum_{i\in\mathcal{I}_k}\widehat{\boldsymbol{\psi}}_{0[l]}(\mathbf{Z}_i)\{\widehat{\boldsymbol{\psi}}_{k[l]}(\mathbf{Z}_i;\mathbf{P}_r) - \widehat{\boldsymbol{\psi}}_{0[l]}(\mathbf{Z}_i)\},\\
\widehat{\sigma}_{l,22} &= n^{-1}\textstyle\sum_{k=1}^{\mathbb{K}}\sum_{i\in\mathcal{I}_k}\{\widehat{\boldsymbol{\psi}}_{k[l]}(\mathbf{Z}_i;\mathbf{P}_r) - \widehat{\boldsymbol{\psi}}_{0[l]}(\mathbf{Z}_i)\}^2,
\end{aligned}$$

and $\widehat{\delta}_l = \widehat{\sigma}_{l,12}/(\widehat{\sigma}_{l,22} + \epsilon_n)$ for some sequence $\epsilon_n \to 0$ with $n^{\frac{1}{2}}\epsilon_n \to \infty$. Then, we estimate $\overline{\boldsymbol{\Delta}}$ and $\boldsymbol{\Sigma}_{\mathbf{P}_r}(\overline{\boldsymbol{\Delta}})$ respectively as: $\widehat{\boldsymbol{\Delta}} = \text{diag}(\widehat{\delta}_1,...,\widehat{\delta}_{p+1})$ and

$$\widehat{\boldsymbol{\Sigma}}_{\mathbf{P}_r}(\widehat{\boldsymbol{\Delta}}) = n^{-1}\sum_{k=1}^{\mathbb{K}}\sum_{i\in\mathcal{I}_k}\widehat{\boldsymbol{\psi}}_k(\mathbf{Z}_i;\mathbf{P}_r,\widehat{\boldsymbol{\Delta}})\widehat{\boldsymbol{\psi}}_k'(\mathbf{Z}_i;\mathbf{P}_r,\widehat{\boldsymbol{\Delta}}),$$

where $\widehat{\boldsymbol{\psi}}_k(\mathbf{Z};\mathbf{P}_r,\widehat{\boldsymbol{\Delta}}) = \widehat{\boldsymbol{\psi}}_0(\mathbf{Z}) + \widehat{\boldsymbol{\Delta}}\{\widehat{\boldsymbol{\psi}}_k(\mathbf{Z};\mathbf{P}_r) - \widehat{\boldsymbol{\psi}}_0(\mathbf{Z})\}\ \forall\ k \in \{1,\ldots,\mathbb{K}\}$. Normal confidence intervals (CIs) for the parameters of interest can also be constructed accordingly based on these variance estimates.

1.5 Implementation Based on Kernel Smoothing (KS)

We next detail the specific implementation of the SNP imputation based on KS estimators. With $\mathcal{T} :=$ KS, the target function for the smoothing is given by: $m(\mathbf{x};\mathbf{P}_r) = m_{\mathbf{P}_r}(\mathbf{P}_r'\mathbf{x}) \equiv \mathbb{E}(Y|\mathbf{X}_{\mathbf{P}_r} = \mathbf{P}_r'\mathbf{x})$. For simplicity, we assume that $\mathbf{X}_{\mathbf{P}_r}$ is continuous with a density $f_{\mathbf{P}_r}(.)$ and support $\mathcal{X}_{\mathbf{P}_r} \equiv \{\mathbf{P}_r'\mathbf{x} : \mathbf{x} \in \mathcal{X}\} \subseteq \mathbb{R}^r$. Let us now consider the following class of local constant KS estimators for $m(\mathbf{x};\mathbf{P}_r)$:

$$\widehat{m}_k(\mathbf{x};\widehat{\mathbf{P}}_{r,k}) \;=\; \frac{\frac{1}{n_{\mathbb{K}}^- h^r}\sum_{i\in\mathcal{I}_k^-}\{K_h(\widehat{\mathbf{P}}_{r,k}'\mathbf{X}_i, \widehat{\mathbf{P}}_{r,k}'\mathbf{x})\}Y_i}{\frac{1}{n_{\mathbb{K}}^- h^r}\sum_{i\in\mathcal{I}_k^-} K_h(\widehat{\mathbf{P}}_{r,k}'\mathbf{X}_i, \widehat{\mathbf{P}}_{r,k}'\mathbf{x})} \quad \forall\ 1 \le k \le \mathbb{K}, \tag{1.15}$$

where $K_h(.)$ and h are as in section 1.4.1 with $K(.)$ now being a suitable kernel on $\mathbb{R}^r$. In the light of Theorem 1.2, we focus primarily on establishing the uniform consistency of $\widehat{m}(\mathbf{x};\widehat{\mathbf{P}}_r) \equiv \widehat{m}_1(\mathbf{x};\widehat{\mathbf{P}}_{r,1})$ in (1.15) with $\mathbb{K} = 1$, *accounting* for the additional estimation error from $\widehat{\mathbf{P}}_r$. For establishing the desired result, we shall assume the following regularity conditions.

Assumption 1.2. (i) $K(.)$ is a symmetric kernel of order $q \ge 2$ with finite q^{th} moments. (ii) $K(.)$ is bounded, integrable and is either Lipschitz continuous with a compact support or, has a bounded derivative $\boldsymbol{\nabla}K(.)$ which satisfies: $\|\boldsymbol{\nabla}K(\mathbf{z})\| \le \Lambda\|\mathbf{z}\|^{-\rho}\ \forall\ \mathbf{z} \in \mathbb{R}^r$ with $\|\mathbf{z}\| > L$, where $\Lambda > 0$, $L > 0$ and $\rho > 1$ are some fixed constants, and $\|.\|$ denotes the standard L_2 vector norm. (iii) $\mathcal{X}_{\mathbf{P}_r} \subseteq \mathbb{R}^r$ is compact. $\mathbb{E}(|Y|^s) < \infty$ for some $s > 2$. $\mathbb{E}(|Y|^s|\mathbf{X}_{\mathbf{P}_r} = \mathbf{z})f_{\mathbf{P}_r}(\mathbf{z})$ and $f_{\mathbf{P}_r}(\mathbf{z})$ are bounded on $\mathcal{X}_{\mathbf{P}_r}$. (iv) $f_{\mathbf{P}_r}(\mathbf{z})$ is bounded away from 0 on $\mathcal{X}_{\mathbf{P}_r}$. (v) $m_{\mathbf{P}_r}(\mathbf{z})$ and $f_{\mathbf{P}_r}(\mathbf{z})$ are both q times continuously differentiable with bounded q^{th} derivatives on some open set $\mathcal{X}_{0,\mathbf{P}_r} \supseteq \mathcal{X}_{\mathbf{P}_r}$. *Additional Conditions (required only if* $\mathbf{P}_r$ *needs to be estimated):* (vi) $K(.)$ has a bounded and integrable derivative $\boldsymbol{\nabla}K(.)$. (vii) $\boldsymbol{\nabla}K(.)$ satisfies: $\|\boldsymbol{\nabla}K(\mathbf{z}_1) - \boldsymbol{\nabla}K(\mathbf{z}_2)\| \le \|\mathbf{z}_1 - \mathbf{z}_2\|\ \phi(\mathbf{z}_1)\ \forall\ \mathbf{z}_1, \mathbf{z}_2 \in \mathbb{R}^r$ such that $\|\mathbf{z}_1 - \mathbf{z}_2\| \le L^*$,

for some fixed constant $L^* > 0$, and some bounded and integrable function $\phi : \mathbb{R}^r \to \mathbb{R}^+$. (viii) $\boldsymbol{\nabla} K(.)$ is Lipschitz continuous on $\mathbb{R}^r$. (ix) $\mathbb{E}(\mathbf{X}|\mathbf{X}_{\mathbf{P}_r} = \mathbf{z})$ and $\mathbb{E}(\mathbf{X}Y|\mathbf{X}_{\mathbf{P}_r} = \mathbf{z})$ are both continuously differentiable with bounded first derivatives on $\mathcal{X}_{0,\mathbf{P}_r} \supseteq \mathcal{X}_{\mathbf{P}_r}$.

Assumption 1.2, mostly adopted from Hansen (2008), imposes some mild smoothness and moment conditions most of which are fairly standard, except perhaps the conditions on $K(.)$ in (vi)-(viii) all of which are however satisfied by the Gaussian kernel among others. We now propose the following result.

Theorem 1.3. *Suppose* $(\widehat{\mathbf{P}}_r - \mathbf{P}_r) = O_p(\alpha_n)$ *for some* $\alpha_n = o(1)$ *with* $\alpha_n = 0$ *identically if* $\mathbf{P}_r$ *is known. Let* q *be the order of the kernel* $K(.)$ *in (1.15) for some integer* $q \geq 2$*. Define:*

$$a_{n,1} = \alpha_n \left(\frac{log\, n}{nh^{r+2}}\right)^{\frac{1}{2}} + \alpha_n^2 h^{-(r+2)} + \alpha_n, \quad a_{n,2} = \left(\frac{log\, n}{nh^r}\right)^{\frac{1}{2}} + h^q$$

and assume that each of the terms involved in $a_{n,1} = o(1)$ *and* $a_{n,2} = o(1)$*. Then, under Assumption 1.2,* $\widehat{m}(\mathbf{x}; \widehat{\mathbf{P}}_r)$*, based on (1.15), satisfies:*

$$sup_{\mathbf{x}\in\mathcal{X}} \, |\widehat{m}(\mathbf{x}; \widehat{\mathbf{P}}_r) - m(\mathbf{x}; \mathbf{P}_r)| = O_p(a_{n,1} + a_{n,2}). \tag{1.16}$$

Remark 1.6. Theorem 1.3 establishes the L_∞ error rate of $\widehat{m}(\mathbf{x}; \widehat{\mathbf{P}}_r)$ under mild regularity conditions and restrictions on h. Among its various implications, the rate also ensures uniform consistency of $\widehat{m}(\mathbf{x}; \widehat{\mathbf{P}}_r)$ at the optimal bandwidth order: $h_{opt} = O(n^{-1/(2q+r)})$ for any kernel order $q \geq 2$ and any $r \leq p$, as long as $\alpha_n = o(n^{-(r+2)/(4q+2r)})$ which always includes: $\alpha_n = O(n^{-\frac{1}{2}})$ and $\alpha_n = 0$. These two cases are particularly relevant in practice as $\mathbf{P}_r$ being finite dimensional, $n^{\frac{1}{2}}$-consistent estimators of $\mathbf{P}_r$ should typically exist. For both cases, using h_{opt} results in $a_{n,1}$ to be of lower order (for $q > 2$) or the same order (for $q = 2$) compared to that of the main term $a_{n,2}$, so that the usual optimal rate prevails as the overall error rate.

Properties of $\widehat{\boldsymbol{\theta}}_{(\mathbf{P}_r,\mathbb{K})}$ for $\mathbb{K}=1$

We now address the CAN property of $\widehat{\boldsymbol{\theta}}_{(\mathbf{P}_r,\mathbb{K})}$ for $\mathbb{K}=1$ under the KS framework. Based on (1.12) and Remark 1.5, the only step required for this is to effectively control the term $\mathbb{G}_{n,\mathbb{K}}$ in (1.12). We propose the following result in this regard.

Theorem 1.4. *Let $\mathbb{K}=1$, $\mathcal{T}:=$ KS, $\mathbb{G}_{n,\mathbb{K}}$ be as in (1.12), and $\widehat{m}(\mathbf{x};\widehat{\mathbf{P}}_r)$ be the KS estimator based on (1.15). Let α_n, $a_{n,1}$ and $a_{n,2}$ be as in Theorem 1.3 with $(\widehat{\mathbf{P}}_r-\mathbf{P}_r)=O_p(\alpha_n)$. Assume that $a^*_{n,1}$ and $a^*_{n,2}$ are $o(1)$, where*

$$a^*_{n,1}=\alpha_n+\frac{\alpha_n}{n^{\frac{1}{2}}h^{(r+1)}}+n^{\frac{1}{2}}\alpha_n^2h^{-2}+n^{\frac{1}{2}}a^2_{n,1}+n^{\frac{1}{2}}a_{n,1}a_{n,2}\quad\text{and}\quad a^*_{n,2}=n^{\frac{1}{2}}a^2_{n,2}.$$

Then, under Assumption 1.2, $\mathbb{G}_{n,\mathbb{K}}=O_p(a^*_{n,1}+a^*_{n,2})=o_p(1)$. Further, let $c^*_{n,\mathbb{K}}$ be as in Theorem 1.2 with $c_n=(a_{n,1}+a_{n,2})$. Then, using (1.12),

$$n^{\frac{1}{2}}\left(\widehat{\boldsymbol{\theta}}_{(\mathbf{P}_r,\mathbb{K})}-\boldsymbol{\theta}_0\right)=n^{-\frac{1}{2}}\sum_{i=1}^{n}\boldsymbol{\psi}(\mathbf{Z}_i,\mathbf{P}_r)+O_p(c^*_{n,\mathbb{K}}+d_n),\tag{1.17}$$

where $d_n=a^*_{n,1}+a^*_{n,2}$. Hence, $n^{\frac{1}{2}}(\widehat{\boldsymbol{\theta}}_{(\mathbf{P}_r,\mathbb{K})}-\boldsymbol{\theta}_0)\xrightarrow{d}\mathcal{N}_{(p+1)}[\mathbf{0},\boldsymbol{\Sigma}\{\mu(.\,;\mathbf{P}_r)\}]$.

Remark 1.7. Note that the term $a^*_{n,2}$ *always* requires $q>r/2$ in order to converge to 0, thus showing the contrasting behavior of the case $\mathbb{K}=1$ compared to $\mathbb{K}\geq 2$ where no such higher order kernel restriction is required. Nevertheless, when $\alpha_n=O(n^{-\frac{1}{2}})$ or $\alpha_n=0$, the optimal bandwidth order: $h_{opt}=O(n^{-1/(2q+r)})$ can indeed be *still* used as long as $q>r/2$ is satisfied. Despite these facts and all the theoretical guarantee in Theorem 1.4, empirical evidence however seems to suggest that $\widehat{\boldsymbol{\theta}}_{(\mathbf{P}_r,1)}$ can be substantially *biased* in finite samples, in part due to over-fitting. This will be demonstrated via our simulation studies in section 1.7.1.

Remark 1.8. *Technical benefits of refitting and CV:* Suppose that $\mathbf{P}_r=I_p$, so that the SNP imputation with $\mathcal{T}:=$ KS is indeed sufficient. Further, assume that all of Theorems 1.1-1.4 hold, so that the estimators $\widehat{\boldsymbol{\theta}}_{np}$, $\widehat{\boldsymbol{\theta}}_{(\mathbf{P}_r,1)}$, and $\widehat{\boldsymbol{\theta}}_{(\mathbf{P}_r,\mathbb{K})}$ ($\mathbb{K}\geq 2$) are comparable and all

asymptotically optimal. However, their constructions are quite different which can significantly affect their finite sample performances. $\widehat{\boldsymbol{\theta}}_{np}$ is based on KS only, and requires stringent under-smoothing and a kernel of order $q > p$ (Remark 1.2); $\widehat{\boldsymbol{\theta}}_{(\mathbf{P}_r,1)}$ is based on KS and refitting (*although* the KS itself is certain to be sufficient), and requires no under-smoothing but needs a (weaker) kernel order condition ($q > p/2$) (Remark 1.7); while $\widehat{\boldsymbol{\theta}}_{(\mathbf{P}_r,\mathbb{K})}$ ($\mathbb{K} \geq 2$) additionally involves CV, and requires no under-smoothing or higher order kernel conditions (Remark 1.4). This highlights the critical role played by refitting and CV, apart from their primary roles in the SNP imputation, in removing any under-smoothing and/or higher order kernel restrictions when $\mathcal{T} :=$ KS, and this continues to hold for any other $(r, \mathbf{P}_r)$ as well. In particular, it shows, rather surprisingly, that refitting should be performed in order to avoid under-smoothing *even if* the smoothing is known to be sufficient.

1.6 Dimension Reduction Techniques

We next discuss choosing and estimating the matrix $\mathbf{P}_r$ ($r < p$) to be used for dimension reduction, if required, in the SNP imputation, and which can play an important role in the sufficiency of the imputation. Simple choices of $\mathbf{P}_r$ include r leading principal component directions of $\mathbf{X}$ or any r canonical directions of $\mathbf{X}$. Note that under the SS setting, $\mathbf{P}_r$ is effectively known if it only involves the distribution of $\mathbf{X}$. We now focus primarily on the case where $\mathbf{P}_r$ also depends on the distribution of Y and hence, is unknown. Such a choice of $\mathbf{P}_r$ is often desirable to ensure that the imputation is as 'sufficient' as possible for predicting Y. Several reasonable choices of such $\mathbf{P}_r$ and their estimation are possible based on non-parametric dimension reduction methods like Sliced Inverse Regression (SIR) (Li, 1991), Principal Hessian Directions (PHD) (Li, 1992; Cook, 1998), Sliced Average Variance Estimation (SAVE) (Cook and Weisberg, 1991; Cook and Lee, 1999) etc. In particular, we focus here on SIR where the choice of $\mathbf{P}_r$ is $\mathbf{P}_r^0$, the eigenvectors corresponding to the r largest eigenvalues of $\mathbb{M} = \mathrm{Var}\{\mathbb{E}(\mathbf{X}|Y)\}$, which leads to an optimal (in some sense) r-dimensional linear transform of $\mathbf{X}$ that can be predicted by Y (Li, 1991).

For estimating $\mathbf{P}_r^0$, we consider the SIR algorithm of Li (1991) and further propose a SS modification to it. With $\mathbb{K}$ and $\{\mathcal{L}_k^-, \mathcal{I}_k^-, \widehat{\mathbf{P}}_{r,k}\}_{k=1}^{\mathbb{K}}$ as before, the original SIR algorithm, under our setting, estimates $\mathbf{P}_r^0$ based on $\mathcal{L}_k^-$ as follows: (i) divide the range of $\{Y_i\}_{i\in\mathcal{I}_k^-}$ into H slices: $\{I_1, .., I_H\}$, where H may depend on $n_{\mathbb{K}}^-$. For $1 \leq h \leq H$, let $\widehat{p}_{h,k}$ denote the proportion of $\{Y_i\}_{i\in\mathcal{I}_k^-}$ in slice I_h; (ii) for each I_h, let $\widehat{\mathbb{M}}_{h,k}$ denote the sample average of the set: $\{\mathbf{X}_i \in \mathcal{L}_k^- : Y_i \in I_h\}$; (iii) estimate $\mathbb{M}$ as: $\widehat{\mathbb{M}}_k = \sum_{h=1}^H \widehat{p}_{h,k}\widehat{\mathbb{M}}_{h,k}\widehat{\mathbb{M}}'_{h,k}$ and $\mathbf{P}_r^0$ as: $\widehat{\mathbf{P}}_{r,k}^0 \rightarrow$ the eigenvectors corresponding to the r largest eigenvalues of $\widehat{\mathbb{M}}_k$. However, the SIR algorithm often tends to give unstable estimates of $\mathbf{P}_r^0$, especially for the directions corresponding to the smaller eigenvalues of $\mathbb{M}$. To improve the efficiency in estimating $\mathbf{P}_r^0$, we now propose a semi-supervised SIR (SS-SIR) algorithm as follows.

The SS-SIR Algorithm

Step (i) stays the same as in SIR, and in step (ii), for each k, and $j \in \mathcal{I}_{\mathcal{U}} = \{1, ..., N\}$, we impute Y_j as $Y_{j,k}^* = Y_{\widehat{i}_{j,k}}$, where $\widehat{i}_{j,k} = \text{argmin}_{i\in\mathcal{I}_k^-}\|\mathbf{X}_i - \mathbf{X}_j\|^2$. For each I_h, let $\widehat{\mathbb{M}}_{h,k}^*$ be the sample average of the set: $\{\mathbf{X}_i \in \mathcal{L}_k^- : Y_i \in I_h\} \bigcup \{\mathbf{X}_j \in \mathcal{U} : Y_{j,k}^* \in I_h\}$. Then in step (iii), we estimate $\mathbb{M}$ as: $\widehat{\mathbb{M}}_k^* = \sum_{h=1}^H \widehat{p}_{h,k}\widehat{\mathbb{M}}_{h,k}^*\widehat{\mathbb{M}}_{h,k}^{*'}$ and accordingly, $\mathbf{P}_r^0$ as $\widehat{\mathbf{P}}_{r,k}^{0*}$, the eigenvectors corresponding to the r largest eigenvalues of $\widehat{\mathbb{M}}_k^*$.

The SS-SIR algorithm aims to improve the estimation of $\mathbf{P}_r^0$ by making use of $\mathcal{U}$ in step (ii) through a nearest neighbour approximation for the unobserved Y in $\mathcal{U}$ using $\mathcal{L}_k^-$. With $n_{\mathbb{K}}^-$ large enough and $m(.)$ smooth enough, the imputed and the true underlying Y should belong to the same slice with a high probability. Thus, the set of $\mathbf{X}$'s belonging to a particular slice is now 'enriched' and consequently, improved estimation of $\mathbb{M}$ and $\mathbf{P}_r^0$ is expected. The proposed method based on a nearest neighbor approximation is also highly scalable and while other smoothing based approximations may be used, they can be computationally intensive. The SS-SIR algorithm is fairly robust to the choice of H, and $H = O(n^{\frac{1}{2}}\log\ n)$ seems to give fairly satisfactory performance. The slices may be chosen to have equal width or equal number of observations. For SIR, $n^{\frac{1}{2}}$-consistency of the

estimates are well established (Li, 1991; Duan and Li, 1991; Zhu and Ng, 1995) for various formulations under fairly general settings (without any model based assumptions). The theoretical properties of SS-SIR, although not derived here, are expected to follow similarly. Our simulation results (not shown here) further suggest that SS-SIR significantly outperforms SIR, leading to substantially improved estimation of $\boldsymbol{\theta}_0$ from the proposed methods.

1.7 Numerical Studies

1.7.1 Simulation Studies

We conducted extensive simulation studies to examine the finite sample performance of our proposed point and interval estimation procedures. Throughout, we let $n = 500$, $N = 10000$, $\mathbb{K} = 5$ and $r = 2$. We considered $p = 10$, 20 as well as $p = 2$ for which no dimension reduction was used. We generated $\mathbf{X} \sim \mathcal{N}_p[\mathbf{0}, I_p]$ and restricted $\mathbf{X}$ to $[-5, 5]^p$ to ensure its boundedness. Given $\mathbf{X} = \mathbf{x}$, we generated $Y \sim \mathcal{N}_1[m(\mathbf{x}), 1]$ for different choices of $m(\mathbf{x})$ to be discussed below. The dimension reduction step for $p = 10$ and 20 was performed using the SS-SIR algorithm with $H = 100$ slices of equal width. The estimators $\{\widehat{m}(\mathbf{x}, \widehat{\mathbf{P}}_{r,k})\}_{k=1}^{\mathbb{K}}$ were obtained using an r-dimensional local constant KS based on a Gaussian kernel with h estimated through least squares CV. The true values of the target parameter $\boldsymbol{\theta}_0$ were estimated via monte carlo with a large sample size of $50,000$. For each configuration, the results are summarized based on 500 replications.

Choices of $m(\mathbf{x})$: We first considered the case with $p = 10$ and 20, and investigated four different functional forms of $m(\mathbf{x})$ as follows:

(i) *Linear (L)*: $m(\mathbf{x}) = \mathbf{x}'\mathbf{b}_p$;

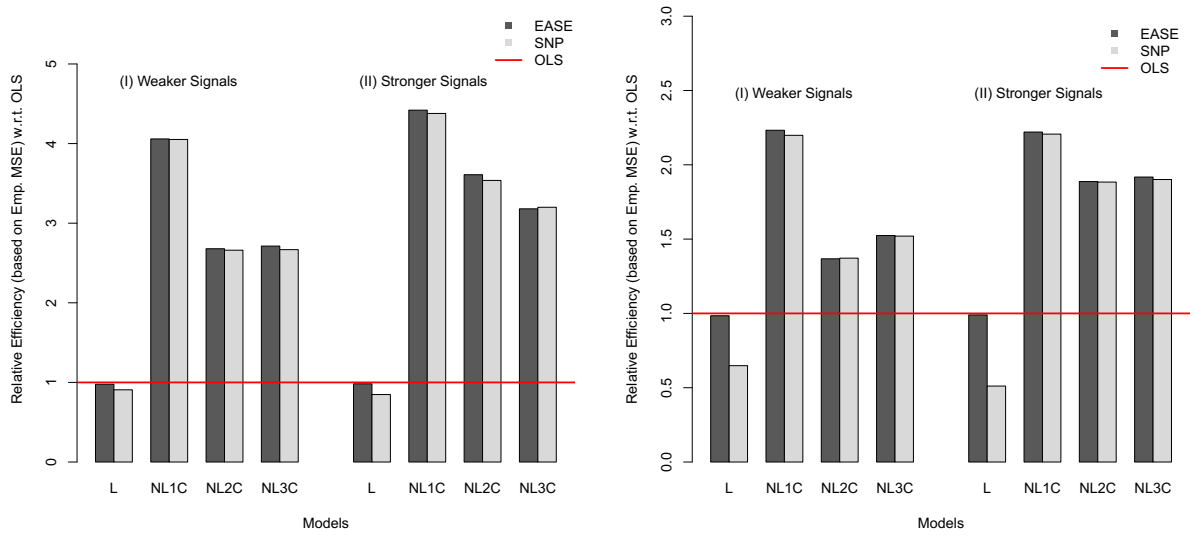
(ii) *Non-linear one component (NL1C)*: $m(\mathbf{x}) = (\mathbf{x}'\mathbf{b}_p) + (\mathbf{x}'\mathbf{b}_p)^2$;

(iii) *Non-linear two component (NL2C)*: $m(\mathbf{x}) = (\mathbf{x}'\mathbf{b}_p)(1 + \mathbf{x}'\boldsymbol{\delta}_p)$; and

(iv) *Non-linear three component (NL3C)*: $m(\mathbf{x}) = (\mathbf{x}'\mathbf{b}_p)(1 + \mathbf{x}'\boldsymbol{\delta}_p) + (\mathbf{x}'\boldsymbol{\omega}_p)^2$.

For each setting, we used two sets of $\mathbf{b}_p$: $\mathbf{b}_p^{(1)} \equiv (\mathbf{1}'_{p/2}, \mathbf{0}'_{p/2})'$ reflecting weaker signals, and $\mathbf{b}_p^{(2)} \equiv \mathbf{1}_p$ reflecting stronger signals; $\boldsymbol{\delta}_p = (\mathbf{0}'_{p/2}, \mathbf{1}'_{p/2})'$; $\boldsymbol{\omega}_p = (1, 0, 1, 0, \ldots, 1, 0)'_{p\times 1}$; and for any a, $\mathbf{1}_a = (1, \ldots, 1)'_{a\times 1}$ and $\mathbf{0}_a = (0, \ldots, 0)'_{a\times 1}$. For the non-linear models (ii)-(iv), note that the corresponding $m(\mathbf{x})$ depends on $\mathbf{x}$ through 1, 2 and 3 dimensional linear transformations of $\mathbf{x}$ respectively. Through appropriate choices of $\mathbf{b}_p$, $\boldsymbol{\delta}_p$ and $\boldsymbol{\omega}_p$, as applicable, these models can incorporate commonly encountered quadratic and interaction effects. Lastly, with $\mathbf{X}$ normally distributed and $\mathbf{P}_r$ being chosen based on SIR, results from (Li, 1991) further imply that the SNP imputation with $r = 2$ is sufficient for models (i)-(iii), and insufficient for model (iv).

We summarize in Figure 1.3 the overall relative efficiency of the proposed estimators,



(a) $p = 10$ (b) $p = 20$

Figure 1.3: Efficiencies of $\widehat{\boldsymbol{\theta}}^E_{(\mathbf{P}_r,\mathbb{K})}$ (EASE) (dark grey bars) and $\widehat{\boldsymbol{\theta}}_{(\mathbf{P}_r,\mathbb{K})}$ (SNP) (light grey bars) relative to $\widehat{\boldsymbol{\theta}}$ (OLS) (red line) with respect to the empirical MSE under models (i), (ii), (iii) and (iv) with weaker signals in the left panels and stronger signals in the right panels of each figure.

$\widehat{\boldsymbol{\theta}}_{(\mathbf{P}_r,\mathbb{K})}$ and $\widehat{\boldsymbol{\theta}}^E_{(\mathbf{P}_r,\mathbb{K})}$, compared to the OLS w.r.t. the empirical mean squared error (MSE), where for any estimator $\widehat{\boldsymbol{\theta}}^*$ of $\boldsymbol{\theta}_0$, the empirical MSE is summarized as the average of

$\|\widehat{\boldsymbol{\theta}}^* - \boldsymbol{\theta}_0\|^2$ over the 500 replications. As expected, both $\widehat{\boldsymbol{\theta}}_{(\mathbf{P}_r,\mathbb{K})}$ and $\widehat{\boldsymbol{\theta}}^E_{(\mathbf{P}_r,\mathbb{K})}$ are substantially more efficient than the OLS under model mis-specification with higher gains observed in the settings with stronger signals. Under the NL1C and NL2C models, the SS-SIR based SNP imputation is expected to be sufficient and thus both $\widehat{\boldsymbol{\theta}}_{(\mathbf{P}_r,\mathbb{K})}$ and $\widehat{\boldsymbol{\theta}}^E_{(\mathbf{P}_r,\mathbb{K})}$ should be achieving the maximal efficiency gain. Although $\widehat{\boldsymbol{\theta}}_{(\mathbf{P}_r,\mathbb{K})}$ and $\widehat{\boldsymbol{\theta}}^E_{(\mathbf{P}_r,\mathbb{K})}$ tend to have similar efficiencies under the non-linear models, their performances differ significantly under the linear setting. The EASE estimator $\widehat{\boldsymbol{\theta}}^E_{(\mathbf{P}_r,\mathbb{K})}$ achieves nearly identical efficiency as the optimal OLS estimator under this setting while $\widehat{\boldsymbol{\theta}}_{(\mathbf{P}_r,\mathbb{K})}$ could be much less efficient than the OLS, particularly with $p = 20$. This, in part attributable to over-fitting and the finite sample variation induced by estimating the SIR directions, highlights the advantage of the EASE estimator. Although the error due to estimation of $\mathbf{P}_r$ is negligible asymptotically, the variation in $\widehat{\mathbf{P}}_r$ *does* have an impact on the efficiency of the proposed estimators in finite samples. When the standard SIR is employed for dimension reduction, the resulting estimators $\widehat{\boldsymbol{\theta}}_{(\mathbf{P}_r,\mathbb{K})}$ and $\widehat{\boldsymbol{\theta}}^E_{(\mathbf{P}_r,\mathbb{K})}$ are substantially less efficient (results not shown) compared to those based on SS-SIR. Results on $p = 2$ are summarized in Figure 1.4. Only $\widehat{\boldsymbol{\theta}}_{(\mathbf{P}_r,\mathbb{K})}$ is considered in this case since the imputation is sufficient and hence $\widehat{\boldsymbol{\theta}}_{(\mathbf{P}_r,\mathbb{K})}$ is expected to achieve full efficiency gain. The efficiency results show a similar pattern to those for the cases $p = 10$ and 20 when $m(.)$ is non-linear, and the efficiency loss in the linear case is less severe compared to the settings with larger p.

To examine the performance of the proposed double CV based inference procedures, we also obtained standard error (SE) estimates and CIs for $\widehat{\boldsymbol{\theta}}_{(\mathbf{P}_r,\mathbb{K})}$ and $\widehat{\boldsymbol{\theta}}^E_{(\mathbf{P}_r,\mathbb{K})}$. In Tables 1.1-1.2, we present the bias, empirical SE (ESE), the average of the estimated SE (ASE) and the coverage probability (CovP) of the 95% CIs for each component of $\widehat{\boldsymbol{\theta}}_{(\mathbf{P}_r,\mathbb{K})}$ and $\widehat{\boldsymbol{\theta}}^E_{(\mathbf{P}_r,\mathbb{K})}$ when $p = 10$ under the linear and NL2C models. In general, we find that both $\widehat{\boldsymbol{\theta}}_{(\mathbf{P}_r,\mathbb{K})}$ and $\widehat{\boldsymbol{\theta}}^E_{(\mathbf{P}_r,\mathbb{K})}$ have negligible biases. The ASEs are close to the ESEs and the CovPs are close to the nominal level of 95%, suggesting that the proposed variance estimation procedure works well in practice with $\mathbb{K} = 5$.

Table 1.1: Bias, ESE, ASE and CovP of $\widehat{\boldsymbol{\theta}}_{(\mathbf{P}_r,\mathbb{K})}$ and $\widehat{\boldsymbol{\theta}}^{E}_{(\mathbf{P}_r,\mathbb{K})}$ for estimating $\boldsymbol{\theta}_0$ with $p = 10$ and $\mathbf{b}_p = \mathbf{b}_p^{(1)}$ under the linear model. Shown also are the bias and ESE of the OLS estimator for comparison. The true parameter value under this model is given by: $\boldsymbol{\theta}_0 = (\alpha_0, \beta_{01}, ..., \beta_{010})' = (0, \mathbf{1}_5', \mathbf{0}_5')'$, as tabulated below.

Parameter	OLS ($\widehat{\boldsymbol{\theta}}$)		SNP ($\widehat{\boldsymbol{\theta}}_{(\mathbf{P}_r,\mathbb{K})}$)				EASE ($\widehat{\boldsymbol{\theta}}^{E}_{(\mathbf{P}_r,\mathbb{K})}$)			
	Bias	ESE	Bias	ESE	ASE	CovP	Bias	ESE	ASE	CovP
$\alpha_0 = 0$	0.003	0.043	0.003	0.045	0.048	**0.96**	0.003	0.044	0.044	**0.96**
$\beta_{01} = 1$	0.000	0.045	0.012	0.047	0.049	**0.95**	-0.004	0.046	0.043	**0.93**
$\beta_{02} = 1$	-0.002	0.045	0.009	0.046	0.049	**0.94**	-0.006	0.045	0.043	**0.93**
$\beta_{03} = 1$	-0.002	0.048	0.008	0.050	0.049	**0.94**	-0.006	0.048	0.043	**0.94**
$\beta_{04} = 1$	-0.002	0.045	0.010	0.047	0.049	**0.95**	-0.006	0.046	0.043	**0.93**
$\beta_{05} = 1$	-0.004	0.044	0.007	0.046	0.049	**0.97**	-0.008	0.045	0.043	**0.93**
$\beta_{06} = 0$	0.002	0.046	0.003	0.048	0.049	**0.95**	0.002	0.046	0.043	**0.93**
$\beta_{07} = 0$	0.003	0.046	0.003	0.048	0.049	**0.94**	0.002	0.046	0.043	**0.93**
$\beta_{08} = 0$	-0.001	0.046	0.000	0.047	0.049	**0.96**	-0.001	0.046	0.044	**0.95**
$\beta_{09} = 0$	0.000	0.045	0.000	0.047	0.049	**0.95**	0.000	0.045	0.043	**0.93**
$\beta_{010} = 0$	0.000	0.045	0.001	0.047	0.049	**0.97**	0.000	0.045	0.044	**0.95**

Table 1.2: Bias, ESE, ASE and CovP of $\widehat{\boldsymbol{\theta}}_{(\mathbf{P}_r,\mathbb{K})}$ and $\widehat{\boldsymbol{\theta}}^{E}_{(\mathbf{P}_r,\mathbb{K})}$ for estimating $\boldsymbol{\theta}_0$ with $p = 10$ and $\mathbf{b}_p = \mathbf{b}_p^{(1)}$ under the NL2C model. Shown also are the bias and ESE of the OLS estimator for comparison. The true parameter value under this model is given by: $\boldsymbol{\theta}_0 = (\alpha_0, \beta_{01}, ..., \beta_{010})' = (0, \mathbf{1}_5', \mathbf{0}_5')'$, as tabulated below.

Parameter	OLS ($\widehat{\boldsymbol{\theta}}$)		SNP ($\widehat{\boldsymbol{\theta}}_{(\mathbf{P}_r,\mathbb{K})}$)				EASE ($\widehat{\boldsymbol{\theta}}^{E}_{(\mathbf{P}_r,\mathbb{K})}$)			
	Bias	ESE	Bias	ESE	ASE	CovP	Bias	ESE	ASE	CovP
$\alpha_0 = 0$	-0.017	0.238	-0.017	0.144	0.138	**0.93**	-0.016	0.144	0.137	**0.94**
$\beta_{01} = 1$	-0.002	0.274	0.012	0.165	0.163	**0.96**	0.014	0.163	0.160	**0.95**
$\beta_{02} = 1$	0.008	0.263	0.011	0.171	0.161	**0.92**	0.013	0.170	0.158	**0.92**
$\beta_{03} = 1$	0.008	0.276	0.023	0.174	0.162	**0.92**	0.026	0.175	0.159	**0.92**
$\beta_{04} = 1$	-0.001	0.282	0.011	0.167	0.162	**0.94**	0.014	0.168	0.159	**0.94**
$\beta_{05} = 1$	-0.009	0.266	0.008	0.167	0.164	**0.96**	0.012	0.167	0.161	**0.94**
$\beta_{06} = 0$	-0.008	0.277	-0.013	0.170	0.160	**0.94**	-0.015	0.170	0.156	**0.94**
$\beta_{07} = 0$	-0.002	0.264	0.000	0.165	0.159	**0.94**	-0.002	0.164	0.155	**0.95**
$\beta_{08} = 0$	0.003	0.264	0.006	0.161	0.160	**0.95**	0.008	0.157	0.157	**0.95**
$\beta_{09} = 0$	0.000	0.283	-0.001	0.170	0.160	**0.95**	-0.001	0.168	0.156	**0.94**
$\beta_{010} = 0$	-0.003	0.270	-0.002	0.156	0.160	**0.95**	-0.002	0.153	0.156	**0.94**

Under the linear model, both $\widehat{\boldsymbol{\theta}}_{(\mathbf{P}_r,\mathbb{K})}$ and $\widehat{\boldsymbol{\theta}}^{E}_{(\mathbf{P}_r,\mathbb{K})}$ have similar magnitudes of biases and standard errors as the OLS, as we expect. Under the NL2C model, compared to OLS, both $\widehat{\boldsymbol{\theta}}_{(\mathbf{P}_r,\mathbb{K})}$ and $\widehat{\boldsymbol{\theta}}^{E}_{(\mathbf{P}_r,\mathbb{K})}$ are overwhelmingly more efficient at the cost of only negligibly larger

biases across all components of $\boldsymbol{\theta}_0$.

To gain further insights into the potential overfitting bias, we also obtained results for $\widehat{\boldsymbol{\theta}}_{(\mathbf{P}_r,1)}$ when no CV was used. The empirical absolute bias for $\widehat{\boldsymbol{\theta}}_{(\mathbf{P}_r,1)}$ ranged from 0 to 0.011 under the linear model, and from 0.002 to 0.08 under the NL2C model. The biases, especially for the NL2C model, are often substantially larger than that of $\widehat{\boldsymbol{\theta}}_{(\mathbf{P}_r,\mathbb{K})}$ and $\widehat{\boldsymbol{\theta}}^{E}_{(\mathbf{P}_r,\mathbb{K})}$, highlighting the importance of CV in reducing over-fitting bias.

Simulation Results for Two-Dimensional Covariates: For $p = 2$, we investigated three functional forms of $m(\mathbf{x})$ as follows: (a) *Linear (Lin.)*: $m(\mathbf{x}) = x_1 + x_2$; (b) *Non-Linear Quadratic (NL-Quad.)*: $m(\mathbf{x}) = x_1 + x_2 + \gamma(x_1^2 + x_2^2)$; and (c) *Non-Linear Interaction (NL-Int.)*: $m(\mathbf{x}) = x_1 + x_2 + \lambda x_1 x_2$; where $\mathbf{x} = (x_1, x_2)'$. For each of γ and λ, we chose two values: $\gamma = \gamma^{(1)} = 0.3$ and $\lambda = \lambda^{(1)} = 0.5$ reflecting weaker signals, and $\gamma = \gamma^{(2)} = 1$ and $\lambda = \lambda^{(2)} = 1$ reflecting stronger signals.

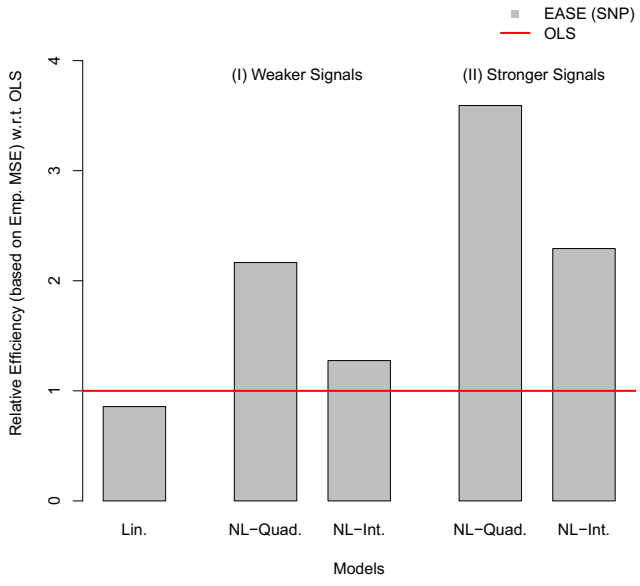


Figure 1.4: Efficiencies of $\widehat{\boldsymbol{\theta}}^{E}_{(\mathbf{P}_r,\mathbb{K})}$ (EASE) $\equiv \widehat{\boldsymbol{\theta}}_{(\mathbf{P}_r,\mathbb{K})}$ (SNP) (grey bars) relative to $\widehat{\boldsymbol{\theta}}$ (OLS) (red line) with respect to the empirical mean squared error under models (a), (b) and (c) for $p = 2$ with weaker signals for the non-linear models in the left panel and stronger signals in the right panel.

1.7.2 Application to EMR Data

We applied our proposed procedures to an EMR genetic study of rheumatoid arthritis (RA), a systemic autoimmune disease (AD), conducted at the Partners HealthCare. The study cohort consists of 4453 patients previously identified as having RA based on a highly accurate algorithm as described in Liao et al. (2010). Multiple ADs are known to be strongly associated with each other through shared genetic risk factors, so that a RA patient may be genetically predisposed towards other ADs like systemic lupus erythematosus (SLE) that can be fatal. Our primary goal here was to understand this genetically shared auto-immunity based on the available data.

The outcome of interest is a SLE genetic risk score (GRS) constructed as a weighted sum of indicators of previously identified SLE risk alleles with weights being the corresponding published log odds ratios as described in Liao et al. (2013). We relate the SLE GRS to a set of 14 clinical variables $\mathbf{X}$ related to ADs. The covariates include gender, race, presence of radiological evidence of bone erosion (erosion), and two lab tests including antibodies to cyclic citrullinated peptide (anti-CCP) and rheumatoid factor (RF) that are routinely checked for RA patients to assess the disease progression. The two lab tests are not always ordered for all patients and hence we coded them as 4 binary variables: (i) ccp indicating anti-CCP positivity, (ii) ccp.miss indicating whether anti-CCP was checked, (iii) rf indicating RF positivity, and (iv) rf.miss indicating if RF was checked. We additionally included 3 binary variables representing ever mentions of anti-tumor necrosis factor (anti-TNF), methotrexate and seropositive, as well as total number of mentions of 4 ADs namely RA, SLE, psoriatic arthritis (PsA) and juvenile rheumatoid arthritis (JRA) in each patient's clinical notes, all extracted via natural language processing (NLP). The count variables were transformed as: $x \to \log(1+x)$ to increase stability of the model fitting. Since obtaining the GRS for a patient would require expensive genotyping, the outcomes were available only for a random subset of $n = 1160$ patients, thereby leading to a SS set-up.

We obtained both the OLS and the EASE estimators based on the observed data. To

implement EASE, we used $\mathbb{K} = 5$, $r = 2$ and $\mathbf{P}_r$ estimated based on SS-SIR using $H = 150$ slices of equal width. In Table 1.3, we present coordinate-wise estimates of the regression parameters along with their estimated SE and the corresponding p-values based on them.

Table 1.3: Estimates (Est.) of the regression coefficients based on the OLS ($\widehat{\boldsymbol{\theta}}$) and the EASE ($\widehat{\boldsymbol{\theta}}^{E}_{(\mathbf{P}_r,\mathbb{K})}$) estimators along with their estimated standard errors (SE) and the corresponding p-values (Pval.) associated with testing the null effect for each of the predictors. Shown also are the coordinate-wise relative efficiencies (RE) of the EASE compared to the OLS.

Predictors	OLS ($\widehat{\boldsymbol{\theta}}$)			EASE ($\widehat{\boldsymbol{\theta}}^{E}_{(\mathbf{P}_r,\mathbb{K})}$)			RE
	Est.	SE	Pval.	Est.	SE	Pval.	
gender	0.002	0.004	0.524	0.003	0.004	0.401	1.104
race	-0.015	0.003	0.000	-0.014	0.003	0.000	1.138
ccp	0.008	0.007	0.288	0.008	0.007	0.235	1.103
rf	-0.010	0.005	0.065	-0.010	0.005	0.043	1.105
ccp.miss	0.003	0.006	0.639	0.003	0.005	0.585	1.081
rf.miss	-0.006	0.005	0.216	-0.007	0.005	0.160	1.079
erosion	-0.002	0.003	0.504	-0.002	0.003	0.463	1.099
anti-TNF	0.001	0.003	0.829	0.001	0.003	0.802	1.088
methotrexate	-0.006	0.003	0.058	-0.005	0.003	0.052	**1.434**
seropositive	0.002	0.004	0.682	0.002	0.004	0.652	1.125
RA	0.001	0.002	0.339	0.001	0.001	0.353	1.096
SLE	0.004	0.005	0.448	0.003	0.005	0.487	1.149
PsA	0.002	0.006	0.746	0.002	0.005	0.652	1.163
JRA	0.007	0.006	0.282	0.007	0.004	0.100	**2.193**

The point estimates of $\boldsymbol{\theta}_0$ based on $\widehat{\boldsymbol{\theta}}$ and $\widehat{\boldsymbol{\theta}}^{E}_{(\mathbf{P}_r,\mathbb{K})}$ in Table 1.3 are all quite close which is desirable and reassuring as it establishes, in a real data, the consistency and stability of EASE. Further, the estimated relative efficiencies are all greater than 1 indicating the improved efficiency of $\widehat{\boldsymbol{\theta}}^{E}_{(\mathbf{P}_r,\mathbb{K})}$ over $\widehat{\boldsymbol{\theta}}$. The efficiency gains for the NLP variables JRA and methotrexate are as high as 120% and 43%. The positive associations between all 4 counts of AD mentions and the SLE GRS, although not statistically significant, suggest that patients with these ADs are also more likely to be genetically predisposed towards SLE. This is consistent with the theory of shared auto-immunity in the literature (Alarcón-Segovia, 2005; Cotsapas et al., 2011). For example, the PTPN22 polymorphism is known to be associated

with multiple ADs including RA, SLE and type 1 diabetes (Chung and Criswell, 2007).

1.8 Discussion

We have developed in this paper an efficient and adaptive estimation strategy for the SS linear regression problem. The adaptive property possessed by the proposed EASE is crucial for advocating 'safe' use of the unlabeled data and is often unaddressed in the existing literature. In general, the magnitude of the efficiency gain with EASE depends on the inherent degree of non-linearity in $\mathbb{E}(Y|\mathbf{X})$ and the extent of sufficiency of the underlying SNP imputation. In particular, if the imputation is sufficient or the working linear model is correct, $\widehat{\boldsymbol{\theta}}^{E}_{(\mathbf{P}_r,\mathbb{K})}$ is further optimal among a wide class of estimators. We have obtained theoretical results along with influence function expansions for $\widehat{\boldsymbol{\theta}}_{(\mathbf{P}_r,\mathbb{K})}$ and $\widehat{\boldsymbol{\theta}}^{E}_{(\mathbf{P}_r,\mathbb{K})}$ substantiating all our claims and also validated them based on numerical studies. The double CV method further facilitates accurate inference, overcoming potential over-fitting issues in finite samples due to smoothing.

The proposed SNP imputation, the key component of EASE, apart from being flexible and scalable, enjoys several useful properties. The refitting step and CV play a crucial role in reducing the bias of $\widehat{\boldsymbol{\theta}}_{(\mathbf{P}_r,\mathbb{K})}$, and for $\mathcal{T} := \text{KS}$ in particular, eradicate any under-smoothing or higher order kernel requirements: two undesirable, yet often inevitable, conditions required for $n^{\frac{1}{2}}$-consistency of two-step estimators based on a first step of smoothing. Theorem 1.4, apart from showing the distinct behaviour of $\widehat{\boldsymbol{\theta}}_{(\mathbf{P}_r,1)}$ compared to $\widehat{\boldsymbol{\theta}}_{(\mathbf{P}_r,\mathbb{K})}$ for $\mathbb{K} \geq 2$, also highlights the key role of CV in completely removing kernel order restrictions, apart from addressing over-fitting issues. The error rates in the results of Theorems 1.3-1.4 are quite sharp and account for any estimation error from $\widehat{\mathbf{P}}_r$. The regularity conditions required are also fairly mild and standard in the literature. The continuity assumption on $\mathbf{X}$ in sections 1.4.1 and 1.5 is mostly for the convenience of proofs, and the results continue to hold for more general $\mathbf{X}$. Lastly, while we have focussed here on linear regression for simplicity, our methods can indeed be easily adapted to other regression problems such as logistic regression

for binary outcomes.

We end with a comment on the *choice* of $\mathbb{K} \geq 2$ in $\widehat{\boldsymbol{\theta}}_{(\mathbf{P}_r,\mathbb{K})}$. While (1.13) holds for any $\mathbb{K} \geq 2$, the error term in (1.13) depends on $\mathbb{K}$ through $c_{n_{\mathbb{K}}^-}$ and more precisely, through $\widetilde{c}_{n_{\mathbb{K}}^-} = \mathbb{K}^{\frac{1}{2}} c_{n_{\mathbb{K}}^-}$. Since $\mathbb{K}$ is fixed, $c_{n_{\mathbb{K}}^-}$ and $\widetilde{c}_{n_{\mathbb{K}}^-}$ are asymptotically equivalent. But for a given n, $c_{n_{\mathbb{K}}^-}$ is expected to decrease with $\mathbb{K}$, while $\widetilde{c}_{n_{\mathbb{K}}^-}$ is likely to increase. It is however desirable that both are small since $c_{n_{\mathbb{K}}^-}$ inherently controls the efficiency of the SNP imputation, while $\widetilde{c}_{n_{\mathbb{K}}^-}$ directly controls the bias of $\widehat{\boldsymbol{\theta}}_{(\mathbf{P}_r,\mathbb{K})}$. Hence, a reasonable choice of $\mathbb{K} \geq 2$ may be based on minimizing: $(c^2_{n_{\mathbb{K}}^-} + \lambda \widetilde{c}^2_{n_{\mathbb{K}}^-})$ for some $\lambda \geq 0$. Since the (first order) asymptotic variance of $\widehat{\boldsymbol{\theta}}_{(\mathbf{P}_r,\mathbb{K})}$ is independent of $\mathbb{K}$, this is equivalent to a penalized minimization of the asymptotic MSE of $\widehat{\boldsymbol{\theta}}_{(\mathbf{P}_r,\mathbb{K})}$ with λ denoting the weightage of the (lower order) bias relative to the (first order) variance. In general, the optimal $\mathbb{K}$ should be inversely related to λ. Conversely, choice of any $\mathbb{K}$ may be viewed to have an associated regularization effect (through λ) resulting in a 'variance-bias trade-off' with smaller $\mathbb{K}$ leading to lower bias at the cost of some efficiency, and higher $\mathbb{K}$ leading to improved efficiency in lieu of some bias. In practice, we find that $\mathbb{K} = 5$ works well, and $\mathbb{K} = 10$ tends to give slightly smaller MSE at the cost of increased bias.

A Unified Framework for Efficient and Adaptive Semi-Supervised Estimation

Abhishek Chakrabortty and Tianxi Cai

Department of Biostatistics

Harvard University

2.1 Summary

We consider a variety of estimation problems under semi-supervised (SS) settings, wherein the available data typically consists of: (i) a small or moderate sized 'labeled' data, and (ii) a much larger sized 'unlabeled' data. Such data arises naturally from settings where the outcome, unlike the covariates, is expensive to obtain, a frequent scenario in modern studies involving large databases like electronic medical records (EMR). For such SS estimation problems, it is often of interest to investigate if and when the unlabeled data can be exploited to improve estimation of the parameter of interest, compared to supervised approaches that are based on only the labeled data.

In this paper, adopting a semi-parametric perspective, we provide a unified framework for SS M/Z-estimation problems based on general estimating equations (EEs), and propose a family of 'Efficient and Adaptive Semi-Supervised Estimators' (EASE) that are always at least as efficient as the supervised estimator, and more efficient whenever the information from the unlabeled data is actually related to the parameter of interest. This adaptive property, often unaddressed in the existing literature, is quite crucial for advocating 'safe' use of the unlabeled data. The construction of EASE essentially corresponds to a (non-parametric) imputation based approach. For a simpler subclass of EEs, including those corresponding to estimation problems for most standard generalized linear (working) models, we provide a more flexible imputation strategy involving use of dimension reduction techniques, if desired, to avoid high dimensional smoothing. As a special case of our proposed framework, we also address the SS version of the 'sliced inverse regression' (SIR) problem, useful for sufficient dimension reduction. We provide explicit theoretical results including influence function expansions, as well as techniques for inference based on EASE, establishing all our claims (including semi-parametric optimality of EASE under some scenarios), followed by extensive simulation studies for logistic regression as well as SIR, and application to a real EMR dataset using logistic regression.

2.2 Introduction

In recent years, semi-supervised learning (SSL) has emerged as an exciting new area of research in statistics and machine learning (Chapelle et al., 2006; Zhu, 2008). A typical semi-supervised (SS) setting is characterized by two types of available data: (i) a small or moderate sized 'labeled' data, $\mathcal{L}$, containing observations for both an outcome Y and a set of covariates $\mathbf{X}$ of interest, and (ii) an 'unlabeled' data, $\mathcal{U}$, of *much larger* size but having observations *only* for the covariates $\mathbf{X}$. By virtue of its large size, $\mathcal{U}$ essentially gives us the distribution of $\mathbf{X}$, denoted henceforth by $\mathbb{P}_{\mathbf{X}}$. Such a setting arises naturally whenever the covariates are easily available so that unlabeled data is plentiful, but the outcome is costly or difficult to obtain, thereby limiting the size of $\mathcal{L}$. This scenario is directly relevant to a variety of practical problems, especially in the modern 'big data' era, with massive unlabeled datasets (often electronically recorded) becoming increasingly available and tractable. A few familiar examples include machine learning problems like text mining, web page classification, speech recognition, natural language processing, and among biomedical applications, the analysis of electronic medical records (EMR) data, where SSL can be particularly of great use. We refer the interested reader to Chakrabortty and Cai (2015) for further discussions.

SSL primarily distinguishes from standard supervised methods by making use of $\mathcal{U}$, an information that is ignored by the latter. The ultimate question of interest in SSL is to investigate if and when this information can be exploited to improve the efficiency over a given supervised approach. It is important to note that while the SS set-up can be viewed as a missing data problem, it is quite different from a standard missing data setting as the probability of missingness tends to 1 in SSL (so that the 'positivity assumption' typically assumed in the classical missing data literature is violated here). Interestingly, characterization of the missingness mechanism, although quite crucial, has often stayed implicit in the SSL literature (Lafferty and Wasserman, 2007). Nevertheless, it has mostly been assumed as 'missing completely at random' which is typically the case, with the labeled data being obtained from labeling a random subset, selected by design, from a large unlabeled dataset. It is also

worth noting that the analysis of SS settings under more general missingness mechanisms is much more complicated owing to the violation of the positivity assumption. The theoretical nuances of SSL including its scope and the consequences of using the unlabeled data have been studied to some extent by Castelli and Cover (1995, 1996) and later, more generally by Lafferty and Wasserman (2007). In recent years, several graph based non-parametric SSL approaches for prediction have also been proposed (Zhu, 2005; Belkin et al., 2006) relying, implicitly or explicitly, on assumptions relating $\mathbb{P}_{\mathbf{X}}$ to $\mathbb{P}_{Y|\mathbf{X}}$, the conditional distribution of $Y \mid \mathbf{X}$, that have been characterized more formally in Lafferty and Wasserman (2007).

However, SS estimation problems, especially from a semi-parametric perspective, without making unnecessary assumptions regarding the underlying data generating mechanism, has been somewhat less studied. Among existing parametric methods, perhaps the most well-known is the 'generative model' approach for classification (Nigam et al., 2000; Nigam, 2001), based on modeling the joint distribution of $(Y, \mathbf{X})$ as a mixture of parametric models, thereby implicitly relating $\mathbb{P}_{Y|\mathbf{X}}$ to $\mathbb{P}_{\mathbf{X}}$. However, these approaches depend strongly on the validity of the assumed mixture model, violation of which can actually *degrade* their performance compared to supervised approaches (Cozman and Cohen, 2001; Cozman et al., 2003).

In general, it has been noted (Zhang and Oles, 2000; Seeger, 2002) that for SS estimation problems involving parametric regression and/or likelihood based *working* models for $\mathbb{P}_{Y|\mathbf{X}}$, one *cannot* possibly gain through SSL by using the knowledge of $\mathbb{P}_{\mathbf{X}}$ if the assumed working model is correct and the parameter of interest is not related to $\mathbb{P}_{\mathbf{X}}$. On the other hand, under model mis-specification, the target parameter may inherently *depend* on $\mathbb{P}_{\mathbf{X}}$, and thus imply the potential utility of $\mathcal{U}$ in improving the estimation. This notion can be further generalized for any SS estimation problem as follows: if we are interested in estimating a parameter $\boldsymbol{\theta}_0 \equiv \boldsymbol{\theta}_0(\mathbb{P})$, a functional of the underlying data generating distribution $\mathbb{P} \equiv (\mathbb{P}_{Y \mid \mathbf{X}}, \mathbb{P}_{\mathbf{X}})$, then SS approaches based on the use of $\mathcal{U}$ can lead to improved estimation of $\boldsymbol{\theta}_0$ *only if* $\boldsymbol{\theta}_0$ and $\mathbb{P}_{\mathbf{X}}$ are somehow related. However, inappropriate usage of $\mathcal{U}$ may lead to degradation of the estimation precision. This therefore signifies the need for developing *robust* and efficient

SS estimators that are *adaptive* to the knowledge of $\mathbb{P}_{\mathbf{X}}$ being actually helpful for estimating $\boldsymbol{\theta}_0$. To the best of our knowledge, work done along these lines is scarce in the SSL literature, and we hope our results in this paper contributes to some extent towards filling this gap.

In this paper, we make a modest attempt towards a general characterization of SS M/Z-estimation problems from a semi-parametric perspective, and providing a unified framework for construction of corresponding efficient and adaptive estimation strategies, in the sense discussed above. In particular, we consider SS estimation problems based on general estimating equations (EEs) that are routinely encountered in various statistical applications. For such problems, we propose a family of 'Efficient and Adaptive Semi-Supervised Estimators' (EASE) that are always at least as efficient as the supervised estimator, and more efficient whenever the information on $\mathbb{P}_{\mathbf{X}}$, available through $\mathcal{U}$, is actually helpful for the purpose of estimating the parameter of interest. This adaptive property, often unaddressed in the existing literature, is quite crucial for advocating 'safe' use of the unlabeled data. The construction of EASE essentially corresponds to a (non-parametric) imputation based approach. For a simpler subclass of EEs, including those corresponding to estimation problems for most standard generalized linear (working) models, we provide a more flexible imputation strategy involving use of dimension reduction techniques, if desired, to avoid high dimensional smoothing. As a special case of our proposed framework, we also address the SS version of the 'sliced inverse regression' (SIR) problem, useful for sufficient dimension reduction. We provide explicit theoretical results including influence function expansions, as well as techniques for inference based on EASE, establishing all our claims (including semi-parametric optimality of EASE under some scenarios), followed by extensive simulation studies for logistic regression as well as SIR, and application to a real EMR dataset.

The rest of this paper is organized as follows. In section 2.3, we formally introduce and formulate SS M/Z-estimation problems, followed by the general construction of EASE, as well as more flexible constructions, along with necessary modifications, for a simpler subclass of EEs. Techniques based on cross-validation (CV) for inference using EASE are also briefly

discussed, followed by extensive simulation studies and application to a real EMR dataset for logistic regression. In section 2.4, we discuss efficient SS estimators for the SIR problem, as a special case of the framework developed in section 2.3, followed by simulation results. For all the theoretical results obtained in this paper, the proofs are not too difficult, and should be especially straightforward upon use of standard results from M-estimation theory, available in Van der Vaart (2000) for instance, as well as the results obtained and techniques used in Chakrabortty and Cai (2015). The proofs are therefore skipped here for brevity.

2.3 Semi-Supervised M-Estimation

Data Representation: Let $Y \in \mathbb{R}$ denote the outcome random variable with a support $\mathcal{Y} \subseteq \mathbb{R}$ of arbitrary nature (continuous and/or discrete), and $\mathbf{X} \in \mathbb{R}^p$ denote the vector of covariates, where p is fixed, and let $\mathbf{Z} = (Y, \mathbf{X}')'$. Then, the entire data available for analysis in SS settings can be represented as: $\mathbb{S} = (\mathcal{L} \cup \mathcal{U})$, where $\mathcal{L} = \{\mathbf{Z}_i \equiv (Y_i, \mathbf{X}_i')' : i = 1, ..., n\}$ consisting of n independent and identically distributed (i.i.d.) observations from the joint distribution $\mathbb{P} \equiv (\mathbb{P}_{Y \mid \mathbf{X}}, \mathbb{P}_{\mathbf{X}})$ of $\mathbf{Z}$ denotes the labeled data, $\mathcal{U} = \{\mathbf{X}_j : j = 1, ..., N\}$ consisting of N i.i.d. realizations of $\mathbf{X}$ denotes the unlabeled data, and further, $\mathcal{L} \perp\!\!\!\perp \mathcal{U}$.

Basic Assumptions: (a) We assume that $N \gg n$ and hence as $n \to \infty$, the proportion of observed outcomes, n/N, tends to 0, which makes SSL different from a classical missing data problem where this proportion is typically assumed to be bounded away from zero. (b) The underlying Y for subjects in $\mathcal{U}$ are assumed to be 'missing completely at random', so that $\mathbf{Z} \sim \mathbb{P}$ for all subjects in $\mathcal{S}$. (c) We assume throughout that $\mathbf{Z}$ has finite 2^{nd} moments, $\text{Var}(\mathbf{X})$ is positive definite and $\mathbf{X}$ has a compact support $\mathcal{X} \subseteq \mathbb{R}^p$. Let $\mathcal{L}_2(\mathbb{P})$ denote the space of all $\mathbb{R}$-valued measurable functions of $\mathbf{Z}$ having finite L_2 norm with respect to (w.r.t.) $\mathbb{P}$, and let $\mathbb{E}(.)$ denote expectation w.r.t. $\mathbb{P}$. (d) Since the moments of $\mathbf{X}$ are essentially known due to the large (potentially infinite) size of $\mathcal{U}$, we also assume without loss of generality (w.l.o.g.) that $\mathbb{E}(\mathbf{X}) = \mathbf{0}$ and $\text{Var}(\mathbf{X}) = I_p$, where I_p denotes the $(p \times p)$ identity matrix.

Notational Conventions: (a) Throughout, $\|.\|$ will denote the standard L_2 vector norm in $\mathbb{R}^d$ for any $d \geq 1$. For any $\mathbf{u} \in \mathbb{R}^d$, $\mathbf{u}_{[j]}$ will denote its j^{th} coordinate $\forall\ 1 \leq j \leq d$, $\overrightarrow{\mathbf{u}}$ will denote $(1, \mathbf{u}')' \in \mathbb{R}^{d+1}$, and for any matrix $M_{d\times d}$, $M_{[i,j]}$ will denote its $(i,j)^{th}$ entry $\forall\ 1 \leq i,j \leq d$. Further, for any $\epsilon \geq 0$, and any $\mathbf{u} \in \mathbb{R}^d$, we let $\mathcal{B}(\mathbf{u};\epsilon) = \{\mathbf{v} : \|\mathbf{u}-\mathbf{v}\| \leq \epsilon\} \subseteq \mathbb{R}^d$ denote the L_2 ball of radius ϵ around $\mathbf{u}$. (b) For any $d \geq 1$, we denote by $\mathcal{N}_d(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ the $d-$variate normal distribution with mean $\boldsymbol{\mu} \in \mathbb{R}^d$ and dispersion matrix $\boldsymbol{\Sigma}_{d\times d}$. For any measurable (possibly vector-valued) function $f(\mathbf{Z}) \in \mathbb{R}^d$ of $\mathbf{Z}$, for any $d \geq 1$, such that $\mathbb{E}(\|f(\mathbf{Z})\|^2) < \infty$, we denote by $\boldsymbol{\Sigma}\{f(.)\}$ the $d \times d$ matrix $\mathrm{Var}\{f(\mathbf{Z})\}$. (c) Let $\mathbb{P}_n(.)$ denote the empirical probability measure based on $\mathcal{L}$. Further, let $f(\mathbf{Z}) \in \mathbb{R}^d$, for any $d \geq 1$, be any measurable function of $\mathbf{Z}$, where $f(.)$ itself *can* be a random data driven function based on $\mathcal{L}$. Then we define: $\mathbb{P}_n\{f(\mathbf{Z})\} = n^{-1}\sum_{i=1}^n f(\mathbf{Z}_i)$, and $\mathbb{G}_n\{f(\mathbf{Z})\} = n^{\frac{1}{2}}[\mathbb{P}_n\{f(\mathbf{Z})\} - \mathbb{E}_{\mathbf{Z}}\{f(\mathbf{Z})\}]$, the $n^{\frac{1}{2}}$-scaled empirical process indexed by $f(.)$, where $\mathbb{E}_{\mathbf{Z}}\{f(\mathbf{Z})\} \equiv \int f(\mathbf{z})\ d\mathbb{P}(\mathbf{z})$ denotes expectation of $f(\mathbf{Z})$ w.r.t. $\mathbf{Z} \sim \mathbb{P}$ with $f(.)$ treated as fixed (even though it may be random).

The M-Estimation Problem: Let $\boldsymbol{\psi}(Y, \mathbf{X}, \boldsymbol{\theta}) \in \mathbb{R}^d$, for any fixed $d \geq 1$, be an *estimating function* of interest, where $\boldsymbol{\theta} \in \Theta$, for some appropriate parameter space $\Theta \subseteq \mathbb{R}^d$, and assume $\mathbb{E}\{\|\boldsymbol{\psi}(Y, \mathbf{X}, \boldsymbol{\theta})\|^2\} < \infty\ \forall\ \boldsymbol{\theta} \in \Theta$. Let $\boldsymbol{\psi}_0(\boldsymbol{\theta}) = \mathbb{E}\{\boldsymbol{\psi}(Y, \mathbf{X}, \boldsymbol{\theta})\}\ \forall\ \boldsymbol{\theta} \in \Theta$. Suppose, we are interested in estimating the unknown parameter/functional $\boldsymbol{\theta}_0 \equiv \boldsymbol{\theta}_0(\mathbb{P})$ given by:

$$\boldsymbol{\theta}_0 \text{ is the solution in } \boldsymbol{\theta} \in \Theta \text{ to the } \textit{estimating equation}\text{: } \boldsymbol{\psi}_0(\boldsymbol{\theta}) \equiv \mathbb{E}\{\boldsymbol{\psi}(Y, \mathbf{X}, \boldsymbol{\theta})\} = \mathbf{0}. \quad (2.1)$$

Given the generality of the framework considered here, the existence and uniqueness of such a $\boldsymbol{\theta}_0$ will be implicitly assumed. For most commonly encountered M-estimation problems, $\boldsymbol{\psi}_0(.)$ typically arises as the derivative of some convex loss, or concave log-likelihood function, in which case, as long as those functions are smooth enough, the existence and uniqueness of $\boldsymbol{\theta}_0$ can typically be guaranteed quite easily, and moreover can be done so with $\boldsymbol{\theta}_0$ allowed to vary over the whole of $\mathbb{R}^d$ (in which case the parameter space Θ in (2.1) can be chosen to be any $\mathcal{B}(\boldsymbol{\theta}_0, \epsilon)$ for an arbitrary $\epsilon > 0$). We are interested in efficient and adaptive SS estimation

of $\boldsymbol{\theta}_0$ based on the *entire* available data $\mathcal{S}$, compared to the supervised M-estimator based on $\mathcal{L}$ only, which is given by: $\widehat{\boldsymbol{\theta}}$, the solution in $\boldsymbol{\theta} \in \Theta$ to: $\boldsymbol{\psi}_n(\boldsymbol{\theta}) \equiv n^{-1} \sum_{i=1}^{n} \boldsymbol{\psi}(Y_i, \mathbf{X}_i, \boldsymbol{\theta}) = \mathbf{0}$. It might be also helpful to explicitly define the underlying semi-parametric (almost non-parametric) model that effectively characterizes our SS set-up, which is given by:

$$\mathcal{M}_{\mathbf{X}} = \{\mathbb{P} : \mathbb{P}_{\mathbf{X}} \text{ is known, and } \mathbb{P}_{Y \mid \mathbf{X}} \text{ unrestricted (upto our basic assumptions)}\}. \tag{2.2}$$

Noting that the knowledge of $\mathbb{P}_{\mathbf{X}}$ is indeed (almost) available to us through $\mathcal{U}$, the model $\mathcal{M}_{\mathbf{X}}$ therefore essentially represents the most unrestricted (upto our basic assumptions) class of distributions $\mathbb{P}$ we can have under the SS set-up. We wish to clarify that $\mathcal{M}_{\mathbf{X}}$ would be the underlying model (apart from some mild smoothness restrictions to be imposed later on, as required) we consider for the estimation of the functional $\boldsymbol{\theta}_0$, and the goal would be to obtain efficient SS estimators of $\boldsymbol{\theta}_0$ for *every* sub-model in $\mathcal{M}_{\mathbf{X}}$. We next introduce an important definition regarding the notion of $\mathbb{P}_{\mathbf{X}}$ being actually related/helpful for the estimation of $\boldsymbol{\theta}_0$.

Definition 2.1. Let $\boldsymbol{\phi}(\mathbf{X}, \boldsymbol{\theta}) = \mathbb{E}\{\boldsymbol{\psi}(Y, \mathbf{X}, \boldsymbol{\theta}) \mid \mathbf{X}\} \; \forall \; \boldsymbol{\theta} \in \Theta$, so that $\boldsymbol{\psi}_0(\boldsymbol{\theta}) = \mathbb{E}_{\mathbf{X}}\{\boldsymbol{\phi}(\mathbf{X}, \boldsymbol{\theta})\}$, where $\mathbb{E}_{\mathbf{X}}(.)$ denotes expectation w.r.t. $\mathbb{P}_{\mathbf{X}}$. Then, with $\boldsymbol{\theta}_0$ as in (2.1), we define $\mathbb{P}_{\mathbf{X}}$ to be *informative* for estimating $\boldsymbol{\theta}_0$ under $\mathcal{M}_{\mathbf{X}}$, if for some measurable set $\mathcal{A} \subseteq \mathcal{X}$ with $\mathbb{P}_{\mathbf{X}}(\mathcal{A}) > 0$, $\boldsymbol{\phi}(\mathbf{x}, \boldsymbol{\theta}_0) \neq 0 \; \forall \; \mathbf{x} \in \mathcal{A}$, and *non-informative* if $\boldsymbol{\phi}(\mathbf{X}, \boldsymbol{\theta}_0) = \mathbf{0}$ almost surely (a.s.) $[\mathbb{P}_{\mathbf{X}}]$.

With $\boldsymbol{\theta}_0$ in (2.1) being the unique solution to $\boldsymbol{\psi}_0(\boldsymbol{\theta}) \equiv \mathbb{E}_{\mathbf{X}}\{\boldsymbol{\phi}(\mathbf{X}, \boldsymbol{\theta})\} = \mathbf{0}$ over $\boldsymbol{\theta} \in \Theta$, note that when $\boldsymbol{\phi}(\mathbf{X}, \boldsymbol{\theta}_0) = \mathbf{0}$ a.s. $[\mathbb{P}_{\mathbf{X}}]$, then indeed $\mathbb{P}_{\mathbf{X}}$ no longer plays any role in the definition of $\boldsymbol{\theta}_0$, as even without the outside $\mathbb{E}_{\mathbf{X}}(.)$ in the representation of $\boldsymbol{\psi}_0(\boldsymbol{\theta})$ above, the expression is $\mathbf{0}$ a.s. $[\mathbb{P}_{\mathbf{X}}]$ at $\boldsymbol{\theta} = \boldsymbol{\theta}_0$, and this expression essentially depends only on $\mathbb{P}_{Y \mid \mathbf{X}}$ which is not necessarily related to $\mathbb{P}_{\mathbf{X}}$, certainly not under $\mathcal{M}_{\mathbf{X}}$, unless further assumptions are made. On the other hand, if $\mathbb{P}_{\mathbf{X}}\{\boldsymbol{\phi}(\mathbf{x}, \boldsymbol{\theta}_0) \neq 0\} > 0$, then $\mathbb{P}_{\mathbf{X}}$ indeed plays a role in defining $\boldsymbol{\theta}_0$ as the outside $\mathbb{E}_{\mathbf{X}}(.)$ in the representation of $\boldsymbol{\psi}_0(\boldsymbol{\theta})$ above is now required to ensure that $\boldsymbol{\psi}_0(\boldsymbol{\theta}) = \mathbf{0}$ at $\boldsymbol{\theta} = \boldsymbol{\theta}_0$. For most common parametric regression (working) models imposing a restriction on $\mathbb{P}_{Y \mid \mathbf{X}}$, through the conditional mean $m(\mathbf{X}) \equiv \mathbb{E}(Y \mid \mathbf{X})$, given by: $m(\mathbf{X}) = g(\mathbf{X}, \boldsymbol{\theta})$ for

some $g(.)$, and $\boldsymbol{\theta} \in \mathbb{R}^d$, the underlying EEs used for estimating the unknown regression parameter typically corresponds to: $\boldsymbol{\psi}(Y, \mathbf{X}, \boldsymbol{\theta}) = \mathbf{h}(\mathbf{X}, \boldsymbol{\theta})\{Y - g(\mathbf{X}, \boldsymbol{\theta})\}$, for some $\mathbf{h}(.)$. In these cases, the notion of $\mathbb{P}_{\mathbf{X}}$ being related or unrelated to $\boldsymbol{\theta}_0$, and therefore being helpful or not for efficient SS estimation of $\boldsymbol{\theta}_0$, essentially boils down to the more familiar notion of the underlying working model for $m(.)$ being *correct* or *mis-specified*. The definition in 2.1 is largely inspired from, and is a generalization of, this familiar (but more restrictive) notion. We next present the general construction of our EASE estimators, and their properties.

Efficient and Adaptive Semi-Supervised Estimators (EASE): Let $\widehat{\boldsymbol{\phi}}(\mathbf{x}, \boldsymbol{\theta})$ denote *any* reasonable estimator, based on $\mathcal{L}$, of $\boldsymbol{\phi}(\mathbf{x}, \boldsymbol{\theta}) \; \forall \; \mathbf{x} \in \mathcal{X}, \boldsymbol{\theta} \in \Theta$. We then define a *modified* EE, $\boldsymbol{\psi}^*_{n,N}(\boldsymbol{\theta})$, and then the EASE estimator $\widehat{\boldsymbol{\theta}}^*$ as the solution to that EE, as follows.

$$\text{Let } \boldsymbol{\psi}^*_{n,N}(\boldsymbol{\theta}) \; = \; \frac{1}{N}\sum_{j=1}^{N} \widehat{\boldsymbol{\phi}}(\mathbf{X}_j, \boldsymbol{\theta}) - \frac{1}{n}\sum_{i=1}^{n} \Big\{ \widehat{\boldsymbol{\phi}}(\mathbf{X}_i, \boldsymbol{\theta}) - \boldsymbol{\psi}(Y_i, \mathbf{X}_i, \boldsymbol{\theta}) \Big\} \quad \forall \; \boldsymbol{\theta} \in \Theta, \text{ and}$$
$$\text{define the EASE estimator } \widehat{\boldsymbol{\theta}}^* \text{ as the solution in } \boldsymbol{\theta} \in \Theta \text{ to: } \; \boldsymbol{\psi}^*_{n,N}(\boldsymbol{\theta}) = \mathbf{0}. \tag{2.3}$$

The construction of the modified EE $\boldsymbol{\psi}^*_{n,N}(\boldsymbol{\theta})$ in (2.3) is fairly intuitive, especially for the first term, where we simply try to mimic the definition of $\boldsymbol{\psi}_0(\boldsymbol{\theta}) = \mathbb{E}_{\mathbf{X}}\{\boldsymbol{\phi}(\mathbf{X}, \boldsymbol{\theta})\}$ by plugging in the estimator $\widehat{\boldsymbol{\phi}}(.)$ of $\boldsymbol{\phi}(.)$ inside, and then (near-perfectly) estimating the outside $\mathbb{E}_{\mathbf{X}}(.)$ through Monte-Carlo based on $\mathcal{U}$. The second term in the definition of $\boldsymbol{\psi}^*_{n,N}(\boldsymbol{\theta})$ is essentially a 'de-biasing' term that turns out to be a natural estimator of the bias of $\widehat{\boldsymbol{\phi}}(.)$ as an estimator of $\boldsymbol{\phi}(.)$. With the construction of $\widehat{\boldsymbol{\phi}}(.)$, to be discussed shortly, allowed to be possibly based on non-parametric smoothing techniques, wherein the convergence rate of $\widehat{\boldsymbol{\phi}}(.)$ is expected to be slower than $O(n^{-\frac{1}{2}})$, this de-biasing term can be quite useful to ensure $n^{\frac{1}{2}}$-consistency of the EASE estimator $\widehat{\boldsymbol{\theta}}^*$ under fairly reasonable conditions. In particular, if $\widehat{\boldsymbol{\phi}}(.)$ is based on kernel smoothing (KS), the de-biasing term plays a crucial role in *avoiding* any *under-smoothing* requirement, so that a bandwidth of the standard 'optimal' order, instead of a smaller order, can be used for the smoothing. The under-smoothing requirement, although

not desirable (as it results in sub-optimal performance of the smoother), is often unavoidable in two-step semi-parametric estimation involving KS in the first step. We refer the interested reader to Newey et al. (1998) for further discussions. Next, we also note that for both $\boldsymbol{\psi}_n(\boldsymbol{\theta})$ and $\boldsymbol{\psi}^*_{n,N}(\boldsymbol{\theta})$, the corresponding solutions $\widehat{\boldsymbol{\theta}}$ and $\widehat{\boldsymbol{\theta}}^*$ do not need to be exact solutions. In fact, they only need to solve: $\boldsymbol{\psi}_n(\boldsymbol{\theta}) = o_p(1)$ and $\boldsymbol{\psi}^*_{n,N}(\boldsymbol{\theta}) = o_p(1)$ respectively (see Van der Vaart (2000) for more details). However, for the sake of simplicity, we prefer to stick to our formulations above, wherein they are the exact solutions to the respective EEs. Finally, we provide some clarifications regarding the construction of the estimator $\widehat{\boldsymbol{\phi}}(\mathbf{x}, \boldsymbol{\theta})$. In general, it can be based on *any* reasonable non-parametric smoothing based approach, including KS for instance. They can also be based on more flexible semi-parametric or semi-non-parametric methods, under more restrictions on the underlying data generating mechanism. For the sake of illustration, we provide *one* particular choice of $\widehat{\boldsymbol{\phi}}(\mathbf{x}, \boldsymbol{\theta})$ based on KS as follows.

$$\widehat{\boldsymbol{\phi}}(\mathbf{x}, \boldsymbol{\theta}) \equiv \widehat{\boldsymbol{\phi}}_{KS}(\mathbf{x}, \boldsymbol{\theta}) = \frac{\frac{1}{nh^p}\sum_{i=1}^n \boldsymbol{\psi}(Y_i, \mathbf{x}, \boldsymbol{\theta})\, K\{(\mathbf{x} - \mathbf{X}_i)/h\}}{\frac{1}{nh^p}\sum_{i=1}^n K\{(\mathbf{x} - \mathbf{X}_i)/h\}} \quad \forall\, \mathbf{x} \in \mathcal{X},\ \boldsymbol{\theta} \in \Theta, \tag{2.4}$$

where $K(.) : \mathbb{R}^p \to \mathbb{R}$ is some appropriate kernel function (e.g. the Gaussian kernel) of order $q \geq 2$ and $h = h(n) > 0$ denotes the bandwidth sequence. For the subclass of problems based on estimating functions that take the form: $\boldsymbol{\psi}(Y, \mathbf{X}, \boldsymbol{\theta}) = \mathbf{h}(\mathbf{X}, \boldsymbol{\theta})\{Y - g(\mathbf{X}, \boldsymbol{\theta})\}$, for some $g(.)$ in $\mathbb{R}$ and $\mathbf{h}(.)$ in $\mathbb{R}^d$, $\widehat{\boldsymbol{\phi}}(\mathbf{x}, \boldsymbol{\theta})$ based on KS takes a particularly simple form as follows:

$$\begin{aligned} &\widehat{\boldsymbol{\phi}}(\mathbf{x}, \boldsymbol{\theta}) = \mathbf{h}(\mathbf{x}, \boldsymbol{\theta})\{\widehat{m}(\mathbf{x}) - g(\mathbf{x}, \boldsymbol{\theta})\} \quad \forall\, \mathbf{x} \in \mathcal{X},\ \boldsymbol{\theta} \in \Theta, \text{ with } \boldsymbol{\psi}(Y, \mathbf{X}, \boldsymbol{\theta}) \text{ as above,} \\ &\text{where } \widehat{m}(\mathbf{x}) \equiv \widehat{m}_{KS}(\mathbf{x}) = \frac{\frac{1}{nh^p}\sum_{i=1}^n Y_i K\{(\mathbf{x} - \mathbf{X}_i)/h\}}{\frac{1}{nh^p}\sum_{i=1}^n K\{(\mathbf{x} - \mathbf{X}_i)/h\}} \quad \forall\, \mathbf{x} \in \mathcal{X} \end{aligned} \tag{2.5}$$

denotes the corresponding KS estimator of the conditional mean $m(\mathbf{x}) \equiv \mathbb{E}(Y \mid \mathbf{X} = \mathbf{x})$. We now formally characterize the theoretical properties of the EASE estimator $\widehat{\boldsymbol{\theta}}^*$ as well as its comparison to those of $\widehat{\boldsymbol{\theta}}$. We first state the necessary assumptions, followed by the results.

Assumption 2.1. *Smoothness of* $\boldsymbol{\psi}_0(.)$. We assume that at least in a neighbourhood $\mathcal{B}(\boldsymbol{\theta}_0; \epsilon)$ of $\boldsymbol{\theta}_0$ for some $\epsilon > 0$, $\boldsymbol{\psi}_0(.)$ is differentiable with a non-singular derivative $\{\mathcal{J}(\boldsymbol{\theta})\}_{d \times d}$, and

further satisfies a Taylor series expansion of the form: $\boldsymbol{\psi}_0(\boldsymbol{\theta}) = \boldsymbol{\psi}_0(\boldsymbol{\theta}_0) + \{\mathcal{J}(\boldsymbol{\theta}_0)\}(\boldsymbol{\theta} - \boldsymbol{\theta}_0) + \mathbf{r}(\boldsymbol{\theta}, \boldsymbol{\theta}_0)\ \forall\ \boldsymbol{\theta} \in \mathcal{B}(\boldsymbol{\theta}_0; \epsilon)$, for some $\mathbf{r}(\boldsymbol{\theta}, \boldsymbol{\theta}_0)$ satisfying $\|\mathbf{r}(\boldsymbol{\theta}, \boldsymbol{\theta}_0)\| \leq O(\|\boldsymbol{\theta} - \boldsymbol{\theta}_0\|^2)$ as $\boldsymbol{\theta} \to \boldsymbol{\theta}_0$.

Assumption 2.2. *Complexity of the class of estimating functions.* We assume that the class of functions: $\mathcal{G}_0 \equiv \{\boldsymbol{\psi}(Y, \mathbf{X}, \boldsymbol{\theta}) : \boldsymbol{\theta} \in \Theta\}$ lies in a $\mathbb{P}$-Glivenko-Cantelli ($\mathbb{P}$-GC) class, and for some $\epsilon > 0$, the class $\mathcal{D}_{0,\epsilon} \equiv \{\boldsymbol{\psi}(Y, \mathbf{X}, \boldsymbol{\theta}) : \boldsymbol{\theta} \in \mathcal{B}(\boldsymbol{\theta}_0; \epsilon) \subseteq \Theta\}$ lies in a $\mathbb{P}$-Donsker class. More generally, we actually assume that the following conditions hold:

$$\sup_{\boldsymbol{\theta} \in \Theta} \|\mathbb{P}_n\{\boldsymbol{\psi}(Y, \mathbf{X}, \boldsymbol{\theta})\} - \boldsymbol{\psi}_0(\boldsymbol{\theta})\| \xrightarrow{P} 0, \quad \text{and} \tag{2.6}$$

$$\mathbb{G}_n\{\boldsymbol{\psi}(Y, \mathbf{X}, \widetilde{\boldsymbol{\theta}}) - \boldsymbol{\psi}(Y, \mathbf{X}, \boldsymbol{\theta}_0)\} \xrightarrow{P} \mathbf{0}, \ \text{ for any (random) sequence } \widetilde{\boldsymbol{\theta}} \xrightarrow{P} \boldsymbol{\theta}_0. \tag{2.7}$$

Assumption 2.3. *Behavior of $\widehat{\boldsymbol{\phi}}(.)$ as an estimator of $\boldsymbol{\phi}(.)$.* We assume first of all that:

$$\sup_{\mathbf{x} \in \mathcal{X},\ \boldsymbol{\theta} \in \Theta} \left\| \widehat{\boldsymbol{\phi}}(\mathbf{x}, \boldsymbol{\theta}) - \boldsymbol{\phi}(\mathbf{x}, \boldsymbol{\theta}) \right\| \xrightarrow{P} 0. \tag{2.8}$$

Further, we assume that for some $\epsilon, \delta > 0$, the (random) sequence of function classes given by: $\mathcal{D}_{n,\epsilon,\delta} \equiv \{\widehat{\boldsymbol{\phi}}(\mathbf{X}, \boldsymbol{\theta}) : \boldsymbol{\theta} \in \mathcal{B}(\boldsymbol{\theta}_0; \epsilon),\ \sup_{\mathbf{x} \in \mathcal{X}} \|\widehat{\boldsymbol{\phi}}(\mathbf{x}, \boldsymbol{\theta}) - \boldsymbol{\phi}(\mathbf{x}, \boldsymbol{\theta})\| \leq \delta\ \ \forall\ \boldsymbol{\theta} \in \mathcal{B}(\boldsymbol{\theta}_0; \epsilon)\}$ lies in a $\mathbb{P}$-Donsker class with probability converging to 1. More generally, we assume that:

$$\sup_{\boldsymbol{\theta} \in \mathcal{B}(\boldsymbol{\theta}_0; \epsilon)} \left\| \frac{1}{N} \sum_{j=1}^{N} \widehat{\boldsymbol{\phi}}(\mathbf{X}_j, \boldsymbol{\theta}) - \mathbb{E}_{\mathbf{X}} \left\{ \widehat{\boldsymbol{\phi}}(\mathbf{X}, \boldsymbol{\theta}) \right\} \right\| = O_p\left(N^{-\frac{1}{2}}\right) \ \text{ and} \tag{2.9}$$

$$\mathbb{G}_n \left\{ \widehat{\boldsymbol{\phi}}(\mathbf{X}, \widetilde{\boldsymbol{\theta}}) - \boldsymbol{\phi}(\mathbf{X}, \boldsymbol{\theta}_0) \right\} \xrightarrow{P} \mathbf{0}, \ \text{ for any (random) sequence } \widetilde{\boldsymbol{\theta}} \xrightarrow{P} \boldsymbol{\theta}_0. \tag{2.10}$$

Theorem 2.1. *Under assumptions 2.1-2.3, and letting $\mathbf{G}^*_{n,1} = \mathbb{G}_n\{\boldsymbol{\psi}(Y, \mathbf{X}, \widehat{\boldsymbol{\theta}}^*) - \boldsymbol{\psi}(Y, \mathbf{X}, \boldsymbol{\theta}_0)\}$ and $\mathbf{G}^*_{n,2} = \mathbb{G}_n\{\widehat{\boldsymbol{\phi}}(\mathbf{X}, \widehat{\boldsymbol{\theta}}^*) - \boldsymbol{\phi}(\mathbf{X}, \boldsymbol{\theta}_0)\}$, the EASE estimator $\widehat{\boldsymbol{\theta}}^*$ satisfies the expansion:*

$$n^{\frac{1}{2}} \left(\widehat{\boldsymbol{\theta}}^* - \boldsymbol{\theta}_0 \right) = n^{-\frac{1}{2}} \sum_{i=1}^{n} \boldsymbol{\xi}_{\mathit{eff}}(\mathbf{Z}_i) + O_p\left(\mathbf{G}^*_{n,1} + \mathbf{G}^*_{n,2}\right) + O_p\left(\frac{n}{N}\right)^{\frac{1}{2}} + o_p(1), \tag{2.11}$$

$$\textit{where } \boldsymbol{\xi}_{\mathit{eff}}(\mathbf{Z}) = -\{\mathcal{J}(\boldsymbol{\theta}_0)\}^{-1} \boldsymbol{\Psi}_{\mathit{eff}}(\mathbf{Z}) \equiv -\{\mathcal{J}(\boldsymbol{\theta}_0)\}^{-1} \{\boldsymbol{\psi}(Y, \mathbf{X}, \boldsymbol{\theta}_0) - \boldsymbol{\phi}(\mathbf{X}, \boldsymbol{\theta}_0)\}.$$

Hence, $n^{\frac{1}{2}}(\widehat{\boldsymbol{\theta}}^{*}-\boldsymbol{\theta}_0) \xrightarrow{d} \mathcal{N}_d[\mathbf{0}, \boldsymbol{\Sigma}\{\boldsymbol{\xi}_{\mathit{eff}}(.)\}]$*, and* $\widehat{\boldsymbol{\theta}}^{*}$ *is a regular and asymptotically linear (RAL) estimator of* $\boldsymbol{\theta}_0$ *with influence function (IF) given by:* $\boldsymbol{\xi}_{\mathit{eff}}(\mathbf{Z})$*. Further, under assumptions 2.1-2.2, and letting* $\overline{\mathbf{G}}_n = \mathbb{G}_n\{\boldsymbol{\psi}(Y, \mathbf{X}, \widehat{\boldsymbol{\theta}}) - \boldsymbol{\psi}(Y, \mathbf{X}, \boldsymbol{\theta}_0)\}$*, the supervised M-estimator* $\widehat{\boldsymbol{\theta}}$ *satisfies:*

$$
\begin{aligned}
n^{\frac{1}{2}}\left(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0\right) \;&=\; n^{-\frac{1}{2}} \sum_{i=1}^{n} \boldsymbol{\xi}_0(\mathbf{Z}_i) + O_p\left(\overline{\mathbf{G}}_n\right) + + o_p(1), \qquad (2.12)\\
\textit{where } \boldsymbol{\xi}_0(\mathbf{Z}) \;&=\; -\{\mathcal{J}(\boldsymbol{\theta}_0)\}^{-1}\boldsymbol{\Psi}_0(\mathbf{Z}) \;\equiv\; -\{\mathcal{J}(\boldsymbol{\theta}_0)\}^{-1}\{\boldsymbol{\psi}(Y, \mathbf{X}, \boldsymbol{\theta}_0)\}.
\end{aligned}
$$

Hence, $n^{\frac{1}{2}}(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) \xrightarrow{d} \mathcal{N}_d[\mathbf{0}, \boldsymbol{\Sigma}\{\boldsymbol{\xi}_0(.)\}]$*, and* $\widehat{\boldsymbol{\theta}}$ *is a RAL estimator of* $\boldsymbol{\theta}_0$ *with IF:* $\boldsymbol{\xi}_0(\mathbf{Z})$*.*

Moreover, when $\boldsymbol{\psi}(Y, \mathbf{X}, \boldsymbol{\theta})$ *takes the particular form:* $\boldsymbol{\psi}(Y, \mathbf{X}, \boldsymbol{\theta}) = \mathbf{h}(\mathbf{X}, \boldsymbol{\theta})\{Y - g(\mathbf{X}, \boldsymbol{\theta})\}$ *for some* $\mathbf{h}(.)$ *in* $\mathbb{R}^d$ *and some* $g(.)$ *in* $\mathbb{R}$*, and* $\widehat{\boldsymbol{\phi}}(\mathbf{X}, \boldsymbol{\theta}) = \mathbf{h}(\mathbf{X}, \boldsymbol{\theta})\{\widehat{m}(\mathbf{X}) - g(\mathbf{X}, \boldsymbol{\theta}\}$ *for any estimator* $\widehat{m}(.)$ *of* $m(.)$*, the results in (2.11)-(2.12) continue to hold under assumptions 2.1-2.2, and some slightly simpler conditions, instead of assumption 2.3, as follows.*

$$
\begin{aligned}
&\sup_{\mathbf{x}\in\mathcal{X}, \boldsymbol{\theta}\in\Theta} \|\mathbf{h}(\mathbf{x}, \boldsymbol{\theta})\| < \infty, \quad \sup_{\mathbf{x}\in\mathcal{X}, \boldsymbol{\theta}\in\Theta} |g(\mathbf{x}, \boldsymbol{\theta})| < \infty, \quad \textit{and for some } \epsilon, \widetilde{\epsilon} > 0,\\
&\|\mathbf{h}(\mathbf{x}, \boldsymbol{\theta}) - \mathbf{h}(\mathbf{x}, \boldsymbol{\theta}_0)\| \;\leq\; \overline{h}(\mathbf{x})\|\boldsymbol{\theta} - \boldsymbol{\theta}_0\| \;\; \forall\, \boldsymbol{\theta} \in \mathcal{B}(\boldsymbol{\theta}_0; \epsilon) \;\; \textit{for some } \overline{h}(.) \in \mathcal{L}_2(\mathbb{P}_{\mathbf{X}}),\\
&|g(\mathbf{x}, \boldsymbol{\theta}) - g(\mathbf{x}, \boldsymbol{\theta}_0)| \;\leq\; \overline{g}(\mathbf{x})\|\boldsymbol{\theta} - \boldsymbol{\theta}_0\| \;\; \forall\, \boldsymbol{\theta} \in \mathcal{B}(\boldsymbol{\theta}_0; \widetilde{\epsilon}) \;\; \textit{for some } \overline{g}(.) \in \mathcal{L}_2(\mathbb{P}_{\mathbf{X}}), \;\; \textit{and}\\
&\sup_{\mathbf{x}\in\mathcal{X}} |\widehat{m}(\mathbf{x}) - m(\mathbf{x})| \;\xrightarrow{P}\; 0, \;\; \textit{and} \;\; \mathbb{G}_n[\mathbf{h}(\mathbf{X}, \boldsymbol{\theta}_0)\{\widehat{m}(\mathbf{X}) - m(\mathbf{X})\}] \;\xrightarrow{P}\; \mathbf{0}. \qquad (2.13)
\end{aligned}
$$

Apart from establishing $n^{\frac{1}{2}}$-consistency and asymptotic normality (CAN) as well as RAL properties for $\widehat{\boldsymbol{\theta}}^{*}$ and $\widehat{\boldsymbol{\theta}}$, under every model in $\mathcal{M}_{\mathbf{X}}$, theorem 2.1 also establishes the efficient and adaptive nature of $\widehat{\boldsymbol{\theta}}^{*}$, as desired. In particular, the asymptotic variances $\boldsymbol{\Sigma}\{\boldsymbol{\xi}_{\text{eff}}(.)\}$ and $\boldsymbol{\Sigma}\{\boldsymbol{\xi}_0(.)\}$ of $\widehat{\boldsymbol{\theta}}^{*}$ and $\widehat{\boldsymbol{\theta}}$ respectively satisfy: $\boldsymbol{\Sigma}\{\boldsymbol{\xi}_{\text{eff}}(.)\} \preceq \boldsymbol{\Sigma}\{\boldsymbol{\xi}_0(.)\}$ for every model in $\mathcal{M}_{\mathbf{X}}$, and further, the inequality is strict whenever $\mathbb{P}_{\mathbf{X}}$ is informative for estimating $\boldsymbol{\theta}_0$ under $\mathcal{M}_{\mathbf{X}}$, in the sense of definition 2.1. Moreover, it is not difficult to show that the IF: $\boldsymbol{\xi}_{\text{eff}}(.)$, achieved by $\widehat{\boldsymbol{\theta}}^{*}$, is also the so-called 'efficient' IF for estimating $\boldsymbol{\theta}_0$ under $\mathcal{M}_{\mathbf{X}}$, so that $\widehat{\boldsymbol{\theta}}^{*}$ *achieves* the semi-parametric efficiency bound globally under $\mathcal{M}_{\mathbf{X}}$ (i.e. for every model in

$\mathcal{M}_{\mathbf{X}}$), and is therefore asymptotically optimal among all RAL estimators of $\boldsymbol{\theta}_0$ under $\mathcal{M}_{\mathbf{X}}$ (i.e. for any RAL estimator of $\boldsymbol{\theta}_0$ with IF: $\boldsymbol{\xi}(\mathbf{Z})$, $\text{Var}\{\boldsymbol{\xi}_{\text{eff}}(\mathbf{Z})\} \preceq \text{Var}\{\boldsymbol{\xi}(\mathbf{Z})\}$ under every model in $\mathcal{M}_{\mathbf{X}}$).

Regarding the assumptions required for theorem 2.1, assumptions 2.1-2.2 are both fairly mild and standard in the M-estimation literature, and similar (or equivalent) conditions can be found in Van der Vaart (2000) for instance. Assumption 2.3 essentially imposes some desirable convergence properties for $\widehat{\boldsymbol{\phi}}(.)$ as an estimator of $\boldsymbol{\phi}(.)$, and should be expected to hold for most reasonable (and smooth enough) choices of $\widehat{\boldsymbol{\phi}}(.)$ and $\boldsymbol{\phi}(.)$. Note also that the conditions in assumptions 2.2-2.3 involving $\mathbb{P}$-GC or $\mathbb{P}$-Donsker classes from empirical process theory are mostly to ensure the validity of (2.6)-(2.10), which are really all that is needed for theorem 2.1 to hold. Hence, more general and/or different sufficient conditions can also be used as long as (2.6)-(2.10) are satisfied. However, GC and Donsker classes are typically known to be some of the most general and mild enough requirements ensuring these type of conditions, and they include a wide variety of function classes as examples. We refer the interested reader to Van der Vaart (2000) and Van Der Vaart and Wellner (1996) for more discussions regarding the various properties and examples of GC and Donsker classes.

Lastly, for the slightly simpler subclass of EEs introduced in the final part of theorem 2.1, the conditions required are still quite mild, and yet relatively easier to verify. In particular, for smoothing based on KS, the last condition in (2.13) has been verified in details, under fairly mild regularity conditions and without using empirical process theory, in Theorem 4.1 and Lemma A.2 of Chakrabortty and Cai (2015), where it has been shown that the required convergence results can be achieved without any under-smoothing requirement, so that a bandwidth of the 'optimal' order: $O(n^{-1/(2q+p)})$ can be used as long as the kernel order q satisfies: $q > p/2$. These kind of requirements are also known to be encountered elsewhere in other forms, especially in empirical process theory where it is well known (Van der Vaart, 2000) that standard classes of smooth functions are Donsker only if they have a smoothness of order greater than half the ambient dimension. For a simpler subclass of problems, including

those based on generalized linear (working) models (GLwMs) with canonical link functions (e.g. linear, logistic or poisson regression), this requirement can be avoided using a sample-splitting or cross-validation (CV) based technique that was demonstrated in Chakrabortty and Cai (2015) for the special case of linear regression. However, given the generality of the framework considered herein, we prefer not to delve any further into this aspect, as the analysis for such approaches can be somewhat involved and is beyond the scope of this paper.

2.3.1 'Separable' Estimating Equations: Flexible SS Estimators

In this section, we consider a subclass of M-estimation problems, based on what we call 'separable' EEs, where the underlying estimating function $\boldsymbol{\psi}(Y, \mathbf{X}, \boldsymbol{\theta})$ takes the simple form:

$$\boldsymbol{\psi}(Y, \mathbf{X}, \boldsymbol{\theta}) = \mathbf{h}(\mathbf{X})\{Y - g(\mathbf{X}, \boldsymbol{\theta})\} \quad \in \mathbb{R}^d, \quad \forall\, \boldsymbol{\theta} \in \Theta \subseteq \mathbb{R}^d. \tag{2.14}$$

In particular, with $d = (p+1)$ and $\mathbf{h}(\mathbf{X}) = \overrightarrow{\mathbf{X}}$, these EEs include, as special cases, the ones encountered in estimation problems for standard GLwMs based on the so-called 'canonical' link functions (e.g. linear, logistic or poisson regression), with $g(\mathbf{X}, \boldsymbol{\theta}) = g_0(\overrightarrow{\mathbf{X}}'\boldsymbol{\theta})$, where $g_0(.)$ denotes the appropriate link function. We aim to provide more flexible SS estimation strategies for these subclass of problems, wherein, unlike the previous approach, it won't be necessary to use a non-parametric estimator of the conditional mean $m(.)$ which, if we are using KS for instance, could be quite undesirable in practice if p is even moderately large, owing to slow convergence rates due to the curse of dimensionality, and moreover, substantial finite sample over-fitting bias. We first make a fundamental definition in this regard.

Definition 2.2. Let $\mu(\mathbf{X}) \in \mathbb{R}$ be any measurable function of $\mathbf{X}$ with $\mu(.) \in \mathcal{L}_2(\mathbb{P}_{\mathbf{X}})$. With the 'separable' estimating function $\boldsymbol{\psi}(.)$ as in (2.14), and the parameter of interest being still $\boldsymbol{\theta}_0$ as defined generally in (2.1), we then define $\mu(.)$ to be an *admissible imputation function* (AIF) w.r.t. $\boldsymbol{\psi}(.)$ for estimating $\boldsymbol{\theta}_0$ under $\mathcal{M}_{\mathbf{X}}$, if $\mathbb{E}_{\mathbf{X}}[\boldsymbol{\psi}\{\mu(\mathbf{X}), \mathbf{X}, \boldsymbol{\theta}_0\}] = \mathbf{0}$, so that

$$\mathbb{E}_{\mathbf{X}}[\mathbf{h}(\mathbf{X})\{\mu(\mathbf{X}) - g(\mathbf{X}, \boldsymbol{\theta}_0)\}] = \mathbf{0}, \text{ and hence, } \mathbb{E}[\mathbf{h}(\mathbf{X})\{Y - \mu(\mathbf{X})\}] = \mathbf{0}. \tag{2.15}$$

The name 'imputation function' for $\mu(.)$ is purely inspired from the fact that we are aiming to replace the Y in the definition of $\boldsymbol{\psi}(.)$ by $\mu(\mathbf{X})$, and still trying to ensure that $\boldsymbol{\theta}_0$ satisfies the EE determined by this 'imputed' version of the underlying estimating function. Of course, one choice of $\mu(.)$ is the conditional mean $m(.)$ itself, but we would be interested in more diverse choices of such $\mu(.)$, and their estimators, that would lead to a corresponding family of flexible SS estimators, indexed by the respective choice of $\mu(.)$. For the case of SS logistic regression, a particular construction of (a family of) $\mu(.)$, as well as its estimator would also be provided, involving effective use of dimension reduction techniques (if desired), which can be quite helpful, especially if smoothing based methods like KS are involved in the estimation, that do not scale well with the underlying smoothing dimension. A corresponding version of such constructions were also provided in Chakrabortty and Cai (2015) for the linear regression problem, which also included, as a special case, the optimal estimator (for that problem) that we have developed here in section 2.3 under a much more general framework. We next present the general construction and properties of a family of SS estimators of $\boldsymbol{\theta}_0$, under the setting introduced above, indexed by the choice of the AIF $\mu(.)$.

A Flexible Family of SS Estimators for 'Separable' Estimating Equations: With $\mu(.)$ as defined in 2.2, let $\widehat{\mu}(.)$ denote *any* reasonable estimator, based on $\mathcal{L}$, of $\mu(.)$, and define $\widehat{\boldsymbol{\phi}}_{\mu}(\mathbf{x}, \boldsymbol{\theta}) = \mathbf{h}(\mathbf{x})\{\widehat{\mu}(\mathbf{x}) - g(\mathbf{x}, \boldsymbol{\theta})\}\ \forall\ \mathbf{x} \in \mathcal{X}, \boldsymbol{\theta} \in \Theta$. Using $\widehat{\mu}(.)$, we now define a μ-*modified* EE, $\boldsymbol{\psi}^{(\mu)}_{n,N}(\boldsymbol{\theta})$, and then a corresponding SS estimator $\widehat{\boldsymbol{\theta}}_{\mu}$ as the solution to that EE, as follows.

$$\text{Let } \boldsymbol{\psi}^{(\mu)}_{n,N}(\boldsymbol{\theta}) \ = \ \frac{1}{N}\sum_{j=1}^{N} \widehat{\boldsymbol{\phi}}_{\mu}(\mathbf{X}_j, \boldsymbol{\theta}) - \frac{1}{n}\sum_{i=1}^{n}\Big\{\widehat{\boldsymbol{\phi}}_{\mu}(\mathbf{X}_i, \boldsymbol{\theta}) - \boldsymbol{\psi}(Y_i, \mathbf{X}_i, \boldsymbol{\theta})\Big\} \quad \forall\ \boldsymbol{\theta} \in \Theta,$$

$$= \frac{1}{N}\sum_{j=1}^{N} \mathbf{h}(\mathbf{X}_j)\{\widehat{\mu}(\mathbf{X}_j) - g(\mathbf{X}_j, \boldsymbol{\theta})\} - \frac{1}{n}\sum_{i=1}^{n} \mathbf{h}(\mathbf{X}_i)\{\widehat{\mu}(\mathbf{X}_i) - Y_i\}, \ \text{ and} \tag{2.16}$$

$$\text{define the SS estimator } \widehat{\boldsymbol{\theta}}_{\mu} \text{ as the solution in } \boldsymbol{\theta} \in \Theta \text{ to: } \boldsymbol{\psi}^{(\mu)}_{n,N}(\boldsymbol{\theta}) = \mathbf{0}. \tag{2.17}$$

The construction of the μ-modified EE $\boldsymbol{\psi}^{(\mu)}_{n,N}(\boldsymbol{\theta})$ in (2.17) is more or less based on the same ideas and intuitions underlying the construction of the modified EE in (2.3) for the general EASE. The first term tries to mimic the definition of $\boldsymbol{\psi}_0(\boldsymbol{\theta}) \equiv \mathbb{E}[\mathbf{h}(\mathbf{X})\{Y - g(\mathbf{X}, \boldsymbol{\theta})\}]$ which equals $\mathbb{E}_{\mathbf{X}}[\mathbf{h}(\mathbf{X})\{\boldsymbol{\mu}(\mathbf{X}) - g(\mathbf{X}, \boldsymbol{\theta}) \; \forall \; \boldsymbol{\theta} \in \Theta$, owing to (2.15), wherein we simply plug in the estimator $\widehat{\mu}(.)$ of $\mu(.)$ inside, and then (near-perfectly) estimate the outside $\mathbb{E}_{\mathbf{X}}(.)$ through Monte-Carlo based on $\mathcal{U}$. The second term, similar to (2.3), corresponds to a 'de-biasing' term that turns out to be a natural estimator of the bias of $\widehat{\mu}(.)$ as an estimator of $\mu(.)$, and in particular, helps avoiding any under-smoothing requirement if KS is used for constructing $\widehat{\mu}(.)$. We next characterize the theoretical properties of $\widehat{\boldsymbol{\theta}}_\mu$ through the following result.

Theorem 2.2. *Let $\boldsymbol{\psi}(.)$, $\boldsymbol{\theta}_0$, $\mu(.)$, $\widehat{\mu}(.)$ and $\widehat{\boldsymbol{\theta}}_\mu$ be as introduced through (2.14)-(2.17), and suppose assumptions 2.1-2.2 hold. Assume further that for some $\epsilon > 0$, the function class: $\mathcal{F}_\epsilon = \{\mathbf{h}(\mathbf{X})g(\mathbf{X}, \boldsymbol{\theta}) : \boldsymbol{\theta} \in \mathcal{B}(\boldsymbol{\theta}_0; \epsilon)\}$ lies in a $\mathbb{P}$-Donsker class, or more generally, assume that*

$$\sup_{\boldsymbol{\theta} \in \mathcal{B}(\boldsymbol{\theta}_0, \epsilon)} \left\| \frac{1}{N} \sum_{j=1}^{N} \mathbf{h}(\mathbf{X}_j) g(\mathbf{X}_j, \boldsymbol{\theta}) - \mathbb{E}_{\mathbf{X}}\{\mathbf{h}(\mathbf{X}) g(\mathbf{X}, \boldsymbol{\theta})\} \right\| = O_p\left(N^{-\frac{1}{2}}\right), \quad \text{for some } \epsilon > 0. \tag{2.18}$$

Further, assume the following convergence properties of $\widehat{\mu}(.)$ as an estimator of $\mu(.)$:

$$\sup_{\mathbf{x} \in \mathcal{X}} \|\mathbf{h}(\mathbf{x})\{\widehat{\mu}(\mathbf{x}) - \mu(\mathbf{x})\}\| \xrightarrow{P} 0, \quad \text{and} \quad \mathbf{G}^{(\mu)}_n \equiv \mathbb{G}_n\left[\mathbf{h}(\mathbf{X})\left\{\widehat{\mu}(\mathbf{X}) - \mu(\mathbf{X})\right\}\right] \xrightarrow{P} \mathbf{0}. \tag{2.19}$$

Then, the SS estimator $\widehat{\boldsymbol{\theta}}_\mu$ in (2.17), satisfies the following expansion:

$$n^{\frac{1}{2}}\left(\widehat{\boldsymbol{\theta}}_\mu - \boldsymbol{\theta}_0\right) = n^{-\frac{1}{2}} \sum_{i=1}^{n} \boldsymbol{\xi}_\mu(\mathbf{Z}_i) + O_p\left(\mathbf{G}^{(\mu)}_n\right) + O_p\left(\frac{n}{N}\right)^{\frac{1}{2}} + o_p(1), \tag{2.20}$$

$$\text{where } \boldsymbol{\xi}_\mu(\mathbf{Z}) = -\{\mathcal{J}(\boldsymbol{\theta}_0)\}^{-1}\boldsymbol{\Psi}_\mu(\mathbf{Z}) \equiv -\{\mathcal{J}(\boldsymbol{\theta}_0)\}^{-1}\left[\mathbf{h}(\mathbf{X})\left\{Y - \mu(\mathbf{X})\right\}\right]. \tag{2.21}$$

Hence, $n^{\frac{1}{2}}(\widehat{\boldsymbol{\theta}}_\mu - \boldsymbol{\theta}_0) \xrightarrow{d} \mathcal{N}_d[\mathbf{0}, \boldsymbol{\Sigma}\{\boldsymbol{\xi}_\mu(.)\}]$, and $\widehat{\boldsymbol{\theta}}_\mu$ is a RAL (and CAN) estimator of $\boldsymbol{\theta}_0$ with IF: $\boldsymbol{\xi}_\mu(\mathbf{Z})$, while the supervised M-estimator $\widehat{\boldsymbol{\theta}}$, using theorem 2.1, is a RAL (and CAN) estimator of $\boldsymbol{\theta}_0$ with IF: $\boldsymbol{\xi}_0(\mathbf{Z}) = -\{\mathcal{J}(\boldsymbol{\theta}_0)\}^{-1}\boldsymbol{\Psi}_0(\mathbf{Z}) \equiv -\{\mathcal{J}(\boldsymbol{\theta}_0)\}^{-1}[\mathbf{h}(\mathbf{X})\{Y - g(\mathbf{X}, \boldsymbol{\theta}_0)\}]$.

Theorem 2.2 therefore equips us a with a family of RAL and CAN estimators $\widehat{\boldsymbol{\theta}}_\mu$ of $\boldsymbol{\theta}_0$, indexed by the choice of the AIF $\mu(.)$. If somehow the AIF is 'sufficient' in the sense that $\mu(.) = m(.)$, then $\boldsymbol{\xi}_\mu(.) = \boldsymbol{\xi}_{\text{eff}}(.)$, so that $\widehat{\boldsymbol{\theta}}_\mu$ and the general EASE estimator $\widehat{\boldsymbol{\theta}}^*$ obtained earlier are indeed asymptotically equivalent, and therefore $\widehat{\boldsymbol{\theta}}_\mu$ enjoys the same set of optimality properties as those discussed for $\widehat{\boldsymbol{\theta}}^*$ at the end of theorem 2.1. However, if the AIF is not sufficient i.e. $\mu(.) \neq m(.)$, then the efficient and adaptive property (or any of the other optimality properties) of $\widehat{\boldsymbol{\theta}}_\mu$ w.r.t. $\widehat{\boldsymbol{\theta}}$ is no longer guaranteed. We address this next.

Construction of Flexible EASE Estimators Based on $\widehat{\boldsymbol{\theta}}_\mu$: To ensure adaptivity even when $\mu(.) \neq m(.)$, we now define the final EASE estimator, based on $\widehat{\boldsymbol{\theta}}_\mu$, as an optimal linear combination of $\widehat{\boldsymbol{\theta}}$ and $\widehat{\boldsymbol{\theta}}_\mu$. Specifically, for any fixed $d \times d$ matrix $\boldsymbol{\Delta}$, $\widehat{\boldsymbol{\theta}}_\mu(\boldsymbol{\Delta}) = \widehat{\boldsymbol{\theta}} + \boldsymbol{\Delta}(\widehat{\boldsymbol{\theta}}_\mu - \widehat{\boldsymbol{\theta}})$ is a CAN and RAL estimator of $\boldsymbol{\theta}_0$ whenever $\widehat{\boldsymbol{\theta}}$ and $\widehat{\boldsymbol{\theta}}_\mu$ are, and an optimal $\boldsymbol{\Delta}$ can be selected easily to minimize the asymptotic variance of the combined estimator. For simplicity, we focus here on $\boldsymbol{\Delta}$ being a diagonal matrix with $\boldsymbol{\Delta} = \text{diag}(\delta_1, ..., \delta_d)$. Then, the EASE based on $\widehat{\boldsymbol{\theta}}_\mu$ is defined as $\widehat{\boldsymbol{\theta}}^E_\mu \equiv \widehat{\boldsymbol{\theta}}_\mu(\widehat{\boldsymbol{\Delta}})$ with $\widehat{\boldsymbol{\Delta}}$ being any consistent estimator (see next section for further details) of the minimizer $\overline{\boldsymbol{\Delta}} = \text{diag}(\overline{\delta}_1, ..., \overline{\delta}_d)$, where $\forall\ 1 \leq l \leq d$,

$$\overline{\delta}_l = -\lim_{\epsilon \downarrow 0} \frac{\text{Cov}\left\{\boldsymbol{\xi}_{0[l]}(\mathbf{Z}),\ \boldsymbol{\xi}_{\mu[l]}(\mathbf{Z}) - \boldsymbol{\xi}_{0[l]}(\mathbf{Z})\right\}}{\text{Var}\left\{\boldsymbol{\xi}_{\mu[l]}(\mathbf{Z}) - \boldsymbol{\xi}_{0[l]}(\mathbf{Z})\right\}\ +\ \epsilon}. \tag{2.22}$$

Note that in (2.22), the ϵ and the limit outside are included to formally account for the case: $\boldsymbol{\xi}_{0[l]}(\mathbf{Z}) = \boldsymbol{\xi}_{\mu[l]}(\mathbf{Z})$ a.s. $[\mathbb{P}]$, when we define $\overline{\delta}_l = 0$ for identifiability.

It is straightforward to show that $\widehat{\boldsymbol{\theta}}^E_\mu$ and $\widehat{\boldsymbol{\theta}}_\mu(\overline{\boldsymbol{\Delta}})$ are asymptotically equivalent, so that $\widehat{\boldsymbol{\theta}}^E_\mu$ is a RAL estimator of $\boldsymbol{\theta}_0$ satisfying:

$$n^{\frac{1}{2}}\left(\widehat{\boldsymbol{\theta}}^E_\mu - \boldsymbol{\theta}_0\right) = n^{-\frac{1}{2}}\sum_{i=1}^{n} \boldsymbol{\xi}_\mu(\mathbf{Z}_i, \overline{\boldsymbol{\Delta}}) + o_p(1) \ \xrightarrow{d}\ \mathcal{N}_d[\mathbf{0}, \boldsymbol{\Sigma}_\mu(\overline{\boldsymbol{\Delta}})] \ \text{ as } n \to \infty, \tag{2.23}$$
$$\text{where } \boldsymbol{\xi}_\mu(\mathbf{Z}, \overline{\boldsymbol{\Delta}}) = \boldsymbol{\xi}_0(\mathbf{Z}) + \overline{\boldsymbol{\Delta}}\{\boldsymbol{\xi}_\mu(\mathbf{Z}) - \boldsymbol{\xi}_0(\mathbf{Z})\}, \ \text{ and } \ \boldsymbol{\Sigma}_\mu(\overline{\boldsymbol{\Delta}}) = \text{Var}\{\boldsymbol{\xi}_\mu(\mathbf{Z}, \overline{\boldsymbol{\Delta}})\}.$$

Note that whenever the AIF is sufficient so that $\mu(.) = m(.)$, then $\boldsymbol{\xi}_\mu(\mathbf{Z}, \overline{\boldsymbol{\Delta}}) = \boldsymbol{\xi}_{\text{eff}}(\mathbf{Z})$, so that

$\widehat{\boldsymbol{\theta}}_{\mu}^{E}$ is asymptotically equivalent (and therefore optimal) to $\widehat{\boldsymbol{\theta}}^{*}$. Further, when $\mu(.) \neq m(.)$, then $\widehat{\boldsymbol{\theta}}_{\mu}^{E}$ is no longer optimal, but is *still* efficient and adaptive compared to $\widehat{\boldsymbol{\theta}}$ (as well as to $\widehat{\boldsymbol{\theta}}_{\mu}$). Lastly, if the AIF is certain to be sufficient, we may simply define $\widehat{\boldsymbol{\theta}}_{\mu}^{E} = \widehat{\boldsymbol{\theta}}_{\mu}$.

Inference Based on $\widehat{\boldsymbol{\theta}}_{\mu}$ and the EASE Estimator $\widehat{\boldsymbol{\theta}}_{\mu}^{E}$: We now provide procedures for making inference about $\boldsymbol{\theta}_0$ based on $\widehat{\boldsymbol{\theta}}_{\mu}$ and $\widehat{\boldsymbol{\theta}}_{\mu}^{E}$, through estimation of their asymptotic variances, as well obtaining the estimate $\widehat{\boldsymbol{\Delta}}$ of $\overline{\boldsymbol{\Delta}}$ involved in the definition of $\widehat{\boldsymbol{\theta}}_{\mu}^{E}$ that is required for its implementation in practice. In order to avoid potential over-fitting bias in the inference procedures, since $\widehat{\mu}(.)$ can be possibly obtained using non-parametric smoothing based techniques, we further adopt a $\mathbb{K}$-fold cross-validation (CV) based approach, for any fixed $\mathbb{K} \geq 2$, for constructing the inference procedures. We first introduce a few notations for this purpose. For any fixed $\mathbb{K} \geq 2$, let $\{\mathcal{L}_k\}_{k=1}^{\mathbb{K}}$ denote a random partition of $\mathcal{L}$ into $\mathbb{K}$ disjoint subsets of equal sizes, $n_{\mathbb{K}} = n/\mathbb{K}$, with index sets $\{\mathcal{I}_k\}_{k=1}^{\mathbb{K}}$. Let $\mathcal{L}_k^{-}$ denote the set excluding $\mathcal{L}_k$ with size $n_{\mathbb{K}}^{-} = n - n_{\mathbb{K}}$ and respective index set $\mathcal{I}_k^{-}$. Let $\{\widehat{\mu}_k(.)\}_{k=1}^{\mathbb{K}}$ denote the corresponding estimators of $\mu(.)$ based on $\{\mathcal{L}_k^{-}\}_{k=1}^{\mathbb{K}}$. Further, for the supervised estimator $\widehat{\boldsymbol{\theta}}$, let $\{\widehat{\boldsymbol{\theta}}_k\}_{k=1}^{\mathbb{K}}$ denote its corresponding versions obtained from $\{\mathcal{L}_k^{-}\}_{k=1}^{\mathbb{K}}$. A key step involved in the construction of the inference procedures, as well as in obtaining $\widehat{\boldsymbol{\Delta}}$, is to obtain reasonable (non-over-fitted) estimates of the IFs of $\widehat{\boldsymbol{\theta}}$ and $\widehat{\boldsymbol{\theta}}_{\mu}$: $\boldsymbol{\xi}_0(\mathbf{Z}) \equiv -\{\mathcal{J}(\boldsymbol{\theta}_0)\}^{-1}\boldsymbol{\Psi}_0(\mathbf{Z})$ and $\boldsymbol{\xi}_{\mu}(\mathbf{Z}) \equiv -\{\mathcal{J}(\boldsymbol{\theta}_0)\}^{-1}\boldsymbol{\Psi}_{\mu}(\mathbf{Z})$ respectively, for $\mathbf{Z} \in \{\mathbf{Z}_i\}_{i=1}^{n} \equiv \{\mathbf{Z}_i : i \in \mathcal{I}_k, k = 1, \ldots, \mathbb{K}\}$.

First of all, in order to estimate $\mathcal{J}(\boldsymbol{\theta}_0)$, we will assume here, for simplicity, that $g(\mathbf{x}, \boldsymbol{\theta})$ is continuously differentiable w.r.t. $\boldsymbol{\theta}$ at least in a neighbourhood of $\boldsymbol{\theta}_0$, with a derivative $\boldsymbol{\nabla} g(\mathbf{x}, \boldsymbol{\theta}) \in \mathbb{R}^d$, for every $\mathbf{x} \in \mathcal{X}$, and that $\mathbf{h}(\mathbf{X})g(\mathbf{X}, \boldsymbol{\theta})$ is further regular and/or smooth enough to allow interchange of derivatives (w.r.t. $\boldsymbol{\theta}$) and expectations (w.r.t. $\mathbf{X}$), so that we have: $\mathcal{J}(\boldsymbol{\theta}_0) = -\mathbb{E}_{\mathbf{X}}\{\mathbf{h}(\mathbf{X})\boldsymbol{\nabla} g'(\mathbf{X}, \boldsymbol{\theta}_0)\}$. Using $\widehat{\boldsymbol{\theta}}$ based on $\mathcal{L} \perp\!\!\!\perp \mathcal{U}$, followed by a Monte-Carlo on $\mathcal{U}$, we can then consistently estimate $\mathcal{J}(\boldsymbol{\theta}_0)$ and $\{\mathcal{J}(\boldsymbol{\theta}_0)\}^{-1}$ respectively as:

$$\widehat{\mathcal{J}}(\boldsymbol{\theta}_0) \;=\; -\,\frac{1}{N}\sum_{j=1}^{N}\mathbf{h}(\mathbf{X}_j)\boldsymbol{\nabla}'(\mathbf{X}_j, \widehat{\boldsymbol{\theta}}), \;\text{ and }\; \{\widehat{\mathcal{J}}(\boldsymbol{\theta}_0)\}^{-1} = \left\{-\,\frac{1}{N}\sum_{j=1}^{N}\mathbf{h}(\mathbf{X}_j)\boldsymbol{\nabla}'(\mathbf{X}_j, \widehat{\boldsymbol{\theta}})\right\}^{-1},$$

where we have implicitly assumed that $\widehat{\mathcal{J}}(\boldsymbol{\theta}_0)$ is indeed invertible. Next, for each $\mathbf{Z}_i \in \mathcal{L}_k$ and $k \in \{1, .., \mathbb{K}\}$, we estimate $\{\boldsymbol{\Psi}_0(\mathbf{Z}_i), \boldsymbol{\xi}_0(\mathbf{Z}_i)\}$ and $\{\boldsymbol{\Psi}_\mu(\mathbf{Z}_i), \boldsymbol{\xi}_\mu(\mathbf{Z}_i)\}$ as:

$$\widehat{\boldsymbol{\Psi}}_{0,k}(\mathbf{Z}_i) = \mathbf{h}(\mathbf{X}_i)\{Y_i - g(\mathbf{X}_i, \widehat{\boldsymbol{\theta}}_k)\}, \quad \widehat{\boldsymbol{\xi}}_{0,k}(\mathbf{Z}_i) = -\{\widehat{\mathcal{J}}(\boldsymbol{\theta}_0)\}^{-1}\widehat{\boldsymbol{\Psi}}_{0,k}(\mathbf{Z}_i), \text{ and}$$
$$\widehat{\boldsymbol{\Psi}}_{\mu,k}(\mathbf{Z}_i) = \mathbf{h}(\mathbf{X}_i)\{Y_i - \widehat{\mu}_k(\mathbf{X}_i)\}, \quad \widehat{\boldsymbol{\xi}}_{\mu,k}(\mathbf{Z}_i) = -\{\widehat{\mathcal{J}}(\boldsymbol{\theta}_0)\}^{-1}\widehat{\boldsymbol{\Psi}}_{\mu,k}(\mathbf{Z}_i).$$

Then, $\boldsymbol{\Sigma}\{\boldsymbol{\xi}_\mu(.)\}$, the asymptotic variance of $\widehat{\boldsymbol{\theta}}_\mu$ can be consistently estimated as:

$$\widehat{\boldsymbol{\Sigma}}\{\boldsymbol{\xi}_\mu(.)\} = n^{-1}\sum_{k=1}^{\mathbb{K}}\sum_{i\in\mathcal{I}_k}\widehat{\boldsymbol{\xi}}_{\mu,k}(\mathbf{Z}_i)\widehat{\boldsymbol{\xi}}'_{\mu,k}(\mathbf{Z}_i).$$

To estimate the combination matrix $\overline{\boldsymbol{\Delta}}$ in (2.22) and the asymptotic variance $\boldsymbol{\Sigma}_\mu(\overline{\boldsymbol{\Delta}})$ in (2.23) of the EASE estimator $\widehat{\boldsymbol{\theta}}^E_\mu$ consistently, let us define, $\forall\ 1 \le l \le d$,

$$\begin{aligned}
\widehat{\sigma}_{l,12} &= -\ n^{-1}\textstyle\sum_{k=1}^{\mathbb{K}}\sum_{i\in\mathcal{I}_k}\widehat{\boldsymbol{\xi}}_{0,k[l]}(\mathbf{Z}_i)\{\widehat{\boldsymbol{\xi}}_{\mu,k[l]}(\mathbf{Z}_i) - \widehat{\boldsymbol{\xi}}_{0[l]}(\mathbf{Z}_i)\},\\
\widehat{\sigma}_{l,22} &= n^{-1}\textstyle\sum_{k=1}^{\mathbb{K}}\sum_{i\in\mathcal{I}_k}\{\widehat{\boldsymbol{\xi}}_{\mu,k[l]}(\mathbf{Z}_i) - \widehat{\boldsymbol{\xi}}_{0,k[l]}(\mathbf{Z}_i)\}^2,
\end{aligned}$$

and $\widehat{\delta}_l = \widehat{\sigma}_{l,12}/(\widehat{\sigma}_{l,22} + \epsilon_n)$ for some sequence $\epsilon_n \to 0$ with $n^{\frac{1}{2}}\epsilon_n \to \infty$. Then, we estimate $\overline{\boldsymbol{\Delta}}$ and $\boldsymbol{\Sigma}_\mu(\overline{\boldsymbol{\Delta}})$ respectively as: $\widehat{\boldsymbol{\Delta}} = \text{diag}(\widehat{\delta}_1, ..., \widehat{\delta}_d)$ and

$$\widehat{\boldsymbol{\Sigma}}_\mu(\widehat{\boldsymbol{\Delta}}) = n^{-1}\sum_{k=1}^{\mathbb{K}}\sum_{i\in\mathcal{I}_k}\widehat{\boldsymbol{\xi}}_{\mu,k}(\mathbf{Z}_i, \widehat{\boldsymbol{\Delta}})\widehat{\boldsymbol{\xi}}'_{\mu,k}(\mathbf{Z}_i, \widehat{\boldsymbol{\Delta}}),$$

where $\widehat{\boldsymbol{\xi}}_{\mu,k}(\mathbf{Z}, \widehat{\boldsymbol{\Delta}}) = \widehat{\boldsymbol{\xi}}_{0,k}(\mathbf{Z}) + \widehat{\boldsymbol{\Delta}}\{\widehat{\boldsymbol{\xi}}_{\mu,k}(\mathbf{Z}) - \widehat{\boldsymbol{\xi}}_{0,k}(\mathbf{Z})\}\ \forall\ k \in \{1, \ldots, \mathbb{K}\}$. Normal confidence intervals (CIs) for the parameters of interest can also be constructed accordingly based on these variance estimates, and the asymptotically normal distribution of the estimators.

2.3.2 Construction of a Family of AIFs for SS Logistic Regression

Estimation problems for logistic regression (working) models are typically based on a maximum log-likelihood approach, so that the problem essentially corresponds to an M-estimation

problem, with the estimating equation being obtained as the derivative of the log-likelihood function defined by the working model. The underlying estimating function characterizing this EE is given by: $\boldsymbol{\psi}(Y, \mathbf{X}, \boldsymbol{\theta}) \equiv \mathbf{h}(\mathbf{X})\{Y - g(\mathbf{X}, \boldsymbol{\theta})\} = \overrightarrow{\mathbf{X}}\{Y - g_0(\overrightarrow{\mathbf{X}}'\boldsymbol{\theta}_0)\} \in \mathbb{R}^d$, where $d = (p+1)$, $\overrightarrow{\mathbf{X}} = (1, \mathbf{X}')'$, $\boldsymbol{\theta} = (\alpha, \boldsymbol{\beta}')'$, Y is (typically) a binary outcome $\in \{0, 1\}$, and $g_0(u) = \exp(u)/\{1 + \exp(u)\} \in [0, 1]\ \forall\ u \in \mathbb{R}$ denotes the appropriate 'expit' link function. Let $\overline{g}_0(a) = \log\{a/(1-a)\} \in \mathbb{R}\ \forall\ a \in (0, 1)$ denote the 'logit' function, the inverse of $g_0(.)$.

With $\boldsymbol{\psi}(.)$ satisfying the desired form (2.14) characterizing separable EEs, we use this setting to provide an illustration of the framework we have developed in section 2.3.1 for constructing flexible SS estimators $\widehat{\boldsymbol{\theta}}_\mu$, and corresponding EASE estimators $\widehat{\boldsymbol{\theta}}^E_\mu$, based on an AIF $\mu(.)$ and its estimator $\widehat{\mu}(.)$. We demonstrate here a family of choices of such a $\mu(.)$, and $\widehat{\mu}(.)$, for the SS logistic regression problem. These choices are primarily motivated by the idea of performing a lower dimensional smoothing, if desired, through appropriate use of dimension reduction techniques. For smoothing methods like KS, that can be quite inefficient in finite samples if p is even moderately large, such approaches are often desirable. In particular, they can be quite useful for the practical implementation of all our proposed SS estimators, the constructions for most of whom are based on the possible use of non-parametric (or semi-non-parametric) smoothing methods like KS. We begin with a few notations.

Let $r \leq p$ be a fixed positive integer and $\mathbf{P}_r = [\mathbf{p}_1, .., \mathbf{p}_r]_{p \times r}$ be any rank r transformation matrix. Let $\mathbf{X}_{\mathbf{P}_r} = \mathbf{P}_r'\mathbf{X}$. Given $(r, \mathbf{P}_r)$, we may now consider approximating the regression function $\mathbb{E}(Y|\mathbf{X})$ by smoothing Y over the r dimensional $\mathbf{X}_{\mathbf{P}_r}$ instead of the original $\mathbf{X} \in \mathbb{R}^p$. In general, $\mathbf{P}_r$ can be user-defined and data dependent. A few reasonable choices of $\mathbf{P}_r$ would be discussed shortly. If $\mathbf{P}_r$ depends only on $\mathbb{P}_{\mathbf{X}}$, it may be assumed to be known given the SS setting considered. If $\mathbf{P}_r$ also depends on the distribution of Y, then it needs to be estimated from $\mathcal{L}$ and the smoothing needs to be performed using the estimated $\mathbf{P}_r$.

For approximating $\mathbb{E}(Y|\mathbf{X})$, we may consider *any* reasonable smoothing technique $\mathcal{T}$ including, for instance, KS, kernel machine regression, smoothing splines etc. Let $m(\mathbf{x}; \mathbf{P}_r)$ denote the 'target function' for smoothing Y over $\mathbf{X}_{\mathbf{P}_r}$ using $\mathcal{T}$. For notational simplicity,

the dependence of $m(\mathbf{x};\mathbf{P}_r)$ and other quantities on $\mathcal{T}$ is suppressed throughout. For KS, the appropriate target is given by: $m(\mathbf{x};\mathbf{P}_r) = m_{\mathbf{P}_r}(\mathbf{P}_r'\mathbf{x})$, where $m_{\mathbf{P}_r}(\mathbf{z}) \equiv \mathbb{E}(Y|\mathbf{X}_{\mathbf{P}_r} = \mathbf{z})$. For basis function expansion based methods, $m(\mathbf{x};\mathbf{P}_r)$ will typically correspond to the L_2 projection of $m(\mathbf{x}) \equiv \mathbb{E}(Y|\mathbf{X} = \mathbf{x}) \in \mathcal{L}_2(\mathbb{P}_\mathbf{X})$ onto the functional space spanned by the basis functions associated with $\mathcal{T}$. Note that we do *not* assume $m(\mathbf{x};\mathbf{P}_r) = m(\mathbf{x})$ and hence, this is essentially a 'semi-non-parametric' (SNP) approach. Obviously, the case with $\mathbf{P}_r = I_p$ and $\mathcal{T} :=$ KS reduces to a fully non-parametric approach. With $\mathbf{P}_r$ and $m(\mathbf{x};\mathbf{P}_r)$ as defined above, let $\widehat{\mathbf{P}}_r$ and $\widehat{m}(\mathbf{x};\widehat{\mathbf{P}}_r)$ respectively denote their estimators based on $\mathcal{L}$. With $m(\mathbf{X},\mathbf{P}_r) = m_{\mathbf{P}_r}(\mathbf{P}_r'\mathbf{x})$, *one* choice of $\widehat{m}(\mathbf{x};\widehat{\mathbf{P}}_r)$ based on KS is given by:

$$\widehat{m}(\mathbf{x};\widehat{\mathbf{P}}_r) \equiv \widehat{m}_{KS}(\mathbf{x};\widehat{\mathbf{P}}_r) = \frac{\frac{1}{nh^r}\sum_{i=1}^n K\{(\widehat{\mathbf{P}}_r'\mathbf{X}_i - \widehat{\mathbf{P}}_r'\mathbf{x})/h\}Y_i}{\frac{1}{nh^r}\sum_{i=1}^n K\{(\widehat{\mathbf{P}}_r'\mathbf{X}_i - \widehat{\mathbf{P}}_r'\mathbf{x})/h\}}, \ \forall\, \mathbf{x} \in \mathcal{X},$$

where $K(.) : \mathbb{R}^r \to \mathbb{R}$ is some appropriate kernel function of order $q \geq 2$ and $h = h(n) > 0$ denotes the bandwidth sequence. With $(r, \mathbf{P}_r, \widehat{\mathbf{P}}_r)$ and $\{m(\mathbf{x};\mathbf{P}_r), \widehat{m}(\mathbf{x};\widehat{\mathbf{P}}_r)\}$ well-defined, we now formally define the AIF $\mu(.)$ and its estimator $\widehat{\mu}(.)$ as follows.

$$\mu(\mathbf{x}) \equiv \mu(\mathbf{x};\mathbf{P}_r) = g_0\left[\overrightarrow{\mathbf{x}}'\boldsymbol{\eta}_{\mathbf{P}_r} + \overline{g}_0\left\{m(\mathbf{x};\mathbf{P}_r)\right\}\right], \ \forall\, \mathbf{x} \in \mathcal{X}, \text{ and} \tag{2.24}$$

$$\widehat{\mu}(\mathbf{x}) \equiv \widehat{\mu}(\mathbf{x};\widehat{\mathbf{P}}_r) = g_0\left[\overrightarrow{\mathbf{x}}'\widehat{\boldsymbol{\eta}}_{\mathbf{P}_r} + \overline{g}_0\left\{\widehat{m}(\mathbf{x};\widehat{\mathbf{P}}_r)\right\}\right], \quad \text{where} \tag{2.25}$$

$\boldsymbol{\eta}_{\mathbf{P}_r}$ denotes the solution in η to: $\mathbb{E}\left\{\overrightarrow{\mathbf{X}}\left(Y - g_0\left[\overrightarrow{\mathbf{X}}'\eta + \overline{g}_0\left\{m(\mathbf{X};\mathbf{P}_r)\right\}\right]\right)\right\} = \mathbf{0}$, and

$\widehat{\boldsymbol{\eta}}_{\mathbf{P}_r}$ denotes the solution in η to: $\mathbb{P}_n\left\{\overrightarrow{\mathbf{X}}\left(Y - g_0\left[\overrightarrow{\mathbf{X}}'\eta + \overline{g}_0\left\{\widehat{m}(\mathbf{X};\widehat{\mathbf{P}}_r)\right\}\right]\right)\right\} = \mathbf{0}$.

Note first of all that owing to the definition of $\mu(.)$ in (2.24) and that of $\boldsymbol{\eta}_{\mathbf{P}_r}$, $\mu(.)$ naturally satisfies: $\mathbb{E}[\overrightarrow{\mathbf{X}}\{Y - \mu(\mathbf{X})\}] = \mathbf{0}$ and hence, $\mathbb{E}_\mathbf{X}[\boldsymbol{\psi}\{\mu(\mathbf{X}), \mathbf{X}, \boldsymbol{\theta}_0\}] = \mathbf{0}$, so that $\mu(.)$ is indeed an AIF in the sense of definition 2.2. Next, note that as long as $\overline{g}_0\{m(\mathbf{X};\mathbf{P}_r)\}$ is well-defined a.s. $[\mathbb{P}_\mathbf{X}]$, the existence and uniqueness of $\boldsymbol{\eta}_{\mathbf{P}_r}$ is clear as it essentially solves the expected EE corresponding to a (population based) logistic regression of Y w.r.t. $\mathbf{X}$ using $\overline{g}_0\{m(\mathbf{X};\mathbf{P}_r)\}$ as an *offset*. Similarly, as long as $\overline{g}_0\{\widehat{m}(\mathbf{X};\widehat{\mathbf{P}}_r)\}$ is well defined a.s. $[\mathbb{P}]$, the existence and

uniqueness of $\widehat{\eta}_{\mathbf{P}_r}$ is also clear, as it essentially solves the empirical EE corresponding to a logistic regression of Y w.r.t. $\mathbf{X}$ based on $\mathcal{L}$ using $\{\overline{g}_0\{\widehat{m}(\mathbf{X}_i;\widehat{\mathbf{P}}_r\} : i = 1,\ldots,n\}$ as a vector of *offsets*. In fact, this also reveals a *simple imputation based algorithm* for implementing the construction of the estimator $\widehat{\boldsymbol{\theta}}_\mu$ based on the above choices of $\mu(.)$ and $\widehat{\mu}(.)$ as follows.

Step (i) *Smoothing:* Choose $(r, \mathbf{P}_r, \mathcal{T})$, and use $\mathcal{L}$ to obtain $\widehat{\mathbf{P}}_r$ (if required) and $\widehat{m}(\mathbf{x};\widehat{\mathbf{P}}_r)$ based on $\mathcal{T}$ applied to $\{Y_i, (\widehat{\mathbf{P}}_r'\mathbf{X}_i)\}_{i=1}^n$.

Step (ii) *Refitting:* Fit a logistic regression, based on $\mathcal{L}$, of $\{Y_i\}_{i=1}^n$ w.r.t. $\{\mathbf{X}_i\}_{i=1}^n$ using $[\overline{g}_0\{\widehat{m}(\mathbf{X}_i,\widehat{\mathbf{P}}_r)\}]_{i=1}^n$ as a vector of *offsets*, to obtain the estimator $\widehat{\eta}_{\mathbf{P}_r}$.

Step (iii) *Imputation:* Define the SNP imputation function: $\widehat{\mu}(\mathbf{x}) \equiv \widehat{\mu}(\mathbf{x};\widehat{\mathbf{P}}_r)$, as in (2.25), by combining, in the appropriate scale, the predictions from the smoothing and the refitting steps. Use $\widehat{\mu}(.)$ to impute the missing Y in $\mathcal{U}$ as $\{\widehat{\mu}(\mathbf{X}_j)\}_{j=1}^N$.

Step (iv) *Final Step:* Fit a logistic regression, based on the imputed $\mathcal{U}$, of $\{\widehat{\mu}(\mathbf{X}_j)\}_{j=1}^N$ w.r.t. $\{\mathbf{X}_j\}_{j=1}^N$ to obtain the SS estimator $\widehat{\boldsymbol{\theta}}_\mu$ for the chosen $\mu(.)$ and $\widehat{\mu}(.)$.

It is straightforward to show that for the logistic regression problem, with the AIF $\mu(.)$ and its estimator $\widehat{\mu}(.)$ chosen to be as in (2.24)-(2.25), the SS estimator $\widehat{\boldsymbol{\theta}}_\mu$ obtained from the above algorithm is, in fact, identical to the one obtained from the general construction in (2.17), and therefore satisfies all the consequences of theorem 2.2 as long as the required assumptions hold. Further, all the rest of the developments in section 2.3.1, including the construction of the EASE estimator $\widehat{\boldsymbol{\theta}}_\mu^E$, as well as the associated inference procedures, also apply to the estimator $\widehat{\boldsymbol{\theta}}_\mu$ obtained from the above algorithm. Note further that in this case, the EASE estimator $\widehat{\boldsymbol{\theta}}_\mu^E$ is efficient and adaptive to the *mis-specification* of the underlying logistic regression (working) model given by: $\mathbb{E}(Y \mid \mathbf{X}) = g_0(\overrightarrow{\mathbf{X}}'\boldsymbol{\theta})$ for some $\boldsymbol{\theta} \in \mathbb{R}^p$, so that when either the model holds or the SNP imputation is sufficient, then the IF $\boldsymbol{\xi}_\mu(\mathbf{Z}, \overline{\boldsymbol{\Delta}})$ of $\widehat{\boldsymbol{\theta}}_\mu^E$ equals the 'efficient' IF $\boldsymbol{\xi}_{\text{eff}}(\mathbf{Z})$, so that $\widehat{\boldsymbol{\theta}}_\mu^E$ is indeed asymptotically optimal among all RAL estimators of $\boldsymbol{\theta}_0$ under $\mathcal{M}_{\mathbf{X}}$. Further, when neither cases hold, then $\widehat{\boldsymbol{\theta}}_\mu^E$ is no longer optimal,

but is *still* efficient and adaptive compared to $\widehat{\boldsymbol{\theta}}$ which, in this case, denotes the maximum likelihood estimator (MLE) for logistic regression. Lastly, if the SNP imputation above is certain to be sufficient (e.g. if $r = p$ and $\mathcal{T} := \text{KS}$), we may also simply define $\widehat{\boldsymbol{\theta}}^E_\mu = \widehat{\boldsymbol{\theta}}_\mu$.

It is also worth noting that a similar family of constructions and associated algorithms was studied in detail in Chakrabortty and Cai (2015) for the case of linear regression, where some interesting geometric intuitions and perspectives were also provided in order to motivate their corresponding refitting step. The version of the refitting step used here can be viewed as a natural generalization of their approach to the case of non-linear monotone link functions. The use of the inverse link $\overline{g}_0(.)$ in the construction of the offsets here is essentially inspired from the same geometric intuitions, and aims to ensure that the refitting step is performed in an appropriate scale. Moreover, while we have focussed here on logistic regression, the construction provided essentially applies to any canonical GLwM based on monotone link functions, including poisson regression (the exponential link) for instance. With appropriate modifications, we believe it can also be extended to even more general classes of GLwMs.

Finally, as far as the validity of the assumptions in theorem 2.2 for this case is concerned, assumptions 2.1-2.2, as well as condition (2.18) can be indeed verified to be true in this case. As for (2.19), if $\mathcal{X}$ is compact, and $m(\mathbf{X}; \mathbf{P}_r)$ and $\widehat{m}(\mathbf{X}; \widehat{\mathbf{P}}_r)$ are bounded away from 0/1 a.s. $[\mathbb{P}]$, then with $g_0(.)$ and $\overline{g}_0(.)$ both sufficiently smooth over their respective domains, (2.19) can be shown to hold as long as: $\sup_{\mathbf{x}\in\mathcal{X}} |\widehat{m}(\mathbf{x}; \widehat{\mathbf{P}}_r) - m(\mathbf{x}; \mathbf{P}_r)| \xrightarrow{P} 0$, and $\mathbb{G}_n[\overrightarrow{\mathbf{X}}\{\widehat{m}(\mathbf{x}; \widehat{\mathbf{P}}_r) - m(\mathbf{x}; \mathbf{P}_r)\}] \xrightarrow{P} \mathbf{0}$. For $\mathcal{T} := \text{KS}$, the validity of these conditions have been studied in detail in Theorem 4.1 and Lemmas A.2-A.3 of Chakrabortty and Cai (2015) under fairly mild conditions. In particular, they have shown that if $(\widehat{\mathbf{P}}_r - \mathbf{P}_r) = O_p(n^{-\frac{1}{2}})$, the result will indeed hold at the optimal bandwidth order of $O(n^{-1/(2q+r)})$ provided $q > r/2$. (This condition can be further removed using a CV based technique that is not pursued here for simplicity).

Choices of $\mathbf{P}_r$: We end this section with a brief discussion regarding the *choice* and estimation, if required, of the matrix $\mathbf{P}_r$ $(r < p)$ that may be used for dimension reduction,

if desired, while constructing the AIF $\mu(.)$ and its estimator $\widehat{\mu}(.)$. As noted earlier, the choice of $(r, \mathbf{P}_r)$ is essentially user-defined. While r would be typically chosen based on practical considerations including the size of n, as well as the choice of $\mathcal{T}$, $\mathbf{P}_r$ is allowed to be an unknown functional of $\mathbb{P}$. Simple choices of $\mathbf{P}_r$ include the r leading principal component directions of $\mathbf{X}$ or any r canonical directions of $\mathbf{X}$. Note that under the SS setting, $\mathbf{P}_r$ is effectively known if it only involves $\mathbb{P}_{\mathbf{X}}$. We now focus primarily on the case where $\mathbf{P}_r$ also depends on the distribution of Y and hence, is unknown in practice. Such a choice of $\mathbf{P}_r$ is often desirable to ensure that the smoothing is as 'sufficient' as possible for predicting Y. Several reasonable choices of such $\mathbf{P}_r$ and their estimation are possible based on non-parametric sufficient dimension reduction methods like Sliced Inverse Regression (SIR) (Li, 1991; Duan and Li, 1991), Principal Hessian Directions (PHD) (Li, 1992; Cook, 1998), Sliced Average Variance Estimation (SAVE) (Cook and Weisberg, 1991; Cook and Lee, 1999) etc.

Perhaps the most popular among these methods is the SIR approach, where the choice of $\mathbf{P}_r$ is given by: $\mathbf{P}_r^0$, the r leading eigenvectors of $\mathbb{M}_0 = \text{Var}\{\mathbb{E}(\mathbf{X}|Y)\}$, which leads to an optimal (in some sense) r-dimensional linear transform of $\mathbf{X}$ that can be predicted by Y. We refer the interested reader to Li (1991) and Duan and Li (1991) for more details on SIR and its properties, and to Chakrabortty and Cai (2015) for further discussions on the use of SIR in SS estimation problems, including a heuristic SS modification of the SIR algorithm.

While SIR, in general, is a reasonable approach for dimension reduction, especially for a continuous Y, it is unfortunately not quite suited for binary outcomes (which is what we consider here at least) since for such outcomes, SIR can provide only one non-trivial direction (see Cook and Lee (1999) for more details). A related approach that is more suited to handle binary outcomes is the SAVE approach of Cook and Lee (1999). For binary outcomes, an appropriate choice of $\mathbf{P}_r$ based on SAVE is given by: $\overline{\mathbf{P}}_r$, the r leading eigenvectors of $\overline{\mathbb{M}} = \{\rho(I_p - \Sigma_1)^2 + (1-\rho)(I_p - \Sigma_0)^2\}$, where $\rho = \mathbb{P}(Y = 1)$, and $\Sigma_y = \text{Var}(\mathbf{X}|Y = y)\ \forall\ y \in \{0, 1\}$.

2.3.3 Simulation Studies for SS Logistic Regression

We conducted extensive simulation studies to examine the finite sample performances of our general point and interval estimation procedures proposed in section 2.3.1 indexed by flexible choices of $\mu(.)$ and $\widehat{\mu}(.)$, for the case of SS logistic regression, wherein the $\mu(.)$ and $\widehat{\mu}(.)$ are chosen based on the SNP imputation strategy discussed in section 2.3.2. In all our results, we will denote the supervised estimator $\widehat{\boldsymbol{\theta}}$, the standard MLE for logistic regression, as MLE, and the corresponding SS estimators $\widehat{\boldsymbol{\theta}}_{\mu}$ and $\widehat{\boldsymbol{\theta}}_{\mu}^{E}$ as SNP and EASE resepctively. Throughout, we let $n = 500$, $N = 10000$, and considered two choices of p given by: $p = 10$ and $p = 20$. We generated $\mathbf{X}$ as: $\mathbf{X} \sim \mathcal{N}_p(\mathbf{0}, I_p)$ and restricted $\mathbf{X}$ to $[-5, 5]^p$ to ensure its boundedness. Given $\mathbf{X}$, we generated Y as: $Y \mid \mathbf{X} = \mathbf{x} \sim \text{Ber}\{m(\mathbf{x})\}$ for different choices of $m(.)$ to be discussed below. The SNP imputation was implemented using a dimension reduction step, wherein we used $r = 2$, and $\mathbf{P}_r$ was chosen and estimated based on the SAVE approach of Cook and Lee (1999). The estimator $\widehat{m}(\mathbf{x}, \widehat{\mathbf{P}}_r)$ was obtained using an r-dimensional local constant KS based on a 2^{nd} order Gaussian kernel with h estimated through maximum likelihood CV. The standard error estimates for all the estimators, as well the estimate $\widehat{\boldsymbol{\Delta}}$ of $\overline{\boldsymbol{\Delta}}$ in the definition of $\widehat{\boldsymbol{\theta}}_{\mu}^{E}$, were obtained based on $\mathbb{K}$-fold CV with $\mathbb{K} = 5$. The true values of the target parameter $\boldsymbol{\theta}_0 \equiv (\alpha_0, \boldsymbol{\beta}_0')'$ were estimated via monte carlo with a large sample size of $50,000$. For each configuration, the results are summarized based on 500 replications.

Choices of $m(\mathbf{x})$: For both choices of p, and with $g_0(u) \equiv \exp(u)/\{1 + \exp(u)\}\ \forall\ u \in \mathbb{R}$ denoting the link function, we investigated three different functional forms of $m(\mathbf{x})$ as follows:

(i) *Linear (Lin.)*: $m(\mathbf{x}) = g_0\left(\mathbf{x}'\mathbf{b}_p\right)$;

(ii) *Non-linear two component (NL2C)*: $m(\mathbf{x}) = g_0\left\{(\mathbf{x}'\mathbf{b}_p)(1 + \mathbf{x}'\boldsymbol{\delta}_p)\right\}$; and

(iii) *Non-linear three component (NL3C)*: $m(\mathbf{x}) = g_0\left\{(\mathbf{x}'\mathbf{b}_p)(1 + \mathbf{x}'\boldsymbol{\delta}_p) + (\mathbf{x}'\boldsymbol{\omega}_p)^2\right\}$;

where, we used $\mathbf{b}_p \equiv (\mathbf{1}'_{p/2}, \mathbf{0}'_{p/2})'$ for all models, $\boldsymbol{\delta}_p = (\mathbf{0}'_{p/2}, 2 * \mathbf{1}'_{p/2})'$ for models (ii)-(iii), and $\boldsymbol{\omega}_p = (1, 0, 1, 0, \ldots, 1, 0)'_{p\times 1}$; and for any a, $\mathbf{1}_a = (1, \ldots, 1)'_{a\times 1}$ and $\mathbf{0}_a = (0, \ldots, 0)'_{a\times 1}$. Note that the terms 'linear' and 'non-linear' appearing in the names of model (i) and models (ii)-(iii) respectively are inspired from the nature of the functional forms appearing inside the link $g_0(.)$, and they respectively correspond to the cases of the underlying logistic regression working model being correct or mis-specified. Moreover, for the non-linear models (ii)-(iii), note that the corresponding $m(\mathbf{x})$ depends on $\mathbf{x}$ through 2 and 3 dimensional linear transformations of $\mathbf{x}$ respectively. Through appropriate choices of $\mathbf{b}_p$, $\boldsymbol{\delta}_p$ and $\boldsymbol{\omega}_p$, as applicable, these models can incorporate commonly encountered quadratic and interaction effects. Lastly, with $\mathbf{X}$ normally distributed and $\mathbf{P}_r$ being chosen based on SAVE, results from (Cook and Lee, 1999) further imply that the SNP imputation with $r = 2$ is sufficient for models (i)-(ii), and insufficient for model (iii).

We summarize in tables 2.1-2.2 the overall relative efficiency (RE) of the proposed SNP ($\widehat{\boldsymbol{\theta}}_\mu$) and EASE ($\widehat{\boldsymbol{\theta}}^E_\mu$) estimators, compared to the MLE ($\widehat{\boldsymbol{\theta}}$) w.r.t. the empirical mean squared error (Emp. MSE), where for any estimator $\widetilde{\boldsymbol{\theta}}$ of $\boldsymbol{\theta}_0$, the Emp. MSE is summarized as the average of $\|\widetilde{\boldsymbol{\theta}} - \boldsymbol{\theta}_0\|^2$ over the 500 replications. As expected, both $\widehat{\boldsymbol{\theta}}_\mu$ and $\widehat{\boldsymbol{\theta}}^E_\mu$ are substantially more efficient than the MLE under model mis-specification, as is the case for the NL2C and NL3C models, while they are equally efficient as the MLE when the working model holds, as is the case for the linear model. Under the NL2C model, the SAVE based SNP imputation is expected to be sufficient and thus both $\widehat{\boldsymbol{\theta}}_\mu$ and $\widehat{\boldsymbol{\theta}}^E_\mu$ should be achieving the maximal efficiency gain, while under the NL3C model, with the the SNP imputation being insufficient, the corresponding efficiency gains for $\widehat{\boldsymbol{\theta}}_\mu$ and $\widehat{\boldsymbol{\theta}}^E_\mu$ are slightly less than those under the NL2C model. Lastly, it is also interesting to note that for most of the cases, the EASE and the SNP estimators achieve nearly identical efficiencies. However in practice, the final combination step involved in the construction of EASE should still be performed in order to ensure theoretically the efficient and adaptive property of EASE, a property that is possessed by $\widehat{\boldsymbol{\theta}}^E_\mu$, but not in general by $\widehat{\boldsymbol{\theta}}_\mu$ unless the AIF $\mu(.)$ turns out to be equal to $m(.)$.

Table 2.1: Comparison of the MLE, SNP and EASE estimators based on Emp. MSE under models (i), (ii) and (iii) for $p = 10$. Shown also are the relative efficiencies (RE) of all the estimators w.r.t. the corresponding MLE for each of the three models considered.

Criteria ↓	Lin. Model			NL2C Model			NL3C Model		
	MLE	SNP	EASE	MLE	SNP	EASE	MLE	SNP	EASE
Emp. MSE	0.210	0.208	0.209	0.097	0.040	0.040	0.108	0.046	0.046
RE w.r.t. MLE	1.000	**1.012**	**1.004**	1.000	**2.421**	**2.416**	1.000	**2.357**	**2.345**

Table 2.2: Comparison of the MLE, SNP and EASE estimators based on Emp. MSE under models (i), (ii) and (iii) for $p = 20$. Shown also are the relative efficiencies (RE) of all the estimators w.r.t. the corresponding MLE for each of the three models considered.

Criteria ↓	Lin. Model			NL2C Model			NL3C Model		
	MLE	SNP	EASE	MLE	SNP	EASE	MLE	SNP	EASE
Emp. MSE	0.644	0.608	0.616	0.192	0.104	0.105	0.216	0.133	0.134
RE w.r.t. MLE	1.000	**1.058**	**1.045**	1.000	**1.841**	**1.831**	1.000	**1.631**	**1.619**

To examine the performance of the CV based inference procedures proposed in section 2.3.1, we also obtained the standard error (SE) estimates and the corresponding CIs for the SNP and EASE estimators $\widehat{\boldsymbol{\theta}}_\mu$ and $\widehat{\boldsymbol{\theta}}_\mu^E$. In tables 2.3-2.4, we present the bias, empirical SE (ESE), the average of the estimated SE (ASE) and the coverage probability (CovP) of the 95% CIs for each component of $\widehat{\boldsymbol{\theta}}_\mu$ and $\widehat{\boldsymbol{\theta}}_\mu^E$ under the linear and NL2C models with $p = 10$. In general, we find that both $\widehat{\boldsymbol{\theta}}_\mu$ and $\widehat{\boldsymbol{\theta}}_\mu^E$ have negligible biases and further, they are similar, if not smaller, in magnitudes to the corresponding biases of the MLE under both models. The ASEs are close to the corresponding ESEs, and the CovPs are close to the nominal level of 95% for both $\widehat{\boldsymbol{\theta}}_\mu$ and $\widehat{\boldsymbol{\theta}}_\mu^E$ under each of the models, suggesting that the proposed variance estimation procedure works well in practice with $\mathbb{K} = 5$. Under the linear model, both $\widehat{\boldsymbol{\theta}}_\mu$ and $\widehat{\boldsymbol{\theta}}_\mu^E$ have similar magnitudes of standard errors as the MLE, as we would expect. Under the NL2C model, compared to the MLE, both $\widehat{\boldsymbol{\theta}}_\mu$ and $\widehat{\boldsymbol{\theta}}_\mu^E$ are substantially more efficient, at the cost of virtually no additional biases, across all components of $\boldsymbol{\theta}_0$.

Table 2.3: Bias, ESE, ASE and CovP of $\widehat{\boldsymbol{\theta}}_\mu$ and $\widehat{\boldsymbol{\theta}}_\mu^E$ for estimating $\boldsymbol{\theta}_0$ under the linear model with $p = 10$. Shown also are the bias and ESE of the MLE for comparison. The true parameter value under this model is given by: $\boldsymbol{\theta}_0 = (\alpha_0, \beta_{01}, ..., \beta_{010})'$, as tabulated below.

Parameter	MLE ($\widehat{\boldsymbol{\theta}}$)		SNP ($\widehat{\boldsymbol{\theta}}_\mu$)				EASE ($\widehat{\boldsymbol{\theta}}_\mu^E$)			
	Bias	ESE	Bias	ESE	ASE	CovP	Bias	ESE	ASE	CovP
$\alpha_0 = 0$	0.005	0.126	0.005	0.128	0.126	**0.95**	0.005	0.127	0.125	**0.95**
$\beta_{01} = 1$	0.031	0.156	0.021	0.156	0.147	**0.95**	0.027	0.156	0.146	**0.94**
$\beta_{02} = 1$	0.041	0.142	0.031	0.141	0.148	**0.96**	0.037	0.142	0.147	**0.96**
$\beta_{03} = 1$	0.046	0.139	0.038	0.139	0.148	**0.98**	0.042	0.139	0.147	**0.98**
$\beta_{04} = 1$	0.041	0.143	0.030	0.142	0.148	**0.97**	0.036	0.143	0.147	**0.96**
$\beta_{05} = 1$	0.053	0.150	0.042	0.150	0.149	**0.95**	0.049	0.150	0.148	**0.94**
$\beta_{06} = 0$	0.000	0.136	0.000	0.136	0.126	**0.92**	0.001	0.136	0.125	**0.92**
$\beta_{07} = 0$	0.008	0.123	0.008	0.125	0.126	**0.94**	0.007	0.124	0.125	**0.95**
$\beta_{08} = 0$	0.000	0.119	0.001	0.121	0.127	**0.98**	0.000	0.120	0.126	**0.98**
$\beta_{09} = 0$	0.002	0.126	0.001	0.126	0.126	**0.96**	0.002	0.126	0.125	**0.96**
$\beta_{010} = 0$	-0.006	0.125	-0.006	0.125	0.126	**0.95**	-0.006	0.124	0.125	**0.95**

Table 2.4: Bias, ESE, ASE and CovP of $\widehat{\boldsymbol{\theta}}_\mu$ and $\widehat{\boldsymbol{\theta}}_\mu^E$ for estimating $\boldsymbol{\theta}_0$ under the NL2C model with $p = 10$. Shown also are the bias and ESE of the MLE for comparison. The true parameter value under this model is given by: $\boldsymbol{\theta}_0 = (\alpha_0, \beta_{01}, ..., \beta_{010})'$, as tabulated below.

Parameter	OLS ($\widehat{\boldsymbol{\theta}}$)		SNP ($\widehat{\boldsymbol{\theta}}_\mu$)				EASE ($\widehat{\boldsymbol{\theta}}_\mu^E$)			
	Bias	ESE	Bias	ESE	ASE	CovP	Bias	ESE	ASE	CovP
$\alpha_0 = 0$	0.001	0.093	0.000	0.067	0.064	**0.93**	0.000	0.067	0.064	**0.93**
$\beta_{01} = 0.125$	-0.003	0.097	-0.001	0.060	0.062	**0.96**	-0.001	0.060	0.062	**0.97**
$\beta_{02} = 0.125$	0.003	0.095	0.004	0.059	0.062	**0.96**	0.004	0.060	0.062	**0.96**
$\beta_{03} = 0.125$	-0.000	0.091	0.005	0.058	0.062	**0.95**	0.005	0.058	0.062	**0.95**
$\beta_{04} = 0.125$	0.003	0.091	0.005	0.061	0.062	**0.95**	0.005	0.061	0.062	**0.95**
$\beta_{05} = 0.125$	0.001	0.090	0.004	0.059	0.062	**0.95**	0.004	0.060	0.062	**0.95**
$\beta_{06} = 0$	0.005	0.098	0.001	0.059	0.062	**0.96**	0.001	0.059	0.062	**0.96**
$\beta_{07} = 0$	-0.012	0.092	-0.004	0.061	0.062	**0.95**	-0.005	0.061	0.062	**0.96**
$\beta_{08} = 0$	-0.002	0.099	0.001	0.060	0.062	**0.96**	0.001	0.059	0.062	**0.96**
$\beta_{09} = 0$	-0.000	0.093	-0.003	0.060	0.062	**0.95**	-0.003	0.060	0.062	**0.95**
$\beta_{010} = 0$	0.001	0.094	0.004	0.061	0.062	**0.94**	0.004	0.061	0.062	**0.94**

2.3.4 Application of SS Logistic Regression to EMR Data

We applied our proposed SS estimation procedure for logistic regression, based on the SNP imputation strategy discussed in section 2.3.2, to an EMR study of rheumatoid arthritis (RA), a systemic autoimmune (AI) disease, conducted at the Partners HealthCare. Further

details on this study can be found in Liao et al. (2010, 2013). The study cohort consists of 44014 patients, and the binary outcome of interest in this case was a disease phenotype defined as clinically confirmed diagnosis of RA. The primary goal was to understand and model the disease risk of RA based on several relevant clinical variables, including RA biomarkers, standard medications for RA, as well other relevant AI diseases and/or clinical conditions known to be closely related to RA, rich information for all of which were available through the data for a large number of patients. However, the availability of gold standard outcomes was limited as it required logistically prohibitive manual chart review by the physician. A labeled training data was therefore only available for a random subset of 500 patients, wherein observations for the gold standard outcome were obtained through manual chart review by two expert rheumatologists, thereby leading to a SS set-up. The empirical estimate, based on the labeled data, of the population prevalence of RA was found to be 99/500.

In order to model the disease risk of RA, we related it to a set of 27 covariates altogether available through the dataset, which included: (i) age, gender, (ii) counts of ICD9 diagnostic codes for RA, (iii) counts of mentions, extracted from the physicians' notes via natural language processing (NLP), of other related AI diseases like psoriatic arthritis (PsA) and juvenile rheumatoid arthritis (JRA), (iv) codified test results and/or NLP extracted mentions of positivity for standard RA biomarkers including rheumatoid factor (RF), anti-cyclic citrullinated polypeptide (anti-CCP) and anti-tumor necrosis factors (anti-TNF) that are routinely checked for RA patients to assess the disease progression, (vi) counts of codified and/or NLP extracted mentions of methotrexate and azathioprine (frequently used medications for RA and other AI diseases), seropositivity, as well as several other standard medications and/or relevant clinical conditions that are known to be related to RA, including Anak, Arava, Enb, Gld, Hum, Neo, Pen, Plaq, PMR, Rem, Rit, Sulf and other medications (other meds). A detailed glossary of the abbreviations used above, as well as further explanations regarding the clinical significance of these variables can be found in Liao et al. (2010, 2013). In our tabulated results, all variables representing codified mentions will be

denoted with a suffix '-r', while those corresponding to NLP extracted mentions will be denoted with a suffix '-nlp'. All the count/binary variables were further log-transformed as: $x \to \log(1+x)$, to increase stability of the model fitting. In order to ensure comparability of the point estimates for the regression coefficients across all the predictors, all the covariates were further standardized to have unit variance w.r.t. the full data, and all our results are reported in this standardized scale for the covariates.

Based on these 27 covariates, we implemented the EASE estimator $\widehat{\boldsymbol{\theta}}_{\mu}^{E}$ for logistic regression, using the SNP imputation strategy in section 2.3.2 for constructing the AIF $\mu(.)$ and its estimator $\widehat{\mu}(.)$ in section 2.3.1. The SNP imputation was implemented using a dimension reduction step with $r = 2$, and $\mathbf{P}_r$ was chosen and estimated based on SAVE (Cook and Lee, 1999). The estimator $\widehat{m}(\mathbf{x}, \widehat{\mathbf{P}}_r)$ was obtained using an r-dimensional local constant KS based on a 2^{nd} order Gaussian kernel with h estimated through maximum likelihood CV. For comparison, we also implemented the corresponding supervised estimator $\widehat{\boldsymbol{\theta}}$, the MLE for logistic regression, based on the labeled data. In addition, for both $\widehat{\boldsymbol{\theta}}$ and $\widehat{\boldsymbol{\theta}}_{\mu}^{E}$, we obtained their respective standard error (SE) estimates based on our inference procedure in section 2.3.1, wherein we used a $\mathbb{K}$-fold CV with $\mathbb{K} = 5$. In table 2.5, we present the coordinate-wise estimates (Est.) of the regression parameters based on $\widehat{\boldsymbol{\theta}}$ and $\widehat{\boldsymbol{\theta}}_{\mu}^{E}$, along with their respective estimated SEs and the corresponding p-values (Pval.) based on these estimates. Shown also are the coordinate-wise estimated relative efficiencies (REs) of $\widehat{\boldsymbol{\theta}}_{\mu}^{E}$ w.r.t. $\widehat{\boldsymbol{\theta}}$.

The point estimates of $\boldsymbol{\theta}_0$ based on $\widehat{\boldsymbol{\theta}}$ and $\widehat{\boldsymbol{\theta}}_{\mu}^{E}$ in table 2.5 are all quite close in general, which is desirable and reassuring as it establishes, in a real data, the consistency and stability of EASE. Further, the estimated REs of $\widehat{\boldsymbol{\theta}}_{\mu}^{E}$ w.r.t. $\widehat{\boldsymbol{\theta}}$ are all greater than 1 indicating the improved efficiency of EASE over the MLE. The efficiency gains for most of the variables are notably quite substantial. Apart from a few cases, where the gains range between 10-30%, for most of the other variables, the gains typically range between 50% to 200%, and in some cases, reach exceptionally high values as well.

Table 2.5: Comparison of the SS and supervised logistic regression estimators applied to the EMR dataset. Shown are the estimates (Est.) of the standardized regression coefficients based on the MLE and the EASE, along with their estimated standard errors (SE) and the p-values (Pval.) associated with testing the null effect for each of the predictors. Shown also are the coordinate-wise relative efficiencies (RE) of the EASE compared to the MLE.

Predictors	MLE ($\widehat{\boldsymbol{\theta}}$)			EASE ($\widehat{\boldsymbol{\theta}}^E_\mu$)			RE of EASE
	Est.	SE	Pval.	Est.	SE	Pval.	
Age	0.516	0.273	0.059	0.698	0.253	0.006	**1.159**
Gender	-0.081	0.193	0.675	-0.067	0.172	0.699	**1.260**
RA-r	1.506	0.375	0.000	1.337	0.339	0.000	**1.217**
JRA-nlp	-0.664	0.466	0.154	-0.485	0.285	0.089	**2.667**
PsA-nlp	-0.773	0.243	0.002	-0.804	0.181	0.000	**1.796**
Anti-CCP-r	-0.032	0.190	0.865	-0.027	0.182	0.882	**1.082**
Anti-CCP-nlp	0.320	0.222	0.148	0.385	0.168	0.022	**1.731**
RF-nlp	0.018	0.219	0.936	-0.011	0.177	0.952	**1.517**
Seropositive-nlp	0.661	0.301	0.028	0.510	0.250	0.041	**1.447**
Anti-TNF-r	0.190	0.307	0.536	0.185	0.225	0.410	**1.866**
Anti-TNF-nlp	-0.056	0.452	0.902	0.024	0.241	0.920	**3.524**
Methotrexate-nlp	0.606	0.259	0.019	0.659	0.216	0.002	**1.436**
Anak-nlp	0.030	0.302	0.920	0.098	0.236	0.679	**1.644**
Arava-r	-0.653	0.369	0.077	-0.681	0.281	0.016	**1.724**
Arava-nlp	0.588	0.361	0.103	0.438	0.280	0.119	**1.658**
Azathioprine-nlp	-0.609	0.322	0.059	-0.700	0.196	0.000	**2.693**
Enb-nlp	0.148	0.392	0.706	0.316	0.203	0.119	**3.752**
Gld-r	0.127	0.284	0.655	0.134	0.279	0.632	**1.034**
Hum-nlp	0.148	0.287	0.606	-0.114	0.183	0.535	**2.471**
Neo-nlp	-0.133	0.379	0.726	-0.060	0.228	0.793	**2.765**
Pen-nlp	-0.179	0.220	0.414	0.066	0.081	0.416	**7.292**
Plaq-r	-0.716	0.356	0.045	-0.604	0.285	0.034	**1.561**
Rem-r	0.151	0.355	0.671	0.178	0.162	0.272	**4.809**
Rit-nlp	0.170	0.236	0.472	0.168	0.134	0.210	**3.127**
Sulf-r	-0.104	0.301	0.731	-0.214	0.231	0.354	**1.701**
Other Meds-r	0.697	0.387	0.072	0.665	0.327	0.042	**1.401**
PMR	-0.184	0.130	0.157	-0.186	0.117	0.111	**1.238**
(Intercept)	-6.263	1.056	0.000	-6.578	0.966	0.000	**1.195**

As a consequence of this improved efficiency of $\widehat{\boldsymbol{\theta}}^E_\mu$, we also note that apart from all the variables that are deemed significant (at the 5% level) by both $\widehat{\boldsymbol{\theta}}$ and $\widehat{\boldsymbol{\theta}}^E_\mu$, based on their respective estimated p-values, $\widehat{\boldsymbol{\theta}}^E_\mu$ also additionally found several other variables including

Age, Anti-CCP-nlp, Arava-r, Azathioprine-r, and Other Meds-R to be significant at the 5% level. Most of the variables found significant by both $\widehat{\boldsymbol{\theta}}$ and $\widehat{\boldsymbol{\theta}}_{\mu}^{E}$ are indeed known to be related directly or indirectly to RA, so that their associations detected based on the data are perhaps reasonable and not unexpected. However, some of the additional associations detected by $\widehat{\boldsymbol{\theta}}_{\mu}^{E}$, and not by $\widehat{\boldsymbol{\theta}}$, are particularly interesting, especially as they include CCP, a standard RA biomarker, as well as Arava and Azathioprine which are both frequently used medications for RA and other AI diseases. Hence, these variables are indeed of substantial clinical relevance to RA and therefore, their associations detected by EASE based on the available data are perhaps worth investigating further using other datasets to examine the reproducibility of the detected associations.

2.4 SS Sliced Inverse Regression (SS-SIR)

In this section, we consider a SS modification of the well-known Sliced Inverse Regression (SIR) approach (Li, 1991; Duan and Li, 1991) for sufficient dimension reduction (SDR). The SDR problem has received considerable interest in recent years, and several approaches for SDR have been proposed over the last two decades, including SIR which is one of the earliest and perhaps the most popular one, as well as other related approaches like Principal Hessian Directions (PHD) (Li, 1992; Cook, 1998), Sliced Average Variance Estimation (SAVE) (Cook and Weisberg, 1991; Cook and Lee, 1999) etc. that are also commonly used.

SDR methods provide useful techniques for data visualization as well as understanding lower dimensional structures that are often implicit in otherwise high dimensional data, and may contain all the relevant information (and hence are 'sufficient') for characterizing the relationship between the outcome Y and a set of covariates $\mathbf{X} \in \mathbb{R}^p$, with p possibly high. Consequently, SDR also provides a reasonable gateway to bypass the curse of dimensionality that one inevitably encounters in non-parametric regression based on smoothing methods like KS that do not scale well with the dimension of $\mathbf{X}$, with convergence rates slowing down exponentially with p, thereby often resulting in poor finite sample performance and over-

fitting bias. We refer the interested reader to Cook (2009) for a detailed discussion on SDR and associated methods, as well as a comprehensive overview of the relevant literature. The underlying lower dimensional model, characterizing the relationship between Y and $\mathbf{X}$, that typically motivates the construction of most SDR approaches can be represented as follows.

$$Y = f(\boldsymbol{\beta}_1'\mathbf{X}, \ldots, \boldsymbol{\beta}_r'\mathbf{X}; \epsilon) \quad \text{for some } r \leq p, \text{ and some } \{\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_r\} \in \mathbb{R}^p, \quad \text{where (2.26)}$$

$f(.) : \mathbb{R}^{r+1} \to \mathcal{Y}$ is some *unknown* 'link' function and $\epsilon \perp\!\!\!\perp \mathbf{X}$ denotes a random noise.

Equivalently, (2.26) can also be represented as: $(Y \perp\!\!\!\perp \mathbf{X}) \mid \{\boldsymbol{\beta}_1'\mathbf{X}, \ldots, \boldsymbol{\beta}_r'\mathbf{X}\}$.

The collection $\{\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_r\}$ in (2.26) is assumed to be linearly independent w.l.o.g. Note that since the link function $f(.)$ in (2.26) is allowed to be completely unknown (upto basic measurability and moment based restrictions to comply with our starting assumptions), the directions $\{\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_r\}$ are only identifiable upto scalar multiples. In other words, only the r-dimensional span $\mathcal{B}_r \equiv \text{span}\{\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_r\} \subseteq \mathbb{R}^p$ of $\{\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_r\}$, and $\mathcal{P}_r$, the rank r orthogonal projection matrix onto $\mathcal{B}_r$, are identifiable under (2.26). The essential goal of most SDR approaches, including SIR, is to efficiently estimate $\mathcal{B}_r$ (or $\mathcal{P}_r$).

The central quantity of interest in SIR is the matrix: $\mathbb{M}_0 = \text{Var}_Y\{\mathbb{E}(\mathbf{X} \mid Y)\}$. Under the model (2.26), and with $\mathbb{E}(\mathbf{X}) = \mathbf{0}$ and $\text{Var}(\mathbf{X}) = I_p$ as assumed w.l.o.g., Duan and Li (1991) and Li (1991) have shown that if a 'design linearity condition' (DLC) holds regarding the underlying design distribution of $\mathbf{X}$, wherein $\mathbb{E}(\mathbf{v}'\mathbf{X} \mid \boldsymbol{\beta}_1'\mathbf{X}, \ldots, \boldsymbol{\beta}_r'\mathbf{X})$ is a linear function of $\{\boldsymbol{\beta}_1'\mathbf{X}, \ldots, \boldsymbol{\beta}_r'\mathbf{X}\}$ for each $\mathbf{v} \in \mathbb{R}^p$, a condition that holds for all elliptically symmetric distributions including the multivariate normal distribution, then the span of $\mathbf{P}_r \equiv \{\mathbf{p}_1, \ldots, \mathbf{p}_r\}$, the r leading eigenvectors of $\mathbb{M}_0$, indeed equals the span $\mathcal{B}_r$ of $\{\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_r\}$. Further discussions on these properties of $\mathbb{M}_0$ and $\mathbf{P}_r$, as well as the validity and applicability of the DLC condition, can be found in Duan and Li (1991), Li (1991) and Hall and Li (1993).

Further, *regardless* of whether a model of the form (2.26) actually holds or not, the directions $\{\mathbf{p}_1, \ldots, \mathbf{p}_r\}$ *always* have the strong interpretability of being the r most 'predictable'

directions of $\mathbf{X}$ given Y. Specifically, they correspond to the solutions of a sequence of maximization problems as follows: $\mathbf{p}_j = \arg\max_{\mathbf{v}\in\mathcal{A}_j} \mathrm{Var}\{\mathbb{E}(\mathbf{v}'\mathbf{X} \mid Y)\}/\mathrm{Var}(\mathbf{v}'\mathbf{X})\ \forall\ 1 \leq j \leq r$, where $\mathcal{A}_1 = \mathbb{R}^p$ and $\mathcal{A}_j = \{\mathbf{v} \in \mathbb{R}^p : \mathbf{v} \perp (\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_{j-1}\}\ \forall\ 2 \leq j \leq r$. Further discussions on these optimality properties of $\mathbf{P}_r$ and their interpretations can be found in Li (1991).

The SIR Approach: Under our assumed setting, with $\mathbb{E}(\mathbf{X}) = \mathbf{0}$ and $\mathrm{Var}(\mathbf{X}) = I_p$ w.l.o.g., the original (supervised) SIR algorithm of Li (1991) essentially proceeds as follows:

(i) Partition the range of Y into L mutually disjoint slices: $\{\mathcal{I}_1, \ldots, \mathcal{I}_L\} \equiv \mathfrak{I}_L$ (say), for some given choices of L and $\mathfrak{I}_L$.

(ii) For $1 \leq l \leq L$, let $\widehat{\theta}_l$ denote the proportion of $\{Y_i\}_{i=1}^n$ in slice $\mathcal{I}_l$ i.e.

$$\widehat{\theta}_l = \frac{1}{n}\sum_{i=1}^{n} 1(Y_i \in \mathcal{I}_l),$$

and for each $\mathcal{I}_l$, let $\widehat{\mathbf{m}}_l$ denote the sample average of the set: $\{\mathbf{X}_i : Y_i \in \mathcal{I}_l\}$ i.e.

$$\widehat{\mathbf{m}}_l = \frac{1}{n\widehat{\theta}_l}\sum_{i=1}^{n} \mathbf{X}_i\{1(Y_i \in \mathcal{I}_l)\} = \frac{\widehat{\boldsymbol{\mu}}_l}{\widehat{\theta}_l},\ \text{ where } \widehat{\boldsymbol{\mu}}_l = \frac{1}{n}\sum_{i=1}^{n} \mathbf{X}_i\{1(Y_i \in \mathcal{I}_l)\}.$$

(iii) Then, the SIR method essentially considers the matrix:

$$\widehat{\mathbb{M}}_L \ \equiv\ \widehat{\mathbb{M}}(\mathfrak{I}_L) = \sum_{l=1}^{L} \widehat{\theta}_l(\widehat{\mathbf{m}}_l\widehat{\mathbf{m}}_l') \ \equiv\ \sum_{l=1}^{L} \frac{\widehat{\boldsymbol{\mu}}_l\widehat{\boldsymbol{\mu}}_l'}{\widehat{\theta}_l},$$

and estimates the r most 'predictable' directions of $\mathbf{X}$ given Y as: $\widehat{\mathbf{P}}_r \equiv \{\widehat{\mathbf{p}}_1, \ldots, \widehat{\mathbf{p}}_r\}$, the r leading eigenvectors of $\widehat{\mathbb{M}}_L$ for any $r \leq p$, regardless of the validity of a model of the form (2.26). Moreover, if (2.26) actually does hold, and so does the DLC condition, then the span of $\widehat{\mathbf{P}}_r$ also $n^{\frac{1}{2}}$-consistently estimates the span $\mathcal{B}_r$ of $\{\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_r\}$.

The theoretical properties of SIR, including the $n^{\frac{1}{2}}$-consistency of the SIR estimates, are well established (Li, 1991; Duan and Li, 1991; Hsing and Carroll, 1992; Zhu and Ng, 1995)

under a variety of settings with or without requiring a model of the form (2.26) to hold. In particular, Li (1991) and Duan and Li (1991), in their original papers on SIR, establish the $n^{\frac{1}{2}}$-consistency of (the span of) $\widehat{\mathbf{P}}_r$, as an estimator of $\mathcal{B}_r$, assuming (2.26) and the DLC condition, and under an asymptotic regime where $(L, \mathfrak{I}_L)$ are considered fixed. It needs to be noted that *even* under such a regime, the SIR based estimators $\widehat{\mathbb{M}}_L$ and $\widehat{\mathbf{P}}_r$ above target a matrix $\mathbb{M}_L$ (to be made more precise shortly) and the corresponding collection $\mathbf{P}_{r,\,L}$ of the r leading eigenvectors of $\mathbb{M}_L$ respectively, so that the span of $\mathbf{P}_{r,\,L}$ *still* equals $\mathcal{B}_r$ as long as (2.26) and the DLC condition holds. Later, Hsing and Carroll (1992) and Zhu and Ng (1995) established the $n^{\frac{1}{2}}$-consistency of the SIR estimates under a more general setting, wherein (2.26) or the DLC condition are not required to hold, and $(L, \mathfrak{I}_L)$ follows an asymptotic regime where L is allowed to diverge (slowly enough) with n and each slice contains an equal number of observations, and the results were established by simply treating $\widehat{\mathbb{M}}_L$ as an estimator of $\mathbb{M}_0$, and $\widehat{\mathbf{P}}_r$ as an estimator of $\mathbf{P}_r$. However, it has also been noted in Zhu and Ng (1995) that even though L can be allowed to diverge, too many slices with too few observations tends to degrade the performance of SIR both in terms of asymptotic theory, as well as in finite samples. Further, as noted in Li (1991), the performance of the SIR estimates tends to be fairly robust to the choice of the number of slices (and the slicing pattern as well), as long as they are chosen reasonably enough (i.e. not too large or not too small). Hence for all practical purposes, and regardless of the validity of (2.26) and the DLC condition, it is perhaps not unreasonable to assume the asymptotic regime of Li (1991) and Duan and Li (1991), so that $(L, \mathfrak{I}_L)$ is considered fixed. In fact, this is the underlying regime we shall assume here for proposing and analyzing our semi-supervised SIR (SS-SIR) algorithm.

The Semi-Supervised Sliced Inverse Regression (SS-SIR) Approach: Throughout the rest of this section, we shall adopt the entire set-up introduced in section 2.3, including all notations and basic assumptions regarding the SS setting. The SS-SIR algorithm we propose

here is based on a methodology that is largely borrowed from, and is essentially a special case of, the general framework we have developed in section 2.3 for SS M-estimation problems. In particular, it is a direct consequence of multiple uses of the general EASE estimator $\widehat{\boldsymbol{\theta}}^*$ constructed in (2.3) for some appropriate choices of the corresponding estimating functions. The main property that the SS-SIR provably (under appropriate assumptions) possesses is that it is always at least as efficient as the SIR, and in most cases, which would be more clear once we state the results, it would be more efficient than the SIR approach.

In order to formalize the details of SS-SIR, we first introduce some notations as follows. For any slicing pattern $(L, \mathfrak{I}_L)$, with a given *fixed* choice of the number of slices L and the collection $\mathfrak{I}_L \equiv \{\mathcal{I}_1, \ldots, \mathcal{I}_L\}$, a disjoint partition of the support $\mathcal{Y}$ of Y, let us define:

$$\mathbb{Z}_l \;=\; 1(Y \in \mathcal{I}_l), \;\text{ and }\; \mathbb{X}_l \;=\; \mathbf{X}\mathbb{Z}_l \;\equiv\; \mathbf{X}\{1(Y \in \mathcal{I}_l)\} \quad \forall\, 1 \leq l \leq L.$$

So, if we assume an asymptotic regime where $(L, \mathfrak{I}_L)$ is fixed and given, then the original supervised SIR approach *essentially targets* the following matrix:

$$\mathbb{M}_L \;\equiv\; \mathbb{M}(\mathfrak{I}_L) \;=\; \sum_{l=1}^{L} \frac{\boldsymbol{\mu}_l \boldsymbol{\mu}_l'}{\theta_l} \;\equiv \mathrm{Var}[\mathbb{E}\{\mathbf{X} \,|\, (\mathbb{Z}_1, \ldots, \mathbb{Z}_L)\}], \quad \text{where,} \;\; \forall\, 1 \leq l \leq L,$$
$$\theta_l \;=\; \mathbb{E}(\mathbb{Z}_l) \;\equiv \mathbb{E}\{1(Y \in \mathcal{I}_l)\}, \;\text{ and }\; \boldsymbol{\mu}_l \;=\; \mathbb{E}(\mathbb{X}_l) \;\equiv\; \mathbb{E}[\mathbf{X}\{1(Y \in \mathcal{I}_l)\}].$$

The parameter $\theta_l \equiv \mathbb{P}(Y \in \mathcal{I}_l)$, for each l, is implicitly assumed to be strictly positive w.l.o.g. Our goal would be to obtain efficient SS estimators of the matrix $\mathbb{M}_L$ above, and compare it to the (supervised) estimator $\widehat{\mathbb{M}}_L$ given by the SIR. Note that the parameters $\{\theta_l\}_{l=1}^L$ and $\{\boldsymbol{\mu}_l\}_{l=1}^L$, and consequently $\mathbb{M}_L$, inherently depend on $\mathbb{P}_{\mathbf{X}}$, so that improved SS estimation of $\mathbb{M}_L$, compared to the supervised SIR estimator $\widehat{\mathbb{M}}_L$, is indeed possible. Further, note that the efficient estimation of $\mathbb{M}_L$ essentially boils down to the efficient estimation of the (finite) collections of parameters given by: $\{\theta_l\}_{l=1}^L$ and $\{\boldsymbol{\mu}_l\}_{l=1}^L$, all of which are expectations of some

random variable/vector. For each $1 \leq l \leq L$, the SIR approach estimates them as follows:

$$\theta_l \;=\; \mathbb{E}(\mathbb{Z}_l) \;\text{ as }\; \widehat{\theta}_l \;=\; \frac{1}{n}\sum_{i=1}^{n} \mathbb{Z}_{l,i}\;, \;\; \text{where } \mathbb{Z}_{l,i} = 1(Y_i \in \mathcal{I}_l) \;\; \forall\, 1 \leq i \leq n,$$

$$\boldsymbol{\mu}_l \;=\; \mathbb{E}(\mathbb{X}_l) \;\text{ as }\; \widehat{\boldsymbol{\mu}}_l \;=\; \frac{1}{n}\sum_{i=1}^{n} \mathbb{X}_{l,i}\;, \;\; \text{where } \mathbb{X}_{l,i} = \mathbf{X}_i \mathbb{Z}_{l,i} \;\; \forall\, 1 \leq i \leq n,$$

$$\text{and hence,} \quad \mathbb{M}_L \;\equiv\; \sum_{l=1}^{L} \frac{\boldsymbol{\mu}_l \boldsymbol{\mu}_l'}{\theta_l} \;\text{ as }\; \widehat{\mathbb{M}}_L \;\equiv\; \sum_{l=1}^{L} \frac{\widehat{\boldsymbol{\mu}}_l \widehat{\boldsymbol{\mu}}_l'}{\widehat{\theta}_l}.$$

A key observation regarding the original SIR approach is that it depends on the observed outcomes $\{Y_i\}_{i=1}^n$ *only* through their corresponding slice indicators $\{(\mathbb{Z}_{l,i})_{l=1}^L\}_{i=1}^n$. Hence, if the dependence of these slice indicator variables $\{\mathbb{Z}_l\}_{l=1}^L$ on $\mathbf{X}$, in terms of their respective conditional means for instance, is smooth enough, then this smoothness can be exploited to create a reasonable prediction rule based on $\mathcal{L}$, preferably using some non-parametric methods applied to the observed $\{(\mathbb{Z}_{l,i})_{l=1}^L\}_{i=1}^n$ and $\{\mathbf{X}_i\}_{i=1}^n$ in $\mathcal{L}$, for predicting $\{\mathbb{Z}_l\}_{l=1}^L$ given $\mathbf{X}$. This can then be used to predict the unobserved slice indicator variables in $\mathcal{U}$, and subsequently the imputed $\mathcal{U}$ can now be incorporated into a standard SIR type approach to develop a SS version of SIR. Motivated by these intuitions, a heuristic SS-SIR approach, with promising empirical performance, was proposed in Chakrabortty and Cai (2015) based on a nearest neighbour approximation. The SS-SIR algorithm proposed here is also motivated by similar intuitions, but is based on a more formal and rigorous approach, wherein we directly focus on efficient SS estimation of the parameters $\{\theta_l\}_{l=1}^L$ and $\{\boldsymbol{\mu}_l\}_{l=1}^L$, and consequently $\mathbb{M}_L$, using the general framework developed in section 2.3 for constructing EASE estimators. To this end, we next define the conditional means of $\mathbb{Z}_l$ and $\mathbb{X}_l$ given $\mathbf{X}$, given by:

$$\theta_l(\mathbf{x}) = \mathbb{E}(\mathbb{Z}_l \,|\, \mathbf{X} = \mathbf{x}) \;\text{ and }\; \boldsymbol{\mu}_l(\mathbf{x}) = \mathbb{E}(\mathbb{X}_l \,|\, \mathbf{X} = \mathbf{x}) \;= \mathbf{x}\theta_l(\mathbf{x}) \;\; \forall\, 1 \leq l \leq L; \;\; \forall\, \mathbf{x} \in \mathcal{X}.$$

Further, let $\widehat{\theta}_l(\mathbf{x})$ denote *any* reasonable estimator of $\theta_l(\mathbf{x})$, and let $\widehat{\boldsymbol{\mu}}_l(\mathbf{x}) = \mathbf{x}\widehat{\theta}_l(\mathbf{x})$ denote the corresponding estimator of $\boldsymbol{\mu}_l(\mathbf{x})$, $\forall\, 1 \leq l \leq L$, $\forall\, \mathbf{x} \in \mathcal{X}$. More discussions regarding the

choice of $\widehat{\theta}_l(.)$ would be provided later. We now define the EASE estimators for θ_l and $\boldsymbol{\mu}_l$ as:

$$\widehat{\theta}_l^* = \frac{1}{N}\sum_{j=1}^{N}\widehat{\theta}_l(\mathbf{X}_j) - \frac{1}{n}\sum_{i=1}^{n}\left\{\widehat{\theta}_l(\mathbf{X}_i) - \mathbb{Z}_{l,i}\right\} \ \forall\, 1 \le l \le L, \quad \text{and} \tag{2.27}$$

$$\widehat{\boldsymbol{\mu}}_l^* = \frac{1}{N}\sum_{j=1}^{N}\widehat{\boldsymbol{\mu}}_l(\mathbf{X}_j) - \frac{1}{n}\sum_{i=1}^{n}\{\widehat{\boldsymbol{\mu}}_l(\mathbf{X}_i) - \mathbb{X}_{l,i}\} \ \forall\, 1 \le l \le L. \tag{2.28}$$

Based on these EASE estimators and the pre-defined choices of the slicing pattern $(L, \mathfrak{I}_L)$, we then propose to estimate $\mathbb{M}_L$ through the *SS-SIR* algorithm, based on:

$$\widehat{\mathbb{M}}_L^* = \sum_{l=1}^{L}\frac{\widehat{\boldsymbol{\mu}}_l^*\widehat{\boldsymbol{\mu}}_l^{*'}}{\widehat{\theta}_l^*},$$

and accordingly estimate the r most predictable directions of $\mathbf{X}$ given Y through $\widehat{\mathbf{P}}_r^* \equiv \{\widehat{\mathbf{p}}_r^*, \ldots, \widehat{\mathbf{p}}_r^*\}$, the r leading eigenvectors of $\widehat{\mathbb{M}}_L^*$ for any $r \le p$. We next summarize the properties of the estimators used to construct $\widehat{\mathbb{M}}_L^*$, and their comparison with the corresponding estimators used to construct the supervised SIR estimator $\widehat{\mathbb{M}}_L$, through the following result.

Theorem 2.3. *For any RAL estimator $\widetilde{\theta}$, let As.Var $(\widetilde{\theta})$ denote the asymptotic variance of $\widetilde{\theta}$ i.e. the variance of the IF of $\widetilde{\theta}$. Now, assume that the following conditions hold:*

$$\sup_{\mathbf{x}\in\mathcal{X}}\left|\widehat{\theta}_l(\mathbf{x}) - \theta_l(\mathbf{x})\right| \xrightarrow{P} 0, \qquad \overline{G}_{n,l} \equiv \mathbb{G}_n\left\{\widehat{\theta}_l(\mathbf{X}) - \theta_l(\mathbf{X})\right\} \xrightarrow{P} 0 \quad \forall\, 1 \le l \le L, \ \textit{and}$$

$$\sup_{\mathbf{x}\in\mathcal{X}}\left\|\mathbf{x}\{\widehat{\theta}_l(\mathbf{x}) - \theta_l(\mathbf{x})\}\right\| \xrightarrow{P} 0, \qquad \overline{\mathbf{G}}_{n,l} \equiv \mathbb{G}_n\left[\mathbf{X}\{\widehat{\theta}_l(\mathbf{X}) - \theta_l(\mathbf{X})\}\right] \xrightarrow{P} \mathbf{0} \quad \forall\, 1 \le l \le L.$$

Then, the estimators involved in the SS-SIR algorithm satisfy: $\forall\, 1 \le l \le L$,

$$n^{\frac{1}{2}}\left(\widehat{\theta}_l^* - \theta_l\right) = n^{-\frac{1}{2}}\sum_{i=1}^{n}\{\mathbb{Z}_{l,i} - \mathbb{E}(\mathbb{Z}_l \mid \mathbf{X}_i)\} + O_p\left(\overline{G}_{n,l}\right) + O_p\left(\frac{n}{N}\right)^{\frac{1}{2}} + o_p(1).$$

$$n^{\frac{1}{2}}\left(\widehat{\boldsymbol{\mu}}_l^* - \boldsymbol{\mu}_l\right) = n^{-\frac{1}{2}}\sum_{i=1}^{n}\{\mathbb{X}_{l,i} - \mathbb{E}(\mathbb{X}_l \mid \mathbf{X}_i)\} + O_p\left(\overline{\mathbf{G}}_{n,l}\right) + O_p\left(\frac{n}{N}\right)^{\frac{1}{2}} + o_p(1).$$

On the other hand, the estimators involved in the SIR algorithm satisfy: $\forall\ 1 \leq l \leq L$,

$$\begin{aligned} n^{\frac{1}{2}}\left(\widehat{\theta}_l - \theta_l\right) &= n^{-\frac{1}{2}} \sum_{i=1}^{n} \{\mathbb{Z}_{l,i} - \mathbb{E}(\mathbb{Z}_l)\} + o_p(1). \\ n^{\frac{1}{2}}\left(\widehat{\boldsymbol{\mu}}_l - \boldsymbol{\mu}_l\right) &= n^{-\frac{1}{2}} \sum_{i=1}^{n} \{\mathbb{X}_{l,i} - \mathbb{E}(\mathbb{X}_l)\} + o_p(1). \end{aligned}$$

Hence, the following comparisons hold: $\forall\ 1 \leq l \leq L$,

$$\begin{aligned} &As.Var\left(\widehat{\theta}_l^*\right) \preceq As.Var\left(\widehat{\theta}_l\right), \ \ with\ strict\ inequality\ unless\ \theta_l(\mathbf{X}) = \theta_l\ a.s.\ [\mathbb{P}_{\mathbf{X}}]. \\ &As.Var\left(\widehat{\boldsymbol{\mu}}_l^*\right) \preceq As.Var\left(\widehat{\boldsymbol{\mu}}_l\right), \ \ with\ strict\ inequality\ unless\ \boldsymbol{\mu}_l(\mathbf{X}) = \boldsymbol{\mu}_l\ a.s.\ [\mathbb{P}_{\mathbf{X}}]. \\ &Further,\ with\ \widehat{\mathbf{m}}_l^* = \frac{\widehat{\boldsymbol{\mu}}_l^*}{\widehat{\theta}_l^*}\ \ and\ \ \widehat{\mathbf{m}}_l = \frac{\widehat{\boldsymbol{\mu}}_l}{\widehat{\theta}_l}, \ \ As.Var\left(\widehat{\mathbf{m}}_l^*\right) \preceq As.Var\left(\widehat{\mathbf{m}}_l\right)\ \forall\ l. \end{aligned}$$

Lastly, $\forall\ \mathbf{v} \in \mathbb{R}^p$, $As.Var\left(\widehat{\mathbb{M}}_L^*\mathbf{v}\right) \preceq As.Var\left(\widehat{\mathbb{M}}_L\mathbf{v}\right)$. *Similar inequalities, although somewhat difficult to show, also continue to hold for the corresponding eigenvectors of* $\widehat{\mathbb{M}}_L^*$ *and* $\widehat{\mathbb{M}}_L$.

Theorem 2.3 therefore clearly establishes the efficient and adaptive nature of the SS-SIR algorithm compared to the supervised SIR approach, in terms of every component estimator involved in the two approaches, as well as the main matrix estimators themselves. Finally, regarding the choice of the (possibly) non-parametric smoothing based estimator $\widehat{\theta}_l(.)$ of $\theta_l(.)$, one choice could be based on KS, and this could be based on smoothing over the whole of $\mathbf{X} \in \mathbb{R}^p$, or if an r-dimensional model, as in (2.26), with $r < p$ is actually assumed to hold, then the smoothing can be based on a lower dimensional transformation $\widehat{\mathbf{P}}_r'\mathbf{X} \in \mathbb{R}^r$ where $\widehat{\mathbf{P}}_r$ is obtained from the supervised SIR as an initial $n^{\frac{1}{2}}$-consistent estimator of $\mathcal{B}_r$ (so that the r-dimensional smoothing based on the estimated covariates will indeed be sufficient). As far as the conditions for theorem 2.3 are concerned, as long as $\mathcal{X}$ is compact and the smoothing is based on KS, the requirements can be satisfied under fairly mild regularity conditions. We refer the interested reader to Theorem 4.1 and Lemmas A.1-A.2 in Chakrabortty and Cai (2015) for further details regarding these results. In particular, with r-dimensional KS for any $r \leq p$, under reasonable conditions, the requirements of theorem 2.3 will hold without any

under-smoothing, so that the optimal bandwidth order $O(n^{-1/(2q+r)})$ can be used. However, a kernel order of $q > r/2$ will be needed. As noted earlier in a different context, this condition can indeed be avoided using a CV based technique that is not pursued in this paper.

2.4.1 Simulation Studies for SS-SIR

We conducted extensive simulation studies to compare the performance of our proposed SS-SIR algorithm to that of the SIR approach, under a variety of scenarios involving continuous as well as binary Y, generated based on SDR models, as in (2.26), with choices of r given by: $r = 1$ and $r = 2$ for continuous Y, and $r = 1$ for binary Y, a case where the SIR is known (Cook and Lee, 1999) to be able to recover only one non-trivial direction. Throughout, we set $n = 500$, $N = 10000$, and used two choices of p given by: $p = 10$ and $p = 30$. $\mathbf{X}$ was generated as: $\mathbf{X} \sim \mathcal{N}_p(\mathbf{0}, I_p)$. For continuous Y, we used three choices of L given by: $L = 10, 20$ and 50, and the slices $\{\mathcal{I}_l : l = 1, \ldots, L\}$ were chosen to each have an equal number of observations from $\mathcal{L}$, as is typical in the standard SIR literature. For binary Y, the choice of L is trivially 2 and the two slices correspond to the two distinct values of Y. For all the settings considered above, and for each choice of $(L, \mathfrak{I}_L)$, we implemented both the SIR and the SS-SIR approaches. For SS-SIR, the estimators $\{\widehat{\theta}_l(.) : l = 1, \ldots, L\}$ were obtained from $\mathcal{L}$ based on r-dimensional local constant KS of Y on the transformed covariates $\widehat{\mathbf{P}}_r'\mathbf{X}$, where $\widehat{\mathbf{P}}_r$ denotes the corresponding SIR estimator. (Note that under the SDR model assumption, and with $\mathbf{X}$ normally distributed, such a smoothing would indeed be sufficient). All the KS steps were performed using a 2^{nd} order Gaussian kernel with the bandwidth chosen based on maximum likelihood CV. Additionally, for the SDR models with $r = 1$ (i.e. single index models), we also implemented the OLS estimator for continuous Y, and the MLE based on logistic regression for binary Y. With $\mathbf{X}$ normally distributed, so that the DLC condition holds, results from Li and Duan (1989) imply that $\mathcal{B}_r$, the span of the parameter $\boldsymbol{\beta}_1$ of interest, can also be validly estimated by the (span of) these estimators.

SDR models for Y: We next enumerate the various SDR models used for the generation of Y, depending on its nature (continuous or binary), as well as the choice of r.

(i) With $r = 1$ and Y continuous, we used the following SDR models:

(a) *Model* $\mathcal{M}_1^{(1)}$*:* $Y = \boldsymbol{\beta}_0'\mathbf{X} + \epsilon$, with $\epsilon \perp\!\!\!\perp \mathbf{X}$ and $\epsilon \sim \mathcal{N}_1(0,1)$.

(b) *Model* $\mathcal{M}_2^{(1)}$*:* $Y = \boldsymbol{\beta}_0'\mathbf{X} + \frac{1}{4}(\boldsymbol{\beta}_0'\mathbf{X})^2 + \frac{1}{6}(\boldsymbol{\beta}_0'\mathbf{X})^3 + \epsilon$, with $\epsilon \perp\!\!\!\perp \mathbf{X}$ and $\epsilon \sim \mathcal{N}_1(0,1)$.

(c) *Model* $\mathcal{M}_3^{(1)}$*:* $Y = (\boldsymbol{\beta}_0'\mathbf{X} + \epsilon)^3$, with $\epsilon \perp\!\!\!\perp \mathbf{X}$ and $\epsilon \sim \mathcal{N}_1(0,1)$.

(ii) For the case $r = 2$, with Y continuous, we generated Y using the following SDR models:

(a) *Model* $\mathcal{M}_1^{(2)}$*:* $Y = \boldsymbol{\beta}_0'\mathbf{X} + (\boldsymbol{\gamma}_0'\mathbf{X})^3 + \epsilon$, with $\epsilon \perp\!\!\!\perp \mathbf{X}$ and $\epsilon \sim \mathcal{N}_1(0,1)$.

(b) *Model* $\mathcal{M}_2^{(2)}$*:* $Y = \boldsymbol{\beta}_0'\mathbf{X} + \{\exp(\boldsymbol{\beta}_0'\mathbf{X})\}(\boldsymbol{\gamma}_0'\mathbf{X}) + \epsilon$, with $\epsilon \perp\!\!\!\perp \mathbf{X}$ and $\epsilon \sim \mathcal{N}_1(0,1)$.

(iii) Lastly, for binary Y, with $r = 1$, we used the following SDR models for generating Y:

(a) *Model* $\mathcal{M}_1^{(b)}$*:* $Y = 1\{(\boldsymbol{\beta}_0'\mathbf{X} + \epsilon) > 0\}$, with $\epsilon \perp\!\!\!\perp \mathbf{X}$ and $\epsilon \sim$ Logistic $(0,1)$.

(b) *Model* $\mathcal{M}_2^{(b)}$*:* $Y = 1[\{\boldsymbol{\beta}_0'\mathbf{X} + \frac{1}{3}(\boldsymbol{\beta}_0'\mathbf{X})^3 + \epsilon\} > 0]$, with $\epsilon \perp\!\!\!\perp \mathbf{X}$ and $\epsilon \sim$ Logistic $(0,1)$.

(c) *Model* $\mathcal{M}_3^{(b)}$*:* $Y = 1[\{(\boldsymbol{\beta}_0'\mathbf{X})^3 + \epsilon\} > 0]$, with $\epsilon \perp\!\!\!\perp \mathbf{X}$ and $\epsilon \sim \mathcal{N}_1(0,1)$.

For all the models, the parameter $\boldsymbol{\beta}_0$ was chosen to be: $(\mathbf{1}_{p/2}', \mathbf{0}_{p/2}')'$, and $\boldsymbol{\gamma}_0$, wherever needed, was chosen to be: $(\mathbf{0}_{p/2}', \mathbf{1}_{p/2}')'$. For all the settings, we replicated the simulations over 500 iterations. For convenience of further discussion, let us define $\mathcal{P}_r$ to be the orthogonal projection matrix onto: the span $\mathcal{B}_r$ of $\boldsymbol{\beta}_0$ for the case $r = 1$, or the span $\mathcal{B}_r$ of $\{\boldsymbol{\beta}_0, \boldsymbol{\gamma}_0\}$ for the case $r = 2$. The error measures we used for comparing the performances of the SS-SIR and SIR estimators, as well as those of the OLS or MLE, wherever applicable, are as follows:

(i) For $r = 1$, and for any generic estimator $\widetilde{\boldsymbol{\beta}}$ of the $\boldsymbol{\beta}_0$ direction, that is further sign normalized w.r.t. $\boldsymbol{\beta}_0$ as $\widetilde{\boldsymbol{\beta}}'\boldsymbol{\beta}_0 \geq 0$ w.l.o.g., we used two error measures: (a) the empirical mean squared error (Emp. MSE) of the normalized version of $\widetilde{\boldsymbol{\beta}}$ w.r.t. the corresponding normalized version of $\boldsymbol{\beta}_0$, defined as: the average of $\|\widetilde{\boldsymbol{\beta}}/\|\widetilde{\boldsymbol{\beta}}\| - \boldsymbol{\beta}_0/\|\boldsymbol{\beta}_0\|\|^2$ over the

500 iterations, and (b) the average of $(1 - R_0^2) \equiv \{1 - (\widetilde{\boldsymbol{\beta}}'\mathcal{P}_r\widetilde{\boldsymbol{\beta}})/(\widetilde{\boldsymbol{\beta}}'\widetilde{\boldsymbol{\beta}})\}$ over the 500 iterations. We denote the corresponding averaged measure of $(1 - R_0^2)$ as: $(1 - \bar{R}_0^2)$.

(ii) For $r = 2$, and for any generic estimators $\{\widetilde{\boldsymbol{\beta}}, \widetilde{\boldsymbol{\gamma}}\}$ of the span of $\{\boldsymbol{\beta}_0, \boldsymbol{\gamma}_0\}$, characterizing the underlying SDR model, we used the error measures: (a) the average of $(1 - R_1^2) \equiv \{1 - (\widetilde{\boldsymbol{\beta}}_1'\mathcal{P}_r\widetilde{\boldsymbol{\beta}}_1)/(\widetilde{\boldsymbol{\beta}}_1'\widetilde{\boldsymbol{\beta}}_1)\}$ over the 500 iterations, for the first direction, and (b) the average of $(1 - R_2^2) \equiv \{1 - (\widetilde{\boldsymbol{\beta}}_2'\mathcal{P}_r\widetilde{\boldsymbol{\beta}}_2)/(\widetilde{\boldsymbol{\beta}}_2'\widetilde{\boldsymbol{\beta}}_2)\}$ over the 500 iterations, for the second direction. We denote the respective averaged measures as $(1 - \bar{R}_1^2)$ and $(1 - \bar{R}_2^2)$.

The error measures $(1-\bar{R}_0^2)$ in (i), and $(1-\bar{R}_1^2)$ and $(1-\bar{R}_2^2)$ in (ii) above, are all scale invariant measures, and are fairly standard choices as performance criteria in the SIR literature. We refer the interested reader to Duan and Li (1991) and Li (1991) for further discussions on them. For all the error measures considered above, and for each choice of the slicing pattern $(L, \mathfrak{I}_L)$, we also report the corresponding relative efficiencies (RE) of the SS-SIR estimator w.r.t. all other supervised estimators considered, defined as the inverse ratio of the error measure, based on the respective criteria, for the SS-SIR estimator to those for the supervised estimators. The simulation results for all the settings are summarized next in tables 2.6-2.17.

Table 2.6: Comparison of the SS-SIR, SIR and OLS estimators based on Emp. MSE and $(1 - \bar{R}_0^2)$ under various choices of $(L, \mathfrak{I}_L)$, for model $\mathcal{M}_1^{(1)}$ with $r = 1$ and $p = 10$.

Criteria ↓	Slices = 10			Slices = 20			Slices = 50		
	SS-SIR	SIR	OLS	SS-SIR	SIR	OLS	SS-SIR	SIR	OLS
Emp. MSE	0.005	0.022	0.004	0.004	0.023	0.004	0.005	0.026	0.004
RE of SS-SIR	1.000	**4.652**	0.768	1.000	**5.196**	0.805	1.000	**5.433**	0.762
$(1 - \bar{R}_0^2)$	0.005	0.022	0.004	0.004	0.023	0.004	0.005	0.026	0.004
RE of SS-SIR	1.000	**4.628**	0.768	1.000	**5.166**	0.806	1.000	**5.397**	0.762

Table 2.7: Comparison of the SS-SIR, SIR and OLS estimators based on Emp. MSE and $(1 - \bar{R}_0^2)$ under various choices of $(L, \mathfrak{I}_L)$, for model $\mathcal{M}_2^{(1)}$ with $r = 1$ and $p = 10$.

Criteria ↓	Slices = 10			Slices = 20			Slices = 50		
	SS-SIR	SIR	OLS	SS-SIR	SIR	OLS	SS-SIR	SIR	OLS
Emp. MSE	0.002	0.020	0.007	0.002	0.020	0.007	0.002	0.022	0.008
RE of SS-SIR	1.000	**9.886**	3.487	1.000	**12.390**	4.455	1.000	**13.643**	4.620
$(1 - \bar{R}_0^2)$	0.002	0.020	0.007	0.002	0.020	0.007	0.002	0.022	0.008
RE of SS-SIR	1.000	**9.832**	3.482	1.000	**12.322**	4.446	1.000	**13.558**	4.611

Table 2.8: Comparison of the SS-SIR, SIR and OLS estimators based on Emp. MSE and $(1 - \bar{R}_0^2)$ under various choices of $(L, \mathfrak{I}_L)$, for model $\mathcal{M}_3^{(1)}$ with $r = 1$ and $p = 10$.

Criteria ↓	Slices = 10			Slices = 20			Slices = 50		
	SS-SIR	SIR	OLS	SS-SIR	SIR	OLS	SS-SIR	SIR	OLS
Emp. MSE	0.005	0.022	0.017	0.005	0.022	0.017	0.005	0.026	0.017
RE of SS-SIR	1.000	**4.563**	3.608	1.000	**4.841**	3.749	1.000	**5.292**	3.464
$(1 - \bar{R}_0^2)$	0.005	0.022	0.017	0.005	0.022	0.017	0.005	0.026	0.017
RE of SS-SIR	1.000	**4.540**	3.594	1.000	**4.815**	3.734	1.000	**5.257**	3.451

Table 2.9: Comparison of the SS-SIR, SIR and OLS estimators based on Emp. MSE and $(1 - \bar{R}_0^2)$ under various choices of $(L, \mathfrak{I}_L)$, for model $\mathcal{M}_1^{(1)}$ with $r = 1$ and $p = 30$.

Criteria ↓	Slices = 10			Slices = 20			Slices = 50		
	SS-SIR	SIR	OLS	SS-SIR	SIR	OLS	SS-SIR	SIR	OLS
Emp. MSE	0.010	0.062	0.004	0.009	0.063	0.004	0.009	0.069	0.004
RE of SS-SIR	1.000	**6.459**	0.443	1.000	**7.093**	0.462	1.000	**7.332**	0.436
$(1 - \bar{R}_0^2)$	0.010	0.061	0.004	0.009	0.062	0.004	0.009	0.068	0.004
RE of SS-SIR	1.000	**6.370**	0.444	1.000	**6.990**	0.463	1.000	**7.215**	0.437

Table 2.10: Comparison of the SS-SIR, SIR and OLS estimators based on Emp. MSE and $(1-\bar{R}_0^2)$ under various choices of $(L, \mathfrak{I}_L)$, for model $\mathcal{M}_2^{(1)}$ with $r=1$ and $p=30$.

Criteria ↓	Slices = 10			Slices = 20			Slices = 50		
	SS-SIR	SIR	OLS	SS-SIR	SIR	OLS	SS-SIR	SIR	OLS
Emp. MSE	0.006	0.059	0.031	0.005	0.060	0.032	0.006	0.067	0.032
RE of SS-SIR	1.000	**10.459**	5.524	1.000	**11.866**	6.274	1.000	**11.624**	5.623
$(1-\bar{R}_0^2)$	0.006	0.058	0.031	0.005	0.060	0.032	0.006	0.066	0.032
RE of SS-SIR	1.000	**10.311**	5.485	1.000	**11.691**	6.226	1.000	**11.434**	5.581

Table 2.11: Comparison of the SS-SIR, SIR and OLS estimators based on Emp. MSE and $(1-\bar{R}_0^2)$ under various choices of $(L, \mathfrak{I}_L)$, for model $\mathcal{M}_3^{(1)}$ with $r=1$ and $p=30$.

Criteria ↓	Slices = 10			Slices = 20			Slices = 50		
	SS-SIR	SIR	OLS	SS-SIR	SIR	OLS	SS-SIR	SIR	OLS
Emp. MSE	0.009	0.061	0.045	0.008	0.061	0.044	0.010	0.070	0.045
RE of SS-SIR	1.000	**6.510**	4.825	1.000	**7.217**	5.207	1.000	**7.417**	4.765
$(1-\bar{R}_0^2)$	0.009	0.060	0.044	0.008	0.060	0.044	0.009	0.069	0.045
RE of SS-SIR	1.000	**6.422**	4.776	1.000	**7.116**	5.155	1.000	**7.298**	4.718

Table 2.12: Comparison of the SS-SIR and SIR estimators based on $(1-\bar{R}_1^2)$ and $(1-\bar{R}_2^2)$ under various choices of $(L, \mathfrak{I}_L)$, for model $\mathcal{M}_1^{(2)}$ with $r=2$ and $p=10$.

Criteria ↓	Slices = 10		Slices = 20		Slices = 50	
	SS-SIR	SIR	SS-SIR	SIR	SS-SIR	SIR
First Direction						
$(1-\bar{R}_1^2)$	0.001	0.017	0.001	0.018	0.001	0.019
RE of SS-SIR	1.000	**14.185**	1.000	**17.019**	1.000	**17.059**
Second Direction						
$(1-\bar{R}_2^2)$	0.053	0.122	0.065	0.130	0.121	0.199
RE of SS-SIR	1.000	**2.310**	1.000	**2.005**	1.000	**1.642**

Table 2.13: Comparison of the SS-SIR and SIR estimators based on $(1-\bar{R}_1^2)$ and $(1-\bar{R}_2^2)$ under various choices of $(L, \mathfrak{I}_L)$, for model $\mathcal{M}_2^{(2)}$ with $r=2$ and $p=10$.

Criteria ↓	Slices = 10		Slices = 20		Slices = 50	
	SS-SIR	SIR	SS-SIR	SIR	SS-SIR	SIR
First Direction						
$(1-\bar{R}_1^2)$	0.005	0.023	0.005	0.025	0.006	0.028
RE of SS-SIR	1.000	**4.335**	1.000	**4.829**	1.000	**4.575**
Second Direction						
$(1-\bar{R}_2^2)$	0.018	0.053	0.020	0.055	0.030	0.065
RE of SS-SIR	1.000	**2.866**	1.000	**2.750**	1.000	**2.200**

Table 2.14: Comparison of the SS-SIR and SIR estimators based on $(1-\bar{R}_1^2)$ and $(1-\bar{R}_2^2)$ under various choices of $(L, \mathfrak{I}_L)$, for model $\mathcal{M}_1^{(2)}$ with $r=2$ and $p=30$.

Criteria ↓	Slices = 10		Slices = 20		Slices = 50	
	SS-SIR	SIR	SS-SIR	SIR	SS-SIR	SIR
First Direction						
$(1-\bar{R}_1^2)$	0.006	0.056	0.005	0.058	0.006	0.065
RE of SS-SIR	1.000	**9.565**	1.000	**11.349**	1.000	**11.139**
Second Direction						
$(1-\bar{R}_2^2)$	0.321	0.410	0.416	0.464	0.693	0.643
RE of SS-SIR	1.000	**1.278**	1.000	**1.116**	1.000	**0.928**

Table 2.15: Comparison of the SS-SIR and SIR estimators based on $(1-\bar{R}_1^2)$ and $(1-\bar{R}_2^2)$ under various choices of $(L, \mathfrak{I}_L)$, for model $\mathcal{M}_2^{(2)}$ with $r=2$ and $p=30$.

Criteria ↓	Slices = 10		Slices = 20		Slices = 50	
	SS-SIR	SIR	SS-SIR	SIR	SS-SIR	SIR
First Direction						
$(1-\bar{R}_1^2)$	0.020	0.075	0.016	0.071	0.019	0.081
RE of SS-SIR	1.000	**3.658**	1.000	**4.336**	1.000	**4.308**
Second Direction						
$(1-\bar{R}_2^2)$	0.074	0.149	0.078	0.153	0.115	0.190
RE of SS-SIR	1.000	**2.016**	1.000	**1.968**	1.000	**1.658**

Table 2.16: Comparison of the SS-SIR, SIR and MLE estimators based on Emp. MSE and $(1-\bar{R}_0^2)$ under all models $\mathcal{M}_1^{(b)}$, $\mathcal{M}_2^{(b)}$, $\mathcal{M}_3^{(b)}$ for the binary outcomes with $r=1$ and $p=10$.

Criteria ↓	Model = $\mathcal{M}_1^{(b)}$			Model = $\mathcal{M}_3^{(b)}$			Model = $\mathcal{M}_3^{(b)}$		
	SS-SIR	SIR	MLE	SS-SIR	SIR	MLE	SS-SIR	SIR	MLE
Emp. MSE	0.026	0.043	0.026	0.016	0.035	0.015	0.013	0.032	0.012
RE of SS-SIR	1.000	**1.645**	0.970	1.000	**2.246**	0.950	1.000	**2.434**	0.945
$(1-\bar{R}_0^2)$	0.026	0.043	0.025	0.016	0.035	0.015	0.013	0.032	0.012
RE of SS-SIR	1.000	**1.636**	0.970	1.000	**2.233**	0.950	1.000	**2.420**	0.945

Table 2.17: Comparison of the SS-SIR, SIR and MLE estimators based on Emp. MSE and $(1-\bar{R}_0^2)$ under all models $\mathcal{M}_1^{(b)}$, $\mathcal{M}_2^{(b)}$, $\mathcal{M}_3^{(b)}$ for binary outcomes with $r=1$ and $p=30$.

Criteria ↓	Model = $\mathcal{M}_1^{(b)}$			Model = $\mathcal{M}_3^{(b)}$			Model = $\mathcal{M}_3^{(b)}$		
	SS-SIR	SIR	MLE	SS-SIR	SIR	MLE	SS-SIR	SIR	MLE
Emp. MSE	0.050	0.099	0.046	0.034	0.089	0.028	0.033	0.088	0.025
RE of SS-SIR	1.000	**1.972**	0.907	1.000	**2.604**	0.812	1.000	**2.697**	0.764
$(1-\bar{R}_0^2)$	0.050	0.096	0.045	0.034	0.087	0.028	0.032	0.086	0.025
RE of SS-SIR	1.000	**1.946**	0.908	1.000	**2.566**	0.813	1.000	**2.657**	0.766

Overall, the results in tables 2.6-2.17 are evidently quite satisfactory. The SS-SIR significantly outperforms the SIR, as is expected, under all the settings considered, and w.r.t. all the error measures used, as well as across all the choices of $(L, \mathfrak{I}_L)$ for continuous Y. The efficiency gains of SS-SIR w.r.t. SIR for all the settings with $r=1$, as well as for the first direction in all the settings with $r=2$, are substantially high, and in certain cases, overwhelmingly so. Further, in all the cases with $r=2$, the efficiency gains even for the second direction are in fact significantly high as well, although they may not have been properly reflected in the corresponding RE measures. This also highlights the potential usefulness of SS-SIR in estimating the higher (second, third etc.) directions, for which the SIR estimates are known to often have poor finite sample performances. For continuous Y with $r=1$, it is also worth noting the comparison of SS-SIR and SIR, across all choices of L, to the OLS. While the SIR significantly under-performs w.r.t. the OLS under all the models, the SS-SIR

does so, only slightly, under $\mathcal{M}_1^{(1)}$ when the standard linear model holds and therefore the OLS is the optimal estimator of $\boldsymbol{\beta}_0$. However, for the non-linear models $\mathcal{M}_2^{(1)}$ and $\mathcal{M}_3^{(2)}$, SS-SIR significantly outperforms the OLS in addition to the SIR, for all choices of L. For binary Y with $r = 1$ however, it is interesting to note that while the SIR once again significantly under-performs w.r.t. the MLE in all cases, the latter almost always performs equally well, and in fact slightly better in some cases, compared to the SS-SIR as well, indicating that at least for the settings considered here, it is difficult to beat the performance of the MLE.

Lastly, for continuous Y, the performance of SS-SIR itself seems to be fairly robust to the choice of L as long as it is reasonable (not too small or not too large w.r.t. n), and certainly seems to be more robust than the SIR whose performance degrades significantly for $L = 50$ in most cases. Combining the results over all the cases, a choice of any L in the range of 10 to 20 seems to be quite reasonable for SS-SIR, at least for $n = 500$, and tends to lead to its best performances under all the settings considered here. In general, we believe that for a given n, a choice of L roughly of the order of $n^{\frac{1}{2}}$ should work quite well for SS-SIR. However, the theoretical results provided here are for a fixed $(L, \mathfrak{I}_L)$ regime, and a detailed analysis of SS-SIR under this regime with L diverging is beyond the scope of this paper.

Surrogate Aided Unsupervised Recovery of Sparse Signals in Single Index Models for Binary Outcomes

Abhishek Chakrabortty[1], Matey Neykov Neykov[2], Raymond J. Carroll[3] and Tianxi Cai[1]

[1]Department of Biostatistics

Harvard University

[2]Department of Operations Research and Financial Engineering

Princeton University

[3]Department of Statistics

Texas A&M University

3.1 Summary

We consider the regression of a binary outcome (Y) on a set of (possibly high dimensional) covariates ($\mathbf{X}$) based on a large *unlabeled* data with observations only for $\mathbf{X}$, and additionally, a 'surrogate' (S) which, while not being strongly predictive of Y all throughout its support, can do so with high accuracy when it assumes extreme values. Such data arises naturally in settings where Y, unlike $(\mathbf{X}, S)$, is somewhat difficult or expensive to obtain, a frequent scenario in modern studies involving large databases like electronic medical records (EMR), where an example of Y and S could be some disease outcome of interest and its corresponding diagnostic codes and/or laboratory test results respectively. Assuming Y and S both follow flexible single index models with respect to $\mathbf{X}$, we show that under sparsity assumptions, we can recover the regression parameter of Y versus $\mathbf{X}$ by simply fitting a least squares LASSO estimator to the subset of the observed data restricted to the extreme sets of S with Y imputed using the surrogacy of S. To the best of our knowledge, a problem of this sort has not been considered in the relevant statistical literature, and our associated results are quite novel. We obtain sharp finite sample performance bounds, with several interesting implications, for our estimator, including deterministic deviation bounds and probabilistic guarantees for the bounds to obey satisfactory convergence rates. We demonstrate the effectiveness of our approach through extensive finite sample simulations, followed by application to a real EMR dataset.

3.2 Introduction

Unsupervised classification methods are of great use in a wide variety of scientific applications including image retrieval and processing, document classification in text mining, high dimensional genomic analysis and other problems in biomedical sciences (Ko and Seo, 2000; Merkl and Rauber, 2000; Gllavata et al., 2004; Chen et al., 2005; Henegar et al., 2006). In recent years, many unsupervised statistical and machine learning methods have been pro-

posed to classify categorical outcomes. Examples include clustering, latent class mixture modeling, neural networks and random forest based methods (Merkl and Rauber, 2000; Hofmann, 2001; Shi and Horvath, 2006; Cios et al., 2007; Wei and Kosorok, 2013). Most of the related existing literature however largely focuses on identifying algorithms that can accurately classify the outcomes of interest with less focus on the statistical properties of the estimated model parameters. In this paper, we consider a surrogate aided unsupervised classification problem of a very different and unique nature. Motivated particularly by the problem of automated phenotyping with electronic medical records (EMR) data, we consider a regression modeling approach to unsupervised classification with assistance from surrogate outcomes whose extreme values are highly predictive of the outcome.

Specifically, we consider regressing a binary outcome Y on a set of covariates $\mathbf{X}$ (possibly high dimensional) based on a flexible single index model with an *unknown* link function, so that the regression parameter ($\boldsymbol{\beta}_0$, say) we wish to recover is identifiable only upto scalar multiples. However, the available data, while it is possibly large/massive in size, is completely unlabeled with Y *never* observed, thus leading to an unsupervised set-up. However, while the availability of Y may be limited, it is often possible, especially in datasets like EMR, that one (or more) of the variables, automatically recorded in the database, while not being strongly predictive of Y throughout its support, can do so with high accuracy when it assumes extreme values, thereby serving as a good 'surrogate' (S, say) in its extreme sets. Such data arises naturally in settings where Y, unlike $\mathbf{X}$ and S, is difficult or expensive to obtain, a scenario that is of great practical relevance especially in the modern 'big data' era with massive unlabeled datasets (often electronically recorded) becoming increasingly available and tractable. In particular, they are frequently encountered in modern biomedical studies involving analyses of large databases like EMR, where a typical choice of Y and S could be a disease phenotype like rheumatoid arthritis (RA) and the ICD9 diagnostic codes and/or lab tests for RA respectively. We first briefly discuss the motivating problem of EMR automated phenotyping followed by a brief summary of our contributions in this paper and

the proposed framework for unsupervised learning with extremes of surrogate outcomes.

3.2.1 Automated Phenotyping Using EMR

Endowed with a wealth of de-identified clinical and phenotype data for large patient cohorts, EMR linked with bio-repositories are increasingly gaining popularity as rich resources of data for discovery research (Murphy et al., 2009; Kohane, 2011). Such large scale datasets obtained in a cost-effective and timely manner are of great importance in modern medical research for addressing several questions including the biological role of rare variants and the disease risk profiles of common and rare variants (Kohane, 2011). The availability of detailed patient level phenotypic data from the EMR system linked with a wide range of genomic and biological marker measurements provides a unique opportunity to rigorously study genome-phenome association networks and improve the understanding of disease processes and treatment responses (Wilke et al., 2011; Kohane et al., 2012). For example, as new genetic variants are being increasingly discovered, the scope of their clinical significance can be assessed by examining the range of disease phenotypes that are associated with these variants. Such assessment has been recently proposed using Phenome-wide Association Studies (PheWAS) based on EMR cohorts (Denny et al., 2010). EMR-based cohorts are the key to the success of PheWAS as they contain nearly complete clinical diagnoses for a large group of subjects, broadening the ability to simultaneously test for potential associations between genetic variants and a wide range of disorders, in contrast to traditional prospective cohort studies collecting data only on predetermined outcomes of interest.

However, despite its potential for translational research, one major rate-limiting step in EMR driven PheWAS is the difficulty in extracting accurate phenotype information from the EMR (Bielinski et al., 2011). Since gold standard measurements for the phenotypes typically require manual chart review by physicians which is logistically prohibitive especially for multiple phenotypes, current PheWAS methods primarily rely on ICD9 codes to assess whether a patient has a clinical condition (Denny et al., 2010; Liao et al., 2010). A major limitation

of the ICD9 codes is that they can have highly variable predictive accuracy for identifying many diseases and hence, can introduce substantial noise into the subsequent association studies. For example, based on data from the Partner's Healthcare, among subjects with at least 3 ICD9 rheumatoid arthritis (RA) codes, only 56% of those actually have confirmed RA after manual chart review by a rheumatologist (Liao et al., 2010). However, for subsets of patients where the ICD9 codes assume extreme values (too high or, too low), they can often predict the corresponding phenotype with a high degree of accuracy, thereby serving as an effective surrogate outcome for these subsets. Appropriate and efficient use of such available surrogacy information can prove to be quite useful for creating data-driven labeling methods that can significantly reduce the burden (in time and effort) of manual labeling. In particular for EMR data, such automated phenotyping algorithms can pave the way for high throughput phenotyping (Yu et al., 2015), allowing for phenome-genome association studies that typically require the availability of multiple phenotypes and hence does not scale well with manual labeling methods for obtaining gold standard labels.

3.2.2 Contributions of this Paper: A Brief Summary

With the above motivation and the basic set-up introduced earlier, we now assume the additional availability of observations for such a surrogate S, and also assume that S depends on $\mathbf{X}$ through another single index model with some parameter $\boldsymbol{\alpha}_0$ (say). We now consider the subset of the data restricted to the extreme sets of S (formally characterized through the upper and lower q^{th} quantiles of S with $q \in (0, 1)$ small enough), impute the missing Y deterministically using the surrogacy of S, and try to recover $\boldsymbol{\beta}_0$ through some regression procedure on this data. We demonstrate that while it is, quite sensibly, not possible to recover $\boldsymbol{\beta}_0$ without some further assumptions, it is indeed possible to do so if we additionally assume that $\boldsymbol{\beta}_0$ is sparse (and in fact sparser than $\boldsymbol{\alpha}_0$, along with some other conditions), by simply fitting a least squares LASSO estimator to this restricted (and imputed) data. While the method is clearly quite simple, its success indeed depends on the assumed conditions as

well as appropriate choices of q for creating the restricted data (something that should be typically dictated by domain knowledge), and that of the tuning parameter which should be chosen slightly higher than usual so that sparser solutions like $\boldsymbol{\beta}_0$ are favored.

We obtain explicit finite sample deviation bounds for the performance of our estimator with high probabilistic guarantees (under some conditions on the design distribution) for the bounds to obey satisfactory convergence rates (depending on the choice of q, the extent of the corresponding misclassification error π_q, and the sample size n_q of the restricted data). The results are quite sharp and have several useful implications, including an interesting 'variance-bias' trade-off involving the above three quantities, and for a given order of the misclassification error π_q, the corresponding optimal order of q can also be determined thereof. We also explicitly characterize the behaviour of π_q versus q for one familiar choice of $(Y, \mathbf{X}, S)$, wherein the interplay between $\boldsymbol{\beta}_0$ and $\boldsymbol{\alpha}_0$, and the necessary conditions for our approach to succeed become more explicit. We demonstrate the effectiveness of our approach through extensive finite sample simulations, where its performance (in estimation, prediction or, variable selection) is found to be comparable, if not better in most cases, to that of a supervised estimator based on as many as 500 labels. The sensitivity of our estimator to the choice of q also seems to be fairly robust as long it is chosen reasonably. We also applied our approach to a real dataset obtained from an EMR cohort with promising results, along with validation on a training data that was available to us in this case.

The rest of this paper is organized as follows. In section 3.3, we formally introduce and formulate the problem including all our basic assumptions and notations, followed by exposition of our proposed estimation procedure, including characterization of all the theoretical properties of our estimator. Numerical studies, including extensive simulation results and an application to an EMR dataset are presented in section 3.5, followed by a concluding discussion in section 3.6. Proofs of all the theoretical results are provided in Appendix B.

3.3 Problem Formulation and Proposed Methodology

Let $Y \in \{0,1\}$ denote a binary outcome random variable of interest, $\mathbf{X} \in \mathbb{R}^p$ denote a p dimensional random vector of covariates, and $S \in \mathbb{R}$ denote a 'surrogate' random variable whose role as surrogate would be made more precise shortly. We assume throughout that $(Y, S, \mathbf{X}')'$, defined on a common probability space, has finite 2^{nd} moments. Let $\mathbb{P}(.)$ denote the underlying probability measure characterizing the distribution of $(Y, S, \mathbf{X}')'$, and $\mathbb{E}(.)$ denote expectation with respect to (w.r.t.) $\mathbb{P}(.)$. We next characterize the surrogacy of S.

Role of S as a surrogate: The variable S, in very heuristic terms, satisfies the following property: it is known apriori (typically based on practical experiences and/or domain knowledge) that when S is 'too low' or 'too high', then the corresponding Y is 'very likely' to be 0 or 1 respectively (or the other way around). In order to formalize this notion, we first introduce a few notations as follows.

For any $q \in (0,1]$, let δ_q and $\overline{\delta}_q$ respectively denote the $(q/2)^{th}$ and $(1-q/2)^{th}$ quantiles of the distribution of S. Let $\mathcal{I}_q$ denote the interval: $(-\infty, \delta_q] \cup [\overline{\delta}_q, \infty)$. Let $\mathbb{P}_q(.)$ denote the underlying probability measure characterizing the distribution of $(Y, S, \mathbf{X}')' \mid S \in \mathcal{I}_q$, and let $\mathbb{E}_q(.)$ denote expectation w.r.t. $\mathbb{P}_q(.)$. Hence, for any measurable event A involving $(Y, S, \mathbf{X}')'$, $\mathbb{P}_q(A) = \mathbb{P}(A \mid S \in \mathcal{I}_q)$, and for any measurable and $\mathbb{P}-$integrable function $f(.)$ of $(Y, S, \mathbf{X}')'$, $\mathbb{E}_q\{f(Y, S, \mathbf{X})\} = \mathbb{E}\{f(Y, S, \mathbf{X}) \mid S \in \mathcal{I}_q\}$. Also, let us define: $\pi_q^- = \mathbb{P}(Y = 1 | S \leq \delta_q)$ and $\pi_q^+ = \mathbb{P}(Y = 0 | S \geq \overline{\delta}_q)$. Then the premises of our problem entails that as functions of q, π_q^- and π_q^+ are both small for small enough q. We now formalize this assumption as follows:

Assumption 3.1. *(Surrogacy Assumption)* Conditional on $S \in \mathcal{I}_q$, let us define the *surrogate outcome* Y_q^* as: $Y_q^* \mid S \in \mathcal{I}_q = 1(S \geq \overline{\delta}_q \mid S \in \mathcal{I}_q)$, where $1(.)$ denotes the indicator function. Let $\pi_q = \mathbb{P}_q(Y \neq Y_q^*)$, the *misclassification error* incurred by using Y_q^* as a surrogate for Y, given $S \in \mathcal{I}_q$. Then, we make the following assumption regarding the behavior of π_q:

$$\pi_q \equiv \mathbb{P}_q(Y \neq Y_q^*) = \frac{1}{2}(\pi_q^- + \pi_q^+) \leq Cq^\nu \quad \forall\, q \leq q_0 \in (0,1] \text{ small enough}, \tag{3.1}$$

for some universal constants $\nu > 0$ and $C > 0$, and some $q_0 \in (0, 1]$ small enough.

Note that the choice of the cut-points δ_q and $\overline{\delta}_q$ as precisely being the $(q/2)^{th}$ and $(1 - q/2)^{th}$ quantiles of S is mostly for simplicity, and is not actually necessary. In general, they may correspond to some q_1^{th} and $(1 - q_2)^{th}$ quantiles of S respectively, with q_1 and q_2 being purely determined by the practical considerations of the problem, such that the surrogacy assumption above is satisfied for those choices of q_1, q_2 and $q = (q_1 + q_2)$. Nevertheless, we shall stick to this formulation for the sake of notational convenience. Also, note that without loss of generality (w.lo.g.), we have adopted the convention that the more likely value of Y in the lower tail of S is 0 and that in the upper tail is 1. If the other way round is more likely, then we may simply flip the definition of Y and Y_q^* to $(1 - Y)$ and $(1 - Y_q^*)$ respectively.

Data representation: The data actually observed under the unsupervised setting considered herein can be represented as: $\mathcal{S}_N^* = \{(S_i, \mathbf{X}_i')' : i = 1, \ldots, N\}$ consisting of N independent and identically distributed (i.i.d.) observations from the joint distribution of $(S, \mathbf{X}')'$ only. Let us also define the corresponding ideal sample with the Y observed as: $\mathcal{S}_N = \{(Y_i, S_i, \mathbf{X}_i')' : i = 1, \ldots, N\}$ consisting of N i.i.d. observations from the joint distribution of $(Y, S, \mathbf{X}')'$. For the most part of this paper, our primary focus will be on the subset of $\mathcal{S}_N^*$ (and $\mathcal{S}_N$) consisting of the observations for which the corresponding S satisfies the restriction $S \in \mathcal{I}_q$. We formally define them as follows. For any $q \in (0, 1]$, define:

$$\mathcal{Z}_{n_q}^* = \{(Y_{q,i}^*, S_i, \mathbf{X}_i')' : \ S_i \in \mathcal{I}_q; \ i = 1, \ldots, n_q \equiv Nq\}, \ \text{ and} \tag{3.2}$$

$$\mathcal{Z}_{n_q} = \{(Y_i, S_i, \mathbf{X}_i')' : \ S_i \in \mathcal{I}_q, ; \ i = 1, \ldots, n_q \equiv Nq\}, \tag{3.3}$$

where w.l.o.g., we have re-indexed the observations in both $\mathcal{Z}_{n_q}^*$ and $\mathcal{Z}_{n_q}$ for notational convenience, and also assumed for simplicity that the effective sample size $n_q \equiv Nq$ of $\mathcal{Z}_{n_q}^*$ and $\mathcal{Z}_{n_q}$ is indeed an integer. Note that in the definition of $\mathcal{Z}_{n_q}^*$, we have additionally included the corresponding (deterministic) observations of the surrogate outcome Y_q^* as they would be useful later in the construction of our proposed estimator.

We also wish to point out here that appealing to the potentially abundant availability of unlabeled data, given the practical considerations underlying our problem of interest, the size N of the original sample $\mathcal{S}_N^*$ will be assumed to be massive/substantially large, so that the distribution of $(S, \mathbf{X}')'$ can be presumed to be known for all practical purposes. Consequently, we shall assume for simplicity that the quantiles δ_q and $\overline{\delta}_q$ of S are known since, for each fixed q, they can be near-perfectly estimated from $\mathcal{S}_N^*$ at an ignorable error rate of $O(N^{-1/2})$ with N very large. Another key consequence that the premise of N being very large bears on our approach in this paper is that even for a desirably small enough choice of q, so that the surrogacy assumption is more likely to be satisfied, the sample size n_q of the corresponding restricted datasets $\mathcal{Z}_{n_q}^*$ and $\mathcal{Z}_{n_q}$, which would be of primary interest to us hereafter, is still satisfactorily large enough so as to ensure statistical stability and reasonable convergence rates for estimators constructed based on these datasets. While all results obtained in this paper for our proposed estimators are finite sample results, they are essentially derived keeping in mind the following asymptotic regime: $N \to \infty$, $q = O(N^{-\eta})$ for some constant $\eta \in (0, 1)$, so that $q \to 0$ and $n_q = O\{N^{(1-\eta)}\} \to \infty$, as $N \to \infty$.

Model based assumptions: We assume throughout that the dependence of Y on $\mathbf{X}$ (and S), and that of S on $\mathbf{X}$ are characterized by the following *single index models* (SIMs).

$$Y = f(\boldsymbol{\beta}_0'\mathbf{X}; \epsilon) \quad \text{with } \epsilon \perp\!\!\!\perp (S, \mathbf{X}) \text{ and } f(.) : \mathbb{R}^2 \to \{0, 1\} \textit{ unknown}; \text{ and} \tag{3.4}$$

$$S = g(\boldsymbol{\alpha}_0'\mathbf{X}; \epsilon^*) \quad \text{with } \epsilon^* \perp\!\!\!\perp \mathbf{X}, \epsilon^* \perp\!\!\!\perp \epsilon, \text{ and } g(.) : \mathbb{R}^2 \to \mathcal{X}_S \textit{ unknown}; \tag{3.5}$$

where $\mathcal{X}_S \subseteq \mathbb{R}$ denotes the appropriate support of S, and $\boldsymbol{\beta}_0, \boldsymbol{\alpha}_0 \in \mathbb{R}^p$ are some unknown parameter vectors of interest characterizing the respective models, while ϵ, ϵ^* represent the corresponding random noise components. Note that since the 'link' functions $f(.)$ and $g(.)$ are allowed to be completely unspecified (upto basic measurability and moment based restrictions to comply with our starting assumptions), the corresponding regression parameters $\boldsymbol{\beta}_0$ and $\boldsymbol{\alpha}_0$ are identifiable *only* upto scalar multiples (or in other words, only the span or 'direction'

of $\boldsymbol{\beta}_0$ and $\boldsymbol{\alpha}_0$ are identifiable). Both models (3.4) and (3.5) are highly flexible, and yet easily interpretable, semi-parametric models, and as special cases, they include all commonly used parametric models like standard generalized linear models (glms) with known link functions. For (3.4), it might be more helpful to view the function $f(.)$ as some sort of a 'thresholding' function given by: $f(a;b) = 1\{\bar{f}(a;b) > 0\} \ \forall \ a, b \in \mathbb{R}$, for some $\bar{f}(.;.) : \mathbb{R}^2 \to \mathbb{R}$. For instance, with $f(a,b) = 1[h(a) + b > 0] \ \forall \ a, b \in \mathbb{R}$ for some $h(.) : \mathbb{R} \to \mathbb{R}$, and with $\epsilon \sim$ Logistic $(0,1)$ or Normal $(0,1)$ distributions, the model (3.4) corresponds to the logistic or probit regression models respectively, including linear as well as non-linear functional forms of $\boldsymbol{\beta}_0'\mathbf{X}$ through appropriate choices of $h(.)$. Similarly, for a continuous S, with $g(a;b) = \{g^*(a) + b\} \ \forall \ a, b \in \mathbb{R}$ for some $g^*(.) : \mathbb{R} \to \mathbb{R}$, and with $\epsilon^* \sim$ Normal $(0, \sigma^2)$ for some $\sigma \geq 0$, the model (3.5), through appropriate choices of $g^*(.)$, includes the standard linear model as well as other non-standard models involving non-linear functional forms of $\boldsymbol{\alpha}_0'\mathbf{X}$.

The assumed models (3.4) and (3.5) also imply that $(Y \perp\!\!\!\perp \mathbf{X}) \,|\, \boldsymbol{\beta}_0'\mathbf{X}$ and $(S \perp\!\!\!\perp \mathbf{X}) \,|\, \boldsymbol{\alpha}_0'\mathbf{X}$, so that $\boldsymbol{\beta}_0'\mathbf{X}$ and $\boldsymbol{\alpha}_0'\mathbf{X}$ are the sufficient statistics summarizing the dependencies of Y and S on $\mathbf{X}$ respectively . Moreover, (3.4) also implies that $(Y \perp\!\!\!\perp S) \,|\, \mathbf{X}$, so that Y relates to S only through $\mathbf{X}$ (and hence, S behaves as a so-called 'instrumental variable', a term frequently used in the causal inference literature). Note that this condition does *not* contradict in any way the surrogacy assumption (3.1) which only relates Y and S *marginally* (involving no conditioning w.r.t. $\mathbf{X}$), and that too only when S assumes extreme values (so that S need not be a strong predictor of Y throughout its entire support).

It is also worth noting that the condition $(Y \perp\!\!\!\perp S) \,|\, \mathbf{X}$ is very different (and not necessarily stronger or weaker) from the typical surrogacy assumption (also known as the 'Naive Bayes' assumption) that is frequently made in the measurement error literature, wherein the surrogate and the covariates are assumed to be independent conditional on the true outcome. However, for the problems generally considered in those literatures, a proper surrogate outcome (typically, a systematically noisy version of the true outcome) is *actually* observed for *all* individuals, which is quite unlike the setting considered herein. In our case, over the

entire observed data S_N^*, the subset $\mathcal{Z}_{n_q}^*$ in (3.2), for some appropriate choice of q, provides us with the only (and albeit approximate) access to the corresponding true Y. Our primary goal however is to still recover (upto a scalar multiple) the regression parameter $\boldsymbol{\beta}_0$ in (3.4) given the setting we are provided. It is therefore only reasonable, and almost necessary, to have $(Y \perp\!\!\!\perp S) \mid \mathbf{X}$, in order to make proper sense out of the problem at hand and to have any hope of recovering $\boldsymbol{\beta}_0$ based on $\mathcal{Z}_{n_q}^*$ (or even $\mathcal{Z}_{n_q}$, for that matter), as it ensures that the restriction $S \in \mathcal{I}_q$ underlying the construction of $\mathcal{Z}_{n_q}^*$ and $\mathcal{Z}_{n_q}$ does not alter the relation between Y and $\mathbf{X}$ in (3.4) that defines our parameter of interest $\boldsymbol{\beta}_0$. We next discuss some useful motivations and essential fundamentals underlying our approach for recovering $\boldsymbol{\beta}_0$.

Basic foundations of our approach and some fundamental results: We begin with a few notations. For a given choice of $q \in (0, 1]$, let $p_q = \mathbb{E}_q(Y)$, $p_q^* = \mathbb{E}_q(Y_q^*) = 1/2$ (in our case), $\boldsymbol{\mu}_q = \mathbb{E}_q(\mathbf{X})$, and $\Sigma_q = \text{Var}\ (\mathbf{X} \mid S \in \mathcal{I}_q)$ which is further assumed to be a positive definite (p.d.) matrix. For any given $q \in (0, 1]$, let us now define:

$$\mathbb{L}_q(\mathbf{v}) \equiv \mathbb{E}_q[\{Y - p_q - \mathbf{v}'(\mathbf{X} - \boldsymbol{\mu}_q)\}^2] \ \ \forall\ \mathbf{v} \in \mathbb{R}^p, \quad \overline{\boldsymbol{\beta}}_q = \underset{\mathbf{v} \in \mathbb{R}^p}{\arg\min}\, \mathbb{L}_q(\mathbf{v}); \tag{3.6}$$

$$\mathbb{L}_q^*(\mathbf{v}) \equiv \mathbb{E}_q[\{Y_q^* - p_q^* - \mathbf{v}'(\mathbf{X} - \boldsymbol{\mu}_q)\}^2] \ \ \forall\ \mathbf{v} \in \mathbb{R}^p, \quad \overline{\boldsymbol{\alpha}}_q = \underset{\mathbf{v} \in \mathbb{R}^p}{\arg\min}\, \mathbb{L}_q^*(\mathbf{v}). \tag{3.7}$$

With Σ_q assumed to be p.d., both $\overline{\boldsymbol{\beta}}_q$ and $\overline{\boldsymbol{\alpha}}_q$ are well-defined and unique, and further denote the population target parameters corresponding to the least squares regression of Y and Y_q^* on $\mathbf{X}$ respectively, given $S \in \mathcal{I}_q$. Also, note that since we would be only interested in the slope vector of regression coefficients corresponding to $\mathbf{X}$, we have conveniently removed the need for any intercept parameter by appropriately centering all concerned variables in the definitions of $\mathbb{L}_q(.)$ and $\mathbb{L}_q^*(.)$. Finally, while we have considered the squared loss above, which admittedly is often not the preferred choice of a loss function for binary outcomes, other convex loss functions more suited for binary outcomes, like the logistic loss, can also be similarly considered. However, we do believe that the squared loss is a somewhat safer and more convenient choice in this case, since the inherent restriction $S \in \mathcal{I}_q$ underlying

our setting can correspond to highly flat regions in the curves of these other choices of loss functions, and their minimization can potentially lead to non-identifiability/perfect classification issues on a population scale, as well as finite sample stability/convergence issues. The squared loss on the other hand does not have any such identifiability issues as long as Σ_q is p.d. It is also worth noting that the target parameter (the slope vector of regression coefficients) corresponding to the least squares reqression of any binary outcome on a set of covariates is well known to have the simple and clear interpretation of being proportional to the corresponding LDA (linear discriminant analysis) direction (Hastie et al., 2008).

The main motivation behind our consideration of (3.6) and (3.7) lies in a remarkable and interesting result from Li and Duan (1989), where they show that for any generic outcome $\mathbb{Y}$ satisfying a SIM (with some parameter $\boldsymbol{\gamma} \in \mathbb{R}^p$) w.r.t. some generic set of covariates $\mathbb{X} \in \mathbb{R}^p$, if the following two conditions hold: (i) the underlying design distribution of $\mathbb{X}$ satisfies a certain 'linearity of expectation condition' given by: $\mathbb{E}(\mathbf{v}'\mathbb{X} \mid \boldsymbol{\gamma}'\mathbb{X})$ is a linear function of $\boldsymbol{\gamma}'\mathbb{X} \;\; \forall\, \mathbf{v} \in \mathbb{R}^p$, and (ii) for a loss function $L(\mathbb{Y}; a+\mathbf{v}'\mathbb{X})$ that is convex in the second argument (like the squared loss, for instance), if the minimizer $(\overline{a}, \overline{\mathbf{v}})$ of $\mathbb{E}\{L(\mathbb{Y}; a + \mathbf{v}'\mathbb{X})\}$ over $a \in \mathbb{R}$, $\mathbf{v} \in \mathbb{R}^p$ exists and is unique, then $\overline{\mathbf{v}} \propto \boldsymbol{\gamma}$, so that $\overline{\mathbf{v}}$ recovers the SIM parameter $\boldsymbol{\gamma}$ upto a scalar multiple. As noted earlier, with $(Y \perp\!\!\!\perp S) \mid \mathbf{X}$, the SIM (3.4) continues to hold with the same parameter $\boldsymbol{\beta}_0$ even under the restriction $S \in \mathcal{I}_q$ that underlines the construction of $\mathcal{Z}^*_{n_q}$ and $\mathcal{Z}_{n_q}$. Hence, given this fact, we have considered in (3.6) the expected squared loss for Y and its corresponding minimizer with the hope that if a result of a similar flavor as that obtained by Li and Duan (1989) can be shown to hold for $\overline{\boldsymbol{\beta}}_q$, then at least if the hypothetical data $\mathcal{Z}_{n_q}$ were actually observed, a minimization of the corresponding empirical squared loss for Y based on $\mathcal{Z}_{n_q}$ would lead to a consistent estimator of the $\boldsymbol{\beta}_0$ direction.

Of course, the major issue is that we only observe $\mathcal{Z}^*_{n_q}$ in practice, and therefore can only hope to minimize the empirical squared loss for Y^*_q based on $\mathcal{Z}^*_{n_q}$, which would be a consistent estimator of the minimizer of the corresponding expected squared loss for Y^*_q, as considered in (3.7). However, owing to the surrogacy assumption, $\mathbb{L}_q(.)$ and $\mathbb{L}^*_q(.)$ should be

pointwise quite close to each other, and hence owing to their smoothness and convexity, it is not unreasonable to expect that their minimizers $\overline{\boldsymbol{\alpha}}_q$ and $\overline{\boldsymbol{\beta}}_q$ should also be close. All these questions and intuitions highlight the necessity of a better understanding of the behavior of $\overline{\boldsymbol{\alpha}}_q$ and $\overline{\boldsymbol{\beta}}_q$, their dependencies on $\boldsymbol{\beta}_0$ and $\boldsymbol{\alpha}_0$, as well as their relation to each other. Lastly, another technical, but no less important, issue is the validity of the 'linearity of expectation condition', which can be quite tricky in our case since our essential design distribution of interest is that of $(\mathbf{X} \mid S \in \mathcal{I}_q)$ and not of $\mathbf{X}$. In standard SIM problems, the design distribution is often assumed to be elliptically symmetric (e.g. the normal distribution) for which the condition is known to hold. However, such an assumption seems overly unrealistic for $(\mathbf{X} \mid S \in \mathcal{I}_q)$, and even for the most familiar case of a normally distributed $\mathbf{X}$, it does not seem to hold apart from some trivial scenarios. We would therefore assume a different kind of a 'linearity condition' that is more reasonable and likely to hold in practice for a fairly wide class of distributions. The next assumption is in this regard, and the subsequent result based on it aims to characterize $\overline{\boldsymbol{\beta}}_q$ and $\overline{\boldsymbol{\alpha}}_q$ more explicitly, as well as provide answers to most of the questions raised herein, some of which would be albeit contradicting our intuitions, but nonetheless would be useful for the exposition and understanding of our final approach.

Assumption 3.2. *(Design Linearity Conditions)* We assume that the *marginal* distribution of $\mathbf{X}$ satisfies: for any $\mathbf{v} \in \mathbb{R}^p$, $\mathbb{E}(\mathbf{v}'\mathbf{X} \mid \boldsymbol{\alpha}_0'\mathbf{X}, \boldsymbol{\beta}_0'\mathbf{X})$ is a linear function of $\boldsymbol{\alpha}_0'\mathbf{X}$ and $\boldsymbol{\beta}_0'\mathbf{X}$, i.e.

$$\mathbb{E}(\mathbf{v}'\mathbf{X} \mid \boldsymbol{\alpha}_0'\mathbf{X}, \boldsymbol{\beta}_0'\mathbf{X}) = c_{\mathbf{v}} + a_{\mathbf{v}}(\boldsymbol{\alpha}_0'\mathbf{X}) + b_{\mathbf{v}}(\boldsymbol{\beta}_0'\mathbf{X}) \ \ \forall\, \mathbf{v} \ \in \mathbb{R}^p, \quad \text{and} \tag{3.8}$$

$$\mathbb{E}(\boldsymbol{\beta}_0'\mathbf{X} \mid \boldsymbol{\alpha}_0'\mathbf{X}) = \overline{c} + \overline{a}(\boldsymbol{\alpha}_0'\mathbf{X}), \tag{3.9}$$

for some real constants $c_{\mathbf{v}}, a_{\mathbf{v}}$ and $b_{\mathbf{v}}$ all depending on $\mathbf{v}$, and some real constants $\overline{c}$ and $\overline{a}$.

The constants $c_{\mathbf{v}}$, $a_{\mathbf{v}}$, and $b_{\mathbf{v}}$ can be evaluated explicitly as the regression coefficients obtained from the least squares regression of $\mathbf{v}'\mathbf{X}$ on $\boldsymbol{\alpha}_0'\mathbf{X}$ and $\boldsymbol{\beta}_0'\mathbf{X}$, and are all well-defined and unique as long as $\mathrm{Var}\{(\boldsymbol{\alpha}_0'\mathbf{X}, \boldsymbol{\beta}_0'\mathbf{X})'\}$ is p.d. Similarly, $\overline{c}$ and $\overline{a}$ can be obtained as the regression coefficients from the least squares regression of $\boldsymbol{\beta}_0'\mathbf{X}$ on $\boldsymbol{\alpha}_0'\mathbf{X}$. Note that (3.8) and

(3.9) are restrictions on $\mathbf{X}$ only (and hence, called 'design' conditions), and do not involve Y (or Y_q^*). Moreover, they *only* involve the unconditional marginal distribution of $\mathbf{X}$ and *not* the distribution of $\{\mathbf{X} \,|\, S \in \mathcal{I}_q\}$, which however is actually our underlying design distribution of interest. The condition (3.8) is slightly stronger than the typical linearity conditions that have been usually assumed (Li and Duan, 1989) in the SIM literature, in the sense that it requires a joint linearity in $\boldsymbol{\alpha}_0'\mathbf{X}$ and $\boldsymbol{\beta}_0'\mathbf{X}$, instead of only $\boldsymbol{\beta}_0'\mathbf{X}$. The primary reason for its necessity is the fact that the underlying design distribution $\{\mathbf{X}|S \in \mathcal{I}_q\}$ inherently depends on $\boldsymbol{\alpha}_0'\mathbf{X}$ through S owing to (3.5). Nevertheless, both conditions (3.8) and (3.9) are still satisfied by all elliptically symmetric distributions, including the normal distribution. Moreover, Hall and Li (1993) have also argued that for a wide class of distributions satisfying some mild restrictions, such linearity conditions are 'approximately true' with high probability for most directions $\mathbf{v} \in \mathbb{R}^p$, as long as p is large enough. We now present our first main result that provides a useful characterization of $\overline{\boldsymbol{\beta}}_q$ and $\overline{\boldsymbol{\alpha}}_q$ in terms of $\boldsymbol{\beta}_0$ and $\boldsymbol{\alpha}_0$.

Theorem 3.1. *Let $\overline{\boldsymbol{\beta}}_q$ and $\overline{\boldsymbol{\alpha}}_q$ be as defined in (3.6) and (3.7) respectively. Assume the design linearity condition (3.8) holds from assumption 3.2. Then, we have:*

$$\overline{\boldsymbol{\beta}}_q = a_q \boldsymbol{\alpha}_0 + b_q \boldsymbol{\beta}_0 \quad \textit{and} \tag{3.10}$$

$$\overline{\boldsymbol{\alpha}}_q = a_q^* \boldsymbol{\alpha}_0, \tag{3.11}$$

for some real constants a_q, b_q and a_q^* all depending on q. The constants (a_q, b_q) are explicitly given by: $(a_{\mathbf{v}}, b_{\mathbf{v}})$ respectively from (3.8) with $\mathbf{v} = \overline{\boldsymbol{\beta}}_q$. The constant a_q^* is given by: $(a_{\mathbf{v}} + b_{\mathbf{v}}\overline{a})$ with $\mathbf{v} = \overline{\boldsymbol{\alpha}}_q$, where $\overline{a}$ is from (3.9), and $(a_{\mathbf{v}}, b_{\mathbf{v}})$ are from (3.8) .

Theorem 3.1 establishes the explicit relationships of $\overline{\boldsymbol{\beta}}_q$ and $\overline{\boldsymbol{\alpha}}_q$, the target parameters for the least squares regression of $(Y \mid S \in \mathcal{I}_q)$ and $(Y_q^* \mid S \in \mathcal{I}_q)$ respectively on $(\mathbf{X} \mid S \in \mathcal{I}_q)$, to the original SIM parameters $\boldsymbol{\beta}_0$ and $\boldsymbol{\alpha}_0$. (3.11) essentially shows that $\overline{\boldsymbol{\alpha}}_q \propto \boldsymbol{\alpha}_0$ and therefore, a simple minimization of the empirical squared loss for Y_q^* based on $\mathcal{Z}_{n_q}^*$ would lead to an estimator that can only consistently recover the direction of $\boldsymbol{\alpha}_0$, and *not* $\boldsymbol{\beta}_0$. This finding,

while somewhat contradictory to our original intuitions, does make sense, since due to (3.5), $(Y_q^* \,|\, S \in \mathcal{I}_q)$ and $(\mathbf{X} \,|\, S \in \mathcal{I}_q)$ are after all directly related through a SIM with parameter $\boldsymbol{\alpha}_0$, and further, the design distribution of $(\mathbf{X} \,|\, S \in \mathcal{I}_q)$ *does* satisfy the conventional linearity of expectation condition typically assumed in the SIM literature, i.e. $\mathbb{E}_q(\mathbf{v}'\mathbf{x} \,|\, \boldsymbol{\alpha}_0'\mathbf{X})$ is indeed linear in $\boldsymbol{\alpha}_0'\mathbf{X}$ owing to (3.5) and (3.8)-(3.9). Hence, (3.11) is really a consequence of standard results that are well known in the SIM literature.

On the other hand, (3.10) shows, rather surprisingly, that $\overline{\boldsymbol{\beta}}_q$ lies in the (two-dimensional) span of $\boldsymbol{\alpha}_0$ and $\boldsymbol{\beta}_0$, and therefore, *even if* $(Y \,|\, S \in \mathcal{I}_q)$ were actually observed, the estimator obtained from a simple minimization of the empirical squared loss for Y based on $\mathcal{Z}_{n_q}$ would only be able to consistently recover the direction of its target parameter $\overline{\boldsymbol{\beta}}_q$ that does *not* lie along the $\boldsymbol{\beta}_0$ direction itself. Note that this result is in stark contrast with standard results from the SIM literature, where it would have been expected to lie along the $\boldsymbol{\beta}_0$ direction. This distinction is largely due to the dependence on $\boldsymbol{\alpha}_0$ of the restriction $S \in \mathcal{I}_q$ that characterizes the underlying design distribution, and also makes the conventional linearity condition, requiring $\mathbb{E}_q(\mathbf{v}'\mathbf{x} \,|\, \boldsymbol{\beta}_0'\mathbf{X})$ to be linear in $\boldsymbol{\beta}_0'\mathbf{X}$, unlikely to hold.

Overall, theorem 3.1 clearly shows that some further assumptions are perhaps needed regarding the structure of $\boldsymbol{\beta}_0$ (and possibly $\boldsymbol{\alpha}_0$ as well) in order to have any hope of recovering $\boldsymbol{\beta}_0$ based on $\mathcal{Z}_{n_q}^*$ (or even $\mathcal{Z}_{n_q}$). In this regard, we demonstrate that if $\boldsymbol{\beta}_0$ additionally satisfies some sparsity assumptions, then it is indeed possible to recover $\boldsymbol{\beta}_0$ from $\mathcal{Z}_{n_q}^*$ based on a L_1-penalized least squares regression of Y_q^* on $\mathbf{X}$. The L_1 penalty introduced in the minimization favours solutions with a sparse structure, as possessed by $\boldsymbol{\beta}_0$, and therefore provides the chance to 'push' the solution of the un-penalized regression (that provably recovers the $\boldsymbol{\alpha}_0$ direction) to a sparser solution possibly targeting the $\boldsymbol{\beta}_0$ direction. It is also worth noting that the sparsity assumption, one of the most popular and interpretable structure based assumptions that are used in the statistical literature (especially in high dimensional statistics, although high dimensionality is not the reason for introducing this assumption in our case), is a scale invariant criteria, and therefore, it fits well into our setting

where $\boldsymbol{\beta}_0$ and $\boldsymbol{\alpha}_0$ are identifiable only upto scalar multiples. We next discuss the details of our proposed approach and the associated estimator.

3.3.1 The Unsupervised LASSO (ULASSO) Estimator

For convenience of further discussion, we first introduce some notations that will be used throughout the rest of this paper. For any $\mathbf{v} \in \mathbb{R}^p$, let $\mathbf{v}_{[j]}$ denote its j^{th} coordinate $\forall\ j \in \{1, \ldots, p\}$. Let $\|v\|_1 = \sum_{j=1}^p |\mathbf{v}_{[j]}|$, $\|\mathbf{v}\|_2 = (\sum_{j=1}^p \mathbf{v}_{[j]}^2)^{1/2}$, and $\|\mathbf{v}\|_\infty = \max\limits_{1 \le j \le p} \|\mathbf{v}_{[j]}\|$ respectively denote the L_1, L_2 and L_∞ norms of any $\mathbf{v} \in \mathbb{R}^p$. Let $\mathcal{A}(\mathbf{v}) = \{j : \mathbf{v}_{[j]} \neq 0\} \subseteq \{1, \ldots, p\}$ denote the support of $\mathbf{v}$, and $s_{\mathbf{v}} = \#\{j : \mathbf{v}_{[j]} \neq 0\}$ denote the size of $\mathcal{A}(\mathbf{v})$ for any $\mathbf{v} \in \mathbb{R}^p$. For any $\mathcal{J} \subseteq \{1, \ldots, p\}$ and any $\mathbf{u} \in \mathbb{R}^p$, let $\Pi_{\mathcal{J}}(\mathbf{u}) \in \mathbb{R}^p$ denote the restriction of $\mathbf{u}$ onto $\mathcal{J}$ i.e. $\{\Pi_{\mathcal{J}}(\mathbf{u})\}_{[j]} = \mathbf{u}_{[j]} 1\{j \in \mathcal{J}\}\ \forall\ j \in \{1, \ldots, p\}$, let $\mathcal{J}^c = \{1, \ldots, p\} \backslash \mathcal{J}$, let $\mathcal{M}_{\mathcal{J}} \subseteq \mathbb{R}^p$ denote the subspace: $\{\mathbf{u} \in \mathbb{R}^p : \mathcal{A}(\mathbf{u}) \subseteq \mathcal{J}\}$, and let $\mathcal{M}_{\mathcal{J}}^{\perp} = \{\mathbf{u} \in \mathbb{R}^p : \mathcal{A}(\mathbf{u}) \subseteq \mathcal{J}^c\} \subseteq \mathbb{R}^p$ denote the orthogonal complement (w.r.t. the L_2 inner product) of $\mathcal{M}_{\mathcal{J}}$. For the choices of $\mathcal{J}$ given by: $\mathcal{J} = \mathcal{A}(\mathbf{v})$ or $\mathcal{J} = \mathcal{A}^c(\mathbf{v})$ for some $\mathbf{v} \in \mathbb{R}^p$, we shall use the shorthand $\Pi_{\mathbf{v}}(.)$ and $\Pi_{\mathbf{v}}^c(.)$ to denote $\Pi_{\mathcal{A}(\mathbf{v})}(.)$, and $\Pi_{\mathcal{A}^c(\mathbf{v})}(.)$ respectively. Lastly, let $P_{\mathcal{J}}(\mathbf{v})$ and $P_{\mathcal{J}}^{\perp}(\mathbf{v})$ respectively denote the orthogonal projections of any $\mathbf{v} \in \mathbb{R}^p$ onto $\mathcal{M}_{\mathcal{J}}$ and $\mathcal{M}_{\mathcal{J}}^{\perp}$, for any $\mathcal{J}$ as above. Note that: $P_{\mathcal{A}(\boldsymbol{\beta}_0)}^{\perp}(\overline{\boldsymbol{\beta}}_q) = \Pi_{\mathcal{A}^c(\boldsymbol{\beta}_0)}(\overline{\boldsymbol{\beta}}_q) \equiv \Pi_{\boldsymbol{\beta}_0}^c(\overline{\boldsymbol{\beta}}_q) = \Pi_{\boldsymbol{\beta}_0}^c(a_q \boldsymbol{\alpha}_0)$. These relations would be used later on.

Let $\overline{\mathbf{X}}_{n_q} = n_q^{-1} \sum_{i=1}^{n_q} \mathbf{X}_i$, $\overline{Y}^*_{n_q} = n_q^{-1} \sum_{i=1}^{n_q} Y^*_{q,i}$ respectively denote the sample means of $\mathbf{X}$ and Y_q^* in $\mathcal{Z}^*_{n_q}$, and let $\widehat{\Sigma}_q = {n_q}^{-1} \sum_{i=1}^{n_q} (\mathbf{X}_i - \overline{\mathbf{X}}_{n_q})(\mathbf{X}_i - \overline{\mathbf{X}}_{n_q})'$ denote the corresponding sample covariance matrix of $\mathbf{X}$. We next define the empirical squared loss between Y_q^* and $\mathbf{X}$ based on $\mathcal{Z}^*_{n_q}$, and some related quantities of interest. For any $\boldsymbol{\beta}, \mathbf{v} \in \mathbb{R}^p$, let us define:

$$\mathcal{L}_{n_q}(\mathcal{Z}^*_{n_q}; \boldsymbol{\beta}) = \frac{1}{n_q} \sum_{i=1}^{n_q} \{(Y^*_{q,i} - \overline{Y}^*_{n_q}) - \boldsymbol{\beta}'(\mathbf{X}_i - \overline{\mathbf{X}}_{n_q})\}^2, \tag{3.12}$$

$$\boldsymbol{\nabla}\{\mathcal{L}_{n_q}(\mathbb{Z}^*_{n_q}; \boldsymbol{\beta})\} = \frac{\partial}{\partial \boldsymbol{\beta}} \mathcal{L}_{n_q}(\mathbb{Z}^*_{n_q}; \boldsymbol{\beta}) = -2\, T_{n_q}(\boldsymbol{\beta}), \text{ where} \tag{3.13}$$

$$T_{n_q}(\boldsymbol{\beta}) = \frac{1}{n_q} \sum_{i=1}^{n_q} \{(Y^*_{q,i} - \overline{Y}^*_{n_q}) - \boldsymbol{\beta}'(\mathbf{X}_i - \overline{\mathbf{X}}_{n_q})\}(\mathbf{X}_i - \overline{\mathbf{X}}_{n_q}), \quad \text{and} \tag{3.14}$$

$$\delta\{\mathcal{L}_{n_q}(\mathcal{Z}^*_{n_q}; \boldsymbol{\beta}; \mathbf{v})\} = \mathcal{L}_{n_q}(\mathcal{Z}^*_{n_q}; \boldsymbol{\beta} + \mathbf{v}) - \mathcal{L}_{n_q}(\mathcal{Z}^*_{n_q}; \boldsymbol{\beta}) - \mathbf{v}' \boldsymbol{\nabla}\{\mathcal{L}_{n_q}(\mathcal{Z}^*_{n_q}; \boldsymbol{\beta})\}. \tag{3.15}$$

Since we are only interested in the slope vector of regression coefficients w.r.t. $\mathbf{X}$, we have appropriately centered all the concerned variables in (3.12). We now consider the L_1 penalized (convex) minimization of (3.12). Assuming that $\boldsymbol{\beta}_0$ is indeed sparse, we propose to estimate the $\boldsymbol{\beta}_0$ direction based on the following estimator:

$$\widehat{\boldsymbol{\beta}}_{n_q}(\lambda) \equiv \widehat{\boldsymbol{\beta}}_{n_q}(\lambda; \mathcal{Z}^*_{n_q}) = \underset{\boldsymbol{\beta} \in \mathbb{R}^p}{\arg\min} \left\{ \mathcal{L}_{n_q}(\mathcal{Z}^*_{n_q}; \boldsymbol{\beta}) + \lambda \|\boldsymbol{\beta}\|_1 \right\}, \tag{3.16}$$

where $\lambda \geq 0$ denotes the regularization/tuning parameter controlling the effect of the penalty term. $\widehat{\boldsymbol{\beta}}_{n_q}(\lambda)$ is simply the LASSO estimator corresponding to the least squares regression of Y^*_q on $\mathbf{X}$ based on $\mathcal{Z}^*_{n_q}$, (and can also be viewed as a sparse LDA estimator, owing to the correspondence between LDA and least squares regression for binary outcomes). We shall call this estimator the *Unsupervised LASSO* (ULASSO) estimator. Note that while we have used the standard L_1 norm as the choice of our penalty term, other sparsity friendly penalties such as weighted versions of the L_1 norm (that includes the Adaptive LASSO penalty as a special case) can also be considered. However, we prefer to stick to the formulation in (3.16) for the sake of simplicity. It also needs to be noted that since the LASSO estimator is known to be sensitive to the scaling of the covariates, it might be helpful in practice to standardize the covariates (w.r.t. the underlying design distribution of $\mathbf{X} \mid S \in \mathcal{I}_q$) prior to implementing the estimation procedure. We shall next study the finite sample properties of $\widehat{\boldsymbol{\beta}}_{n_q}(\lambda)$ first in terms of deterministic deviation bounds, followed by probabilistic bounds regarding performance guarantees and convergence rates.

Deterministic deviation bounds: For convenience of further discussion, we introduce a few notations and definitions, followed by stating our required assumptions. Recalling the definition of $T_{n_q}(\boldsymbol{\beta})$ from (3.14), and letting $\mathbb{T}_{n_q} = T_{n_q}(\overline{\boldsymbol{\beta}}_q)$, it is straightforward to note that:

$$\mathbb{T}_{n_q} \equiv T_{n_q}(\overline{\boldsymbol{\beta}}_q) = \mathbb{T}^{(1)}_{n_q} + \mathbb{T}^{(2)}_{n_q,1} - \mathbb{T}^{(2)}_{n_q,2}, \quad \text{where} \tag{3.17}$$

$$\mathbb{T}^{(1)}_{n_q} = \frac{1}{n_q}\sum_{i=1}^{n_q}(Y^*_{q,i} - Y_i)(\mathbf{X}_i - \overline{\mathbf{X}}_q), \tag{3.18}$$

$$\mathbb{T}^{(2)}_{n_q,1} = \frac{1}{n_q}\sum_{i=1}^{n_q}\{Y_i - p_q - \overline{\boldsymbol{\beta}}'_q(\mathbf{X}_i - \boldsymbol{\mu}_q)\}(\mathbf{X}_i - \boldsymbol{\mu}_q), \quad \text{and} \tag{3.19}$$

$$\mathbb{T}^{(2)}_{n_q,2} = \frac{1}{n_q}\sum_{i=1}^{n_q}\{Y_i - p_q - \overline{\boldsymbol{\beta}}'_q(\mathbf{X}_i - \boldsymbol{\mu}_q)\}(\overline{\mathbf{X}}_q - \boldsymbol{\mu}_q), \tag{3.20}$$

where, throughout in (3.18)-(3.20), $\{Y_1, \ldots, Y_{n_q}\}$ denotes the corresponding unobserved true outcomes from $\mathcal{Z}_{n_q}$, as defined in (3.3). We next state the first of our two assumptions required for obtaining the finite sample deviation bounds of $\widehat{\boldsymbol{\beta}}_{n_q}(\lambda)$.

Assumption 3.3. *(Restricted Strong Convexity)* With $\delta\{\mathcal{L}_{n_q}(\mathcal{Z}^*_{n_q}; \boldsymbol{\beta}; \mathbf{v})\}$ as defined in (3.15), we assume that at $\boldsymbol{\beta} = \overline{\boldsymbol{\beta}}_q$, the loss function $\mathcal{L}_{n_q}(\mathcal{Z}^*_{n_q}; \boldsymbol{\beta})$ satisfies a restricted strong convexity property as follows: $\exists$ a (non-random) constant $\kappa_q > 0$, depending on q, such that

$$\delta\{\mathcal{L}_{n_q}(\mathcal{Z}^*_{n_q}; \overline{\boldsymbol{\beta}}_q; \mathbf{v})\} \equiv \mathbf{v}'\widehat{\Sigma}_q\mathbf{v} \;\leq\; \kappa_q\|\mathbf{v}\|_2^2 \quad \forall\, \mathbf{v} \in \mathbb{C}(\boldsymbol{\beta}_0; \overline{\boldsymbol{\beta}}_q), \text{ where} \tag{3.21}$$

$$\mathbb{C}(\boldsymbol{\beta}_0; \overline{\boldsymbol{\beta}}_q) = \{\mathbf{v} \in \mathbb{R}^p : \|\Pi^c_{\boldsymbol{\beta}_0}(\mathbf{v})\|_1 \leq 3\|\Pi_{\boldsymbol{\beta}_0}(\mathbf{v})\|_1 + 4\|\Pi^c_{\boldsymbol{\beta}_0}(\overline{\boldsymbol{\beta}}_q)\|_1\} \subseteq \mathbb{R}^p.$$

Assumption 3.3, largely adopted from Negahban et al. (2012), is one of the several restricted eigenvalue type assumptions that are standard in the high dimensional statistics literature. While we have assumed (3.21) to hold deterministically for any realization of $\mathcal{Z}^*_{n_q}$, it only needs to hold almost surely (a.s.) w.r.t. $\mathbb{P}_q$ for some constant κ_q. It can also be generalized further, wherein it only needs to hold with high probability, in which case that corresponding probability needs to be factored into all our subsequent probabilistic bounds. A somewhat simpler, yet stronger, sufficient condition that ensures (3.21) is that the minimum eigenvalue of $\widehat{\Sigma}_q$ is bounded below a.s. $[\mathbb{P}_q]$ by some constant $\kappa_q > 0$. Even this can be substantially weakened if Σ has a strictly positive minimum eigenvalue, and the underlying design distribution $(\mathbf{X}\,|\,S \in \mathcal{I}_q)$ is sufficiently well behaved with nice concentration properties (e.g. sub-gaussian distributions), in which case, using random matrix theory, (3.21) can be shown to hold for some appropriate choice of κ_q with overwhelming probability. We shall

however stick to the formulation in (3.21) for simplicity, and we refer the interested reader to Vershynin (2010) for further details. We next state our second set of assumptions which relates to structured sparsity conditions on $\boldsymbol{\beta}_0$, $\boldsymbol{\alpha}_0$ and some arbitrary realization of $\widehat{\boldsymbol{\beta}}_{n_q}(\lambda)$.

Assumption 3.4. *(Restricted Sparsity Conditions)* 1. We assume that $\boldsymbol{\beta}_0$ is strictly *sparser* than $\boldsymbol{\alpha}_0$ in the sense that $\mathcal{A}^c(\boldsymbol{\beta}_0) \cap \mathcal{A}(\boldsymbol{\alpha}_0)$ is non-empty, so that $\exists\ j \in \{1, \ldots, p\}$ such that $\boldsymbol{\beta}_{0[j]} = 0$ and $\boldsymbol{\alpha}_{0[j]} \neq 0$. Let $C_{\min}(\boldsymbol{\alpha}_0, \boldsymbol{\beta}_0) = \min\{|\boldsymbol{\alpha}_{0[j]}| : j \in \mathcal{A}^c(\boldsymbol{\beta}_0) \cap \mathcal{A}(\boldsymbol{\alpha}_0)\} > 0$, and let $C_{\max}(\boldsymbol{\alpha}_0, \boldsymbol{\beta}_0) = \max\{|\boldsymbol{\alpha}_{0[j]}| : j \in \mathcal{A}^c(\boldsymbol{\beta}_0) \cap \mathcal{A}_0(\boldsymbol{\alpha}_0)\} > 0$.
2. For a given choice of the tuning parameter $\lambda \geq 0$, we *define* λ to be $(\boldsymbol{\beta}_0, \boldsymbol{\alpha}_0, q)-$*admissible*, if $\exists$ *some* realization $z^*_{n_q}$ (*not* necessarily the observed one) of the data $\mathcal{Z}^*_{n_q}$ such that the corresponding estimator $\widehat{\boldsymbol{\beta}}_{n_q}(\lambda; z^*_{n_q})$ based on $z^*_{n_q}$ and the given choice of λ, satisfies the property: $\widehat{\boldsymbol{\beta}}_{n_q[j]}(z^*_{n_q}; \lambda) = 0$ for some $j \in \mathcal{A}^c(\boldsymbol{\beta}_0) \cap \mathcal{A}(\boldsymbol{\alpha}_0)$ (a non-empty set due to condition 1). In our subsequent results, λ, wherever involved, would be assumed to satisfy this condition.

Assumption 3.4 imposes some mild restrictions on the sparsity patterns of $\boldsymbol{\beta}_0$, $\boldsymbol{\alpha}_0$, and one arbitrary realization of $\widehat{\boldsymbol{\beta}}(\lambda; \mathcal{Z}^*_{n_q})$ for a given choice of λ. In particular, condition 1 needs $\boldsymbol{\beta}_0$ to be sparser than $\boldsymbol{\alpha}_0$ in at least one coordinate, and therefore formally characterizes the very essence and purpose behind our consideration of a penalized regression approach for recovering $\boldsymbol{\beta}_0$, which we motivated earlier through the intuition that the penalized solution would be favoring sparser solutions and try to push it away from the un-regularized solution that recovers the $\boldsymbol{\alpha}_0$ direction. Condition 1 therefore simply ensures that $\boldsymbol{\beta}_0$ is among one of these favorable sparser directions. Condition 2 is a somewhat unusual condition requiring a very mild, but nonetheless critical, assumption to hold regarding the sparsity structure of the solution $\widehat{\boldsymbol{\beta}}(\lambda; \mathcal{Z}^*_{n_q})$ for at least *one* realization of $\mathcal{Z}^*_{n_q}$ over the entire sample space underlying the generation of $\mathcal{Z}^*_{n_q}$. Given a λ, it requires the estimator, at least for one arbitrary sample point, to be sparse in one of the coordinates $\in \mathcal{A}^c(\boldsymbol{\beta}_0) \cap \mathcal{A}(\boldsymbol{\alpha}_0)$. Of course, while this can perhaps be always ensured by making the λ large, the main utility of this condition lies in ensuring the fact that even for choices of λ of a reasonably small enough order (which really determines the convergence rates of the estimators), at least one of the coordinates in $\boldsymbol{\alpha}_0$

that $\in \mathcal{A}^c(\boldsymbol{\beta}_0) \cap \mathcal{A}(\boldsymbol{\alpha}_0)$ should not be too strong so as to be always (over the whole sample space) selected by $\widehat{\boldsymbol{\beta}}(\lambda; \mathcal{Z}^*_{n_q})$. We now propose our deterministic deviation bound result.

Theorem 3.2. *Let $\mathcal{Z}^*_{n_q}$ be given, and suppose assumption 3.3 and condition 1 of assumption 3.4 hold. Let λ be any given choice of the tuning parameter in (3.16) such that $\lambda \geq 4\|\mathbb{T}_{n_q}\|_\infty$, where $\mathbb{T}_{n_q}$ is as in (3.17). Suppose further that the chosen λ is $(\boldsymbol{\beta}_0, \boldsymbol{\alpha}_0, q)$-admissible, as defined in condition 2 of assumption 3.4. Then, the corresponding ULASSO estimator $\widehat{\boldsymbol{\beta}}_{n_q}(\lambda) \equiv \widehat{\boldsymbol{\beta}}(\lambda; \mathcal{Z}^*_{n_q})$ satisfies the following deviation bound w.r.t. the $\boldsymbol{\beta}_0$ direction:*

$$\left\|\widehat{\boldsymbol{\beta}}_{n_q}(\lambda) - b_q\boldsymbol{\beta}_0\right\|_2 \leq \frac{\lambda}{\kappa_q}\left[\left\{9s_{\boldsymbol{\beta}_0} + d_1(\boldsymbol{\alpha}_0, \boldsymbol{\beta}_0)\right\}^{\frac{1}{2}} + d_2(\boldsymbol{\alpha}_0, \boldsymbol{\beta}_0)\right], \quad \textit{where} \tag{3.22}$$

b_q is as in (3.10), κ_q is as in (3.21), $s_{\boldsymbol{\beta}_0}$ is the size of $\mathcal{A}_{\boldsymbol{\beta}_0}$, and $d_1(\boldsymbol{\alpha}_0, \boldsymbol{\beta}_0)$, $d_2(\boldsymbol{\alpha}_0, \boldsymbol{\beta}_0) > 0$ are universal constants depending only on $\boldsymbol{\alpha}_0$ and $\boldsymbol{\beta}_0$ in a scale invariant manner as follows:

$$d_1(\boldsymbol{\alpha}_0, \boldsymbol{\beta}_0) = 4\overline{d}(\boldsymbol{\alpha}_0, \boldsymbol{\beta}_0)\left\|\Pi^c_{\boldsymbol{\beta}_0}(\boldsymbol{\alpha}_0)\right\|_1, \quad d_2(\boldsymbol{\alpha}_0, \boldsymbol{\beta}_0) = \overline{d}(\boldsymbol{\alpha}_0, \boldsymbol{\beta}_0)\|\boldsymbol{\alpha}_0\|_2, \textit{ with} \tag{3.23}$$
$$\overline{d}(\boldsymbol{\alpha}_0, \boldsymbol{\beta}_0) = \frac{4\left\|\Pi^c_{\boldsymbol{\beta}_0}(\boldsymbol{\alpha}_0)\right\|_1 + 3s^{\frac{1}{2}}_{\boldsymbol{\beta}_0} C_{max}(\boldsymbol{\alpha}_0, \boldsymbol{\beta}_0)}{C^2_{min}(\boldsymbol{\alpha}_0, \boldsymbol{\beta}_0)},$$

where $C_{max}(\boldsymbol{\alpha}_0, \boldsymbol{\beta}_0)$ and $C_{min}(\boldsymbol{\alpha}_0, \boldsymbol{\beta}_0)$ are as defined in condition 2 of assumption 3.4.

Theorem 3.2, for any given realization of $\mathcal{Z}^*_{n_q}$ and any choice of a corresponding λ that satisfies the required conditions, establishes a purely *deterministic* deviation bound of $\widehat{\boldsymbol{\beta}}_{n_q}(\lambda)$ w.r.t. a scalar multiple of $\boldsymbol{\beta}_0$ (the scalar multiple b_q being implicitly, and quite sensibly, assumed to be non-zero), as desired. Note that the constants $d_1(\boldsymbol{\alpha}_0, \boldsymbol{\beta}_0)$ and $d_2(\boldsymbol{\alpha}_0, \boldsymbol{\beta}_0)$ appearing in the bound (3.22) are invariant to the scaling of $\boldsymbol{\alpha}_0$ and $\boldsymbol{\beta}_0$, which is quite desirable and sensible given the SIM setting we have considered. Apart from the universal constants and the strong convexity constant, the bound primarily depends on λ whose order will essentially determine the convergence rate of $\widehat{\boldsymbol{\beta}}_{n_q}(\lambda)$. In this regard, the (random) lower bound $4\|\mathbb{T}_{n_q}\|_\infty$ characterizing the choice of λ in theorem 3.2 now becomes the quantity of primary interest. If we can find a non-random sequence a_{n_q} that converges to 0 at a satisfactorily fast

enough rate, and a_{n_q} can be shown to upper bound $\mathbb{T}_{n_q}$ with high probability (w.h.p.), then a choice of $\lambda = 4a_{n_q}$, as long as it satisfies the additional conditions required for theorem 3.2, will guarantee the bound in (3.22) to hold w.h.p. at the satisfactory rate of $O(a_{n_q}/\kappa_q)$.

Probabilistic performance guarantees: We next aim to provide such a probabilistic bound for $\mathbb{T}_{n_q}$. In order to do so, we shall require some mild restrictions on the underlying design distribution of $(\mathbf{X} \mid S \in \mathcal{I}_q)$, so that it is sufficiently well-behaved, and satisfies some nice and desirable concentration properties. In particular, we shall assume that it follows some sub-gaussian distribution. Recall that any random variable Z with $\mathbb{E}(Z) = 0$ is said to follow a sub-gaussian distribution with parameter σ^2, for some $\sigma \geq 0$, if $\mathbb{E}\{\exp(tZ)\} \leq \exp(\sigma^2 t^2/2)\ \forall\ t \in \mathbb{R}$. Further, a random vector $\mathbf{Z} \in \mathbb{R}^p$, with $\mathbb{E}(\mathbf{Z}) = \mathbf{0}$, is said to follow a sub-gaussian distribution with parameter σ^2, for some $\sigma \geq 0$, if for each $\mathbf{t} \in \mathbb{R}^p$, the random variable $\mathbf{t}'\mathbf{Z}$ follows a sub-gaussian distribution with parameter at most $\sigma^2\|\mathbf{t}\|_2^2$. We next define a similar notion of sub-gaussian distributions that is more suitable for our case, wherein we appropriately account for the underlying restriction of $\{S \in \mathcal{I}_q\}$, as follows.

Definition 3.1. *(Sub-gaussian distributions given $S \in \mathcal{I}_q$).* Let $Z \in \mathbb{R}$ and $\mathbf{Z} \in \mathbb{R}^p$ be any scalar and vector valued measurable functions of $(Y, S, \mathbf{X}')'$ respectively. Let $\widetilde{Z}_q = \{Z - \mathbb{E}_q(Z)\}$, and $\widetilde{\mathbf{Z}}_q = \{\mathbf{Z} - \mathbb{E}_q(\mathbf{Z})\}$ denote their corresponding centered versions given $\{S \in \mathcal{I}_q\}$, for any $q \in (0, 1]$. Then, Z is said to follow a $\{S \in \mathcal{I}_q\}$-*restricted sub-gaussian distribution* with parameter σ_q^2 for some constant $\sigma_q > 0$ (allowed to depend on q), to be denoted as: $Z \sim \mathbb{SG}_q(\sigma_q^2)$, if $\mathbb{E}_q\{\exp(t\widetilde{Z}_q)\} \leq \exp(\sigma_q^2 t^2/2)\ \forall\ t \in \mathbb{R}$. Further, $\mathbf{Z}$ is said to follow a $\{S \in \mathcal{I}_q\}$-restricted sub-gaussian distribution, with parameter σ_q^2 for some constant $\sigma_q > 0$ (allowed to depend on q), to be denoted as: $\mathbf{Z} \sim \mathbb{SG}_q(\sigma_q^2)$, if for each $\mathbf{t} \in \mathbb{R}^p$, $\mathbf{t}'\mathbf{Z}$ follows a $\{S \in \mathcal{I}_q\}$-restricted sub-gaussian distribution with parameter at most $\sigma_q^2\|\mathbf{t}\|_2^2$.

The conditions required to be satisfied by $\{S \in \mathcal{I}_q\}$-restricted sub-gaussian distributions, as introduced in definition 3.1 above, are quite mild, and should be expected to hold for a fairly large family of distributions for $(S, \mathbf{X}')'$, especially those where the (unconditional) dis-

tribution of $(S, \mathbf{X}')'$ is itself sub-gaussian. In particular, when $(S, \mathbf{X}')'$ follows a multivariate normal distribution, it can be shown, as would be discussed later, that for most small enough q of interest to us (specifically, $\forall\ q \leq 1/2$), $\mathbf{X} \sim \mathbb{SG}_q(\sigma_q^2)$ indeed with $\sigma_q^2 \leq c_1 \overline{\delta}_q^2 \leq c_2 \log(q^{-1})$ for some universal constants $c_1, c_2 > 0$. (For $q \in (1/2, 1]$, the moment generating function of $(\mathbf{X} \mid S \in \mathcal{I}_q)$ still follows a sub-gaussian type bound, but only upto a scalar multiple). Note also that the parameter σ_q in definition 3.1 is allowed to depend on q, and therefore, possibly diverge (slowly enough) as q decreases (as is seen to be the case when $(S, \mathbf{X}')'$ is normally distributed). We now provide our final result regarding probabilistic bounds for $\mathbb{T}_{n_q}$.

Theorem 3.3. *Suppose $\mathbf{X} \sim \mathbb{SG}_q(\sigma_q^2)$, as defined in 3.1, for some constant $\sigma_q > 0$ allowed to depend on q. For any $a \in [0,1]$, define $\widetilde{a} > 0$ as: $\widetilde{a} = 0$ if $a \in \{0, 1\}$, $\widetilde{a} = 1/2$ if $a = 1/2$, and $\widetilde{a} = [(a - 1/2)/log\{\pi_q/(1-\pi_q)\}]^{1/2}$ if $a \notin \{0, 1, 1/2\}$. Let $\widetilde{p}_q$ and $\widetilde{\pi}_q$ denote $\widetilde{a}$ for $a = p_q$ and $a = \pi_q$ respectively. Further, let $\gamma_q^2 = (\widetilde{p}_q + \sigma_q \|\overline{\boldsymbol{\beta}}_q\|_2)^2$. Then, with $\mathbb{T}_{n_q} = \mathbb{T}_{n_q}^{(1)} + \mathbb{T}_{n_q,1}^{(2)} - \mathbb{T}_{n_q,2}^{(2)}$ as defined in (3.17)-(3.20), we have: for any given $\epsilon_1, \epsilon_2, \epsilon_3, \epsilon_4, \epsilon_5 > 0$,*

$$
\begin{aligned}
&(i)\ \mathbb{P}_q\left\{\left\|\mathbb{T}_{n_q}^{(1)}\right\|_\infty > \epsilon_1(\pi_q + \epsilon_2)\right\} \leq\ 2exp\left\{-\frac{\epsilon_1^2}{2\sigma_q^2} + log\,(n_q p)\right\} + exp\left(-\frac{n_q \epsilon_2^2}{2\widetilde{\pi}_q^2}\right), \\
&(ii)\ \mathbb{P}_q\left\{\left\|\mathbb{T}_{n_q,1}^{(2)}\right\|_\infty > 2\sigma_q\gamma_q(2\sqrt{2}\epsilon_3 + \epsilon_3^2)\right\} \leq\ 2exp\left(-n_q\epsilon_3^2 + log\, p\right),\ and \\
&(iii)\ \mathbb{P}_q\left(\left\|\mathbb{T}_{n_q,2}^{(2)}\right\|_\infty > \epsilon_4\epsilon_5\right) \leq\ 2exp\left(-\frac{n_q\epsilon_4^2}{2\sigma_q^2} + log\, p\right) + 2exp\left(-\frac{n_q\epsilon_5^2}{2\gamma_q^2}\right). \qquad (3.24)
\end{aligned}
$$

In particular, for any universal constants: $c_1, c_2 > 0$ such that $max\,(c_1, c_2) > 1$; and $c_3 > 0$, $c_4, c_5 > 1$, and $c_6 > 0$; and letting $c_0 = (c_4 + c_5 c_6)$; and assuming $\pi_q < 1/2$ w.l.o.g., we have:

$$
\begin{aligned}
&With\ probability\ at\ least\ 1 - \left(\frac{\pi_q}{1-\pi_q}\right)^{c_3} - \frac{2}{p^{(c_1-1)} n_q^{(c_2-1)}} - \frac{2}{p^{(c_4-1)}} - \frac{2}{p^{(c_5-1)}} - \frac{2}{p^{c_6}}, \\
&\left\|\mathbb{T}_{n_q}\right\|_\infty\ \leq a_{n_q} \equiv a_{n_q}(c_1, \ldots, c_6), \quad where\ a_{n_q} \equiv a_{n_q}(c_1, \ldots, c_6)\ is\ given\ by: \\
&\sigma_q\sqrt{2log\,(p^{c_1} n_q^{c_2})}\left\{\pi_q + \sqrt{\frac{(1-2\pi_q)c_3}{n_q}}\right\} + 2\sigma_q\gamma_q\left(2\sqrt{2c_4}\sqrt{\frac{log\, p}{n_q}} + \frac{log\, p}{n_q}c_0\right). \qquad (3.25)
\end{aligned}
$$

Theorem 3.3 provides an explicit finite sample characterization of the behaviour of $\mathbb{T}_{n_q}$ in terms of a flexible probabilistic bound. In particular, it implies that for some suitably chosen

constants $(c_1, \ldots, c_6)$, a choice of $\lambda = 4a_{n_q}$ will ensure that the condition $\lambda \geq 4\|\mathbb{T}_{n_q}\|_\infty$, required for theorem 3.2, holds w.h.p. (characterized more explicitly in the bound above). Consequently, with a choice of $\lambda = 4a_{n_q}$, as long as it satisfies the other conditions required for theorem 3.2, the deviation bound (3.22) holds w.h.p. as well, thereby ensuring a satisfactory convergence rate of $O(a_{n_q}/\kappa_q)$ for $\widehat{\boldsymbol{\beta}}_{n_q}(\lambda)$ as an estimator of the $\boldsymbol{\beta}_0$ direction.

Turning to the convergence rate of a_{n_q} itself, we note that the (dominating) polynomial part of the rate is determined primarily by π_q and $n_q^{-1/2}$, which behave antagonistically w.r.t. each other as q increases or decreases, so that the rate exhibits an interesting phenomenon similar to a 'variance-bias tradeoff'. The misclassification error π_q, expected to increase as q increases, can be viewed as a 'bias' term, while $n_q^{-1/2}$, which decreases as q increases, corresponds to the usual variance (rather, standard deviation) term. In particular, with $\pi_q = O(q^\nu)$ for some given $\nu > 0$, as in (3.1), and $q = O(N^{-\eta})$ for some unknown $\eta \in (0, 1)$, the combined rate: $(\pi_q + n_q^{-1/2}) \equiv O\{N^{-\nu\eta} + N^{(1-\eta)/2}\}$ can be minimized w.r.t. η, leading to an optimal choice given by: $\eta_{opt} = 1/(2\nu+1)$, and a corresponding optimal order of q given by: $q_{opt} = O\{N^{-1/(2\nu+1)}\}$. For $q = q_{opt}$, π_q and $n_q^{-1/2}$ have the same order, so that optimal order of the (polynomial part of the) convergence rate of a_{n_q} is given by: $(a_{n_q})_{opt} = O\{N^{-\nu/(2\nu+1)}\}$.

Practical choice of λ: Finally, while we have so far characterized the theoretical properties of $\widehat{\boldsymbol{\beta}}_{n_q}(\lambda)$ in great detail and generality, an important issue that has not yet been addressed is the choice of λ in practice, required for the actual implementation of $\widehat{\boldsymbol{\beta}}_{n_q}(\lambda)$, since π_q (and ν) would be typically unknown in reality. In this regard, we first note that, the theoretical choice $4a_{n_q}$ for λ is essentially of the order: $O[\{\log\ (n_q p)\}^{1/2}(\pi_q + n_q^{-1/2})]$ (ignoring, for simplicity, the constants σ_q and γ_q). Owing to the additional π_q term (as well as the $\log\ (n_q p)$ term), this is therefore expected to be slightly higher than $O[\{(\log p)/n_q\}^{1/2}]$ which is well known to be the typical choice of the order of λ for standard L_1-penalized estimation methods. This is again a manifestation of the fact that in our case, the choice of the λ should be slightly *higher than usual*, so that sparser solutions are favored. Motivated

by this intuition, we now propose to choose the tuning parameter λ in practice through minimizing a criteria similar to the Bayes Information Criteria (BIC) defined as follows:

$$\text{BIC}(\lambda) \equiv \text{BIC}\{\lambda; \widehat{\boldsymbol{\beta}}_{n_q}(\lambda); \mathcal{Z}^*_{n_q}\} = \mathcal{L}_{n_q}\{\mathcal{Z}^*_{n_q}; \widehat{\boldsymbol{\beta}}_{n_q}(\lambda)\} + \frac{\log(n_q)}{n_q}\|\widehat{\boldsymbol{\beta}}_{n_q}(\lambda)\|_0, \tag{3.26}$$

where, $\forall\ \mathbf{v} \in \mathbb{R}^p$, $\|\mathbf{v}\|_0 = s_{\mathbf{v}}$ denotes the L_0 pseudo-norm, and $\mathcal{L}_{n_q}(.;.)$ is as in (3.12). Compared to other standard criteria for selecting tuning parameters, like the Akaike Information Criteria (AIC) and cross-validation (CV), the BIC is known to penalize more and therefore, select sparser solutions which serves quite well for our purpose. While a detailed theoretical analysis of the merits and demerits of using the BIC for selecting λ in our case is beyond the scope of this paper, we find that, based on our extensive simulation studies as well as applications to real data, the above criteria works quite well in practice, and we believe this continues to hold in general as long as the π_q and the chosen q are reasonable enough.

3.4 Analysis of Key Quantities for a Familiar Choice of (Y, S, X)

We next characterize, for a particularly familiar choice of $(Y, S, \mathbf{X})$, the distributional properties of $\{(S, \mathbf{X}) \,|\, S \in \mathcal{I}_q\}$, as well as the behaviour of π_q, all of which are closely related to the fundamental assumptions underlying our proposed methodology for this problem. Suppose $\mathbf{X} \sim \mathcal{N}_p(\boldsymbol{\mu}, \Sigma)$, the $p-$variate normal distribution with mean $\boldsymbol{\mu} \in \mathbb{R}^p$ and dispersion matrix $\Sigma_{p\times p}$. Further, let $S \,|\, \mathbf{X}$ follow the standard linear model: $S = \boldsymbol{\alpha}_0'\mathbf{X} + \epsilon^*$ for some $\boldsymbol{\alpha}_0 \in \mathbb{R}^p$, where $\epsilon^* \perp\!\!\!\perp \mathbf{X}$ and $\epsilon^* \sim \mathcal{N}_1(a, \sigma^2)$, for some $a \in \mathbb{R}$ and $\sigma \geq 0$. Lastly, let $(Y \,|\, \mathbf{X}) \equiv (Y \,|\, S, \mathbf{X})$ follow the standard logistic model given by: $Y = 1(\boldsymbol{\beta}_0'\mathbf{X} + \epsilon > 0)$ for some $\boldsymbol{\beta}_0 \in \mathbb{R}^p$, where $\epsilon \perp\!\!\!\perp (S, \mathbf{X})$ and $\epsilon \sim$ Logistic $(b, 1)$, the standard logistic distribution with mean b, for some $b \in \mathbb{R}$, and variance 1. For simplicity, we shall assume w.l.o.g. that $\boldsymbol{\mu} = \mathbf{0}$, $a = 0$, and $b = 0$.

Let $\sigma_S^2 \equiv \text{Var}(S) = (\boldsymbol{\alpha}_0'\Sigma\boldsymbol{\alpha}_0 + \sigma^2)$, and let $\psi(\boldsymbol{\beta}_0'\mathbf{X}) = \mathbb{P}(Y = 1 \,|\, \mathbf{X}) \equiv \mathbb{P}(Y = 1 \,|\, \mathbf{X}, S)$, with $\psi(u) = \exp(u)/\{1 + \exp(u)\}\ \forall\ u \in \mathbb{R}$. Note that the assumed set-up also implies that

$(S, \mathbf{X})'$ jointly follows a $(p+1)$-variate normal distribution, and further, $\mathbf{X} \mid S \sim \mathcal{N}_p(\boldsymbol{\gamma}_0, \Gamma)$ with $\boldsymbol{\gamma}_0 = (\Sigma\boldsymbol{\alpha}_0)/\sigma_S^2 \in \mathbb{R}^p$ and $\Gamma_{p\times p} = (\Sigma - \sigma_S^2\boldsymbol{\gamma}_0\boldsymbol{\gamma}_0')$. Let $\rho_0 \equiv \rho_0(\boldsymbol{\alpha}_0, \boldsymbol{\beta}_0) \in [-1, 1]$ denote Corr $(\boldsymbol{\alpha}_0'\mathbf{X}, \boldsymbol{\beta}_0'\mathbf{X})$, and assume w.l.o.g. that the signs of $\boldsymbol{\alpha}_0$ and $\boldsymbol{\beta}_0$ have been appropriately chosen such that $\rho_0 \geq 0$. Let $\widetilde{\rho}_0 \equiv \widetilde{\rho}(\boldsymbol{\alpha}_0, \boldsymbol{\beta}_0) = \rho_0\ (\boldsymbol{\alpha}_0'\Sigma\boldsymbol{\alpha}_0/\sigma_S^2)^{1/2} \in [0, 1]$, the fraction of ρ_0 scaled by the population R^2 value for the regression of S on $\mathbf{X}$. Let $\eta_0 \equiv \eta_0(\boldsymbol{\beta}_0) > 0$ denote $(\boldsymbol{\beta}_0'\Sigma\boldsymbol{\beta}_0)^{1/2}$. Lastly, let $\Phi(.)$ and $\phi(.)$ denote the cumulative distribution function (c.d.f.) and the density function respectively of the standard $\mathcal{N}_1(0, 1)$ distribution, and for any $q \in (0, 1]$, let z_q and $\overline{z}_q$ denote its $(q/2)^{th}$ and $(1 - q/2)^{th}$ quantiles. Hence, $z_q \leq 0, \overline{z}_q \geq 0$, $z_q = -\overline{z}_q$, and further, $\delta_q = \sigma_S z_q$ and $\overline{\delta}_q = \sigma_S \overline{z}_q$. We now propose the following result.

Theorem 3.4. *Consider the particular set-up introduced above for $(Y, S, \mathbf{X})$. Then,*

(i) The expectations of S and $\mathbf{X}$ given $S \in \mathcal{I}_q$ satisfy the following relations:

$$\mathbb{E}(S \mid S \geq \overline{\delta}_q) = -\ \mathbb{E}(S \mid S \leq \delta_q) = \sigma_S \frac{\phi(z_q)}{\Phi(z_q)},\ \textit{and}\ \ \mathbb{E}_q(S) = 0.$$
$$\mathbb{E}(\mathbf{X} \mid S \geq \overline{\delta}_q) = -\ \mathbb{E}(\mathbf{X} \mid S \leq \delta_q) = \boldsymbol{\gamma}_0 \left\{\sigma_S \frac{\phi(z_q)}{\Phi(z_q)}\right\},\ \textit{and}\ \ \mathbb{E}_q(\mathbf{X}) = \mathbf{0}.$$

(ii) The variances of S and $\mathbf{X}$ given $S \in \mathcal{I}_q$, denoted henceforth by $Var_q(.)$, satisfy:

$$\mathbb{E}(S^2 \mid S \geq \overline{\delta}_q) = \mathbb{E}(S^2 \mid S \leq \delta_q) = \sigma_S^2 \left\{1 + \overline{z}_q \frac{\phi(z_q)}{\Phi(z_q)}\right\} \equiv Var_q\ (S).$$
$$\mathbb{E}(\mathbf{X}\mathbf{X}' \mid S \geq \overline{\delta}_q) = \mathbb{E}(\mathbf{X}\mathbf{X}' \mid S \leq \delta_q) = \Sigma + \boldsymbol{\gamma}_0\boldsymbol{\gamma}_0' \left\{\sigma_S^2\ \overline{z}_q \frac{\phi(z_q)}{\Phi(z_q)}\right\} \equiv Var_q\ (\mathbf{X}).$$

(iii) The moment generating functions (m.g.f.s) of S and $\mathbf{X}$ given $S \in \mathcal{I}_q$ satisfy:

$$MGF_{S,q}(t) \equiv \mathbb{E}_q(e^{tS}) = \frac{e^{\sigma_S^2 t^2/2}}{2\Phi(z_q)}\{\Phi(z_q + \sigma_S t) + \Phi(z_q - \sigma_S t)\}\ \ \forall\ t \in \mathbb{R}.$$
$$MGF_{\mathbf{X},q}(\mathbf{t}) \equiv \mathbb{E}_q(e^{\mathbf{t}'\mathbf{X}}) = \frac{e^{\mathbf{t}'\Sigma\mathbf{t}/2}}{2\Phi(z_q)}\{\Phi(z_q + \sigma_S \mathbf{t}'\boldsymbol{\gamma}_0) + \Phi(z_q - \sigma_S \mathbf{t}'\boldsymbol{\gamma}_0)\}\ \ \forall\ \mathbf{t} \in \mathbb{R}^p.$$

(iv) Let $\lambda_{max}(\Sigma) > 0$ denote the maximum eigenvalue of Σ. Then, the m.g.f.s $MGF_{S,q}(.)$

and $MGF_{\mathbf{X},q}(.)$ *further satisfy the following sub-gaussian type bounds:*

$$MGF_{S,q}(t) \leq \ exp\left\{\frac{1}{2}t^2\sigma_S^2(1+2z_q^2)\right\} \ \forall\, t \in \mathbb{R};\ \ \forall\, q \in (0,1/2].$$
$$MGF_{S,q}(t) \leq \ 4\ exp\left(\frac{1}{2}t^2\sigma_S^2\right),\ \ \forall\, t \in \mathbb{R};\ \ \forall\, q \in (1/2,1].$$
$$MGF_{\mathbf{X},q}(\mathbf{t}) \leq \ exp\left[\frac{1}{2}\|t\|_2^2\{\lambda_{max}(\Sigma)+2\sigma_S^2 z_q^2\|\boldsymbol{\gamma}_0\|_2^2\}\right] \ \forall\, \mathbf{t} \in \mathbb{R}^p;\ \ \forall\, q \in (0,1/2].$$
$$MGF_{\mathbf{X},q}(\mathbf{t}) \leq \ 4\ exp\left\{\frac{1}{2}\|t\|_2^2\lambda_{max}(\Sigma)\right\} \ \forall\, \mathbf{t} \in \mathbb{R}^p;\ \ \forall\, q \in (1/2,1].$$

(v) The misclassification error π_q *(as well as* π_q^+ *and* π_q^-*) satisfies the following bound:*

$$\pi_q \equiv \frac{1}{2}(\pi_q^+ + \pi_q^-) \ \leq \ exp\left(\eta_0^2/2\right)\frac{\Phi\left(-\overline{z}_q - \sigma_S\boldsymbol{\beta}_0'\boldsymbol{\gamma}_0\right)}{\Phi\left(-\overline{z}_q\right)} \quad \forall\, q \in (0,1];$$
$$\leq exp\left\{\frac{1}{2}(1-\widetilde{\rho}_0^2)\eta_0^2 - \overline{z}_q\widetilde{\rho}_0\eta_0\right\}\frac{(\overline{z}_q^2+1)}{\overline{z}_q(\overline{z}_q+\widetilde{\rho}_0\eta_0)} \ \lesssim C\,exp\left\{\frac{1}{2}(1-\widetilde{\rho}_0^2)\eta_0^2 - \overline{z}_q\widetilde{\rho}_0\eta_0\right\},$$

where $C > 0$ *is some universal constant, and all other notations are as defined earlier.*

(vi) Lastly, the behavior of $\overline{\delta}_q^2$ *(as well as* δ_q^2*) w.r.t.* q *is reflected by the following bounds:*

$$\overline{\delta}_q^2 \ \equiv \ \sigma_S^2\overline{z}_q^2 \ = \ \sigma_S^2 z_q^2 \ \equiv \ \delta_q^2 \ \leq \ 2\sigma_S^2 log\left(q^{-1}\right);\ \ \forall\, q \in (0,1].$$
$$\overline{\delta}_q^2 \ \equiv \ \sigma_S^2\overline{z}_q^2 \ = \ \sigma_S^2 z_q^2 \ \equiv \ \delta_q^2 \ \geq \ 2\sigma_S^2 log\left\{(5q)^{-1}\right\};\ \ \forall\, q \in [0.0002,1].$$

Theorem 3.4, for the special choice of $(Y, S, \mathbf{X})$ considered herein, explicitly characterizes the distributional properties of $\{(S,\mathbf{X}) \,|\, S \in \mathcal{I}_q\}$, as well as the behavior of the misclassification error π_q w.r.t. q. In particular, result (iv) implies that for most choices of q that are of interest to us, $\mathbf{X} \sim \mathbb{SG}_q(\sigma_q^2)$ indeed, as required in theorem 3.3, for some appropriate choice of the constant $\sigma_q > 0$ depending on q only through $\overline{z}_q$. Further, owing to result (vi), $\overline{z}_q^2$ (and hence, σ_q^2) diverges slowly enough, only at logarithmic orders, as $q \downarrow 0$, thereby showing that, at least in this case, σ_q (and γ_q) appearing in the bound (3.25) only has a minor effect on the convergence rate of $\lambda = 4a_{n_q}$ and consequently, on that of our estimator $\widehat{\boldsymbol{\beta}}_{n_q}(\lambda)$.

Moreover, the strict positive definiteness of Σ_q as shown by result (ii), combined with the fact that $\mathbf{X} \sim \mathbb{SG}(\sigma_q^2)$ for some σ_q, also ensures that our strong convexity assumption 3.3 can be ensured to hold in this case w.h.p., using the results from Vershynin (2010), through some appropriate choice of the constant $\kappa_q > 0$ which, with some more work, can also be shown to be uniformly bounded below w.r.t. q by some universal constant.

Finally, turning our focus onto the bound of π_q obtained in result (v) which, it needs to be noted, is a fairly sharp bound (especially for small enough q), we first observe that apart from $\overline{z}_q$, it depends critically on the constants η_0 and $\widetilde{\rho}_0$ that can be respectively interpreted as the 'strength' of the signal $\boldsymbol{\beta}_0'\mathbf{X}$ underlying the generation of $Y \mid \mathbf{X}$, and that of the correlation ρ_0 (upto a minor scaling constant) between the $\boldsymbol{\beta}_0$ and the $\boldsymbol{\alpha}_0$ directions. In order for this bound to obey a polynomial rate w.r.t. q, as we have assumed in (3.1), it must behave as: $c_1\exp(-c_2\overline{z}_q^2)$, owing to result (vi), for some constants $c_1, c_2 > 0$. However, treating η_0 and $\widetilde{\rho}_0$ as universal constants, the bound clearly behaves only as: $c_1^*\exp(-c_2^*\overline{z}_q)$, for some constants $c_1^*, c_2^* > 0$, thereby leading to a slower rate than desired.

This therefore indicates that it would be perhaps more helpful to envision a 'regime' where η_0 and $\widetilde{\rho}_0$ are allowed to vary with q, so that η_0 increases and $(1-\widetilde{\rho}_0^2)$ decreases, both albeit slowly enough, as q decreases to 0. In particular, for a given choice of q, if η_0 is strong enough such that $\eta_0^2 \gtrsim d_1\overline{z}_q^2$, and at the same time $(1-\widetilde{\rho}_0^2)$ is small enough so that $(1-\widetilde{\rho}_0^2)\eta_0^2 \lesssim d_2$, for some constants $d_1, d_2 > 0$, then clearly, the bound starts behaving as: $c_1\exp(-c_2\overline{z}_q^2)$ for some constants $c_1, c_2 > 0$, as desired. Further, owing to result (vi), note that this setting also necessarily implies that: $(1-\rho_0^2) \leq (1-\widetilde{\rho}_0^2) \lesssim O(\overline{z}_q^{-2}) \lesssim O\{\log(q^{-1})\} \lesssim O\{1/(\log N)\}$, thereby indicating that the $\boldsymbol{\alpha}_0$ and $\boldsymbol{\beta}_0$ directions must be fairly strongly correlated with each other, at least upto $1/(\log N)$ order, under this regime, a regime that is almost necessary, in this case, for our surrogacy assumption 3.1 to hold, and for our approach to be successful.

Intuitively, this makes sense since $\mathbb{E}(S \mid \mathbf{X}) \equiv \boldsymbol{\alpha}_0'\mathbf{X}$ and $\mathbb{P}(Y = 1 \mid \mathbf{X}) \equiv \psi(\boldsymbol{\beta}_0'\mathbf{X})$ are monotone in $\boldsymbol{\alpha}_0'\mathbf{X}$ an $\boldsymbol{\beta}_0'\mathbf{X}$ respectively, so that given $\mathbf{X}$, the tails of S will be closely linked to those of $\boldsymbol{\alpha}_0'\mathbf{X}$, and further owing to the surrogacy assumption in these tails, the corresponding

$\psi(\boldsymbol{\beta}_0'\mathbf{X})$ should be close to 0/1 indicating that $\boldsymbol{\beta}_0'\mathbf{X}$ should also lie in its own tails, thereby implying the necessity of a fairly strong correlation between the $\boldsymbol{\alpha}_0$ and $\boldsymbol{\beta}_0$ directions. Note however that this does *not* trivialize the problem or our proposed methods in any way for this case. While the $\boldsymbol{\alpha}_0$ direction does need to be 'close' to the $\boldsymbol{\beta}_0$ direction in this case, it only needs to be so in $1/(\log N)$ order (and therefore, not necessarily 'too close'). Hence, while the $\boldsymbol{\alpha}_0$ direction can indeed be near perfectly estimated from $\mathcal{S}_N^*$ at a rate of $O(N^{-1/2})$, it may not be a reasonable estimator of the $\boldsymbol{\beta}_0$ direction itself, which it might be only able to estimate at sub-optimal logarithmic rates, instead of the polynomial rates we have ensured, under the additional structured sparsity assumptions, for our estimator based on $\mathcal{Z}_{n_q}^*$.

3.5 Numerical Studies

3.5.1 Simulation Results

We conducted extensive simulation studies to compare the performance of our proposed ULASSO estimator to those of standard supervised estimators based on labeled data of various sample sizes. Throughout, we let $N = 100000$, and use two choices each for p, q and the labeled data size (n) given by: $p = 20$ or $p = 50$, $q = 0.02$ or $q = 0.04$, and $n = 300$ or $n = 500$ (denoted henceforth as n_{300} or n_{500} respectively). $\mathbf{X}$ is generated as: $\mathbf{X} \sim \mathcal{N}_p(\mathbf{0}, \Sigma)$, where we use two choices of Σ, representing two types of correlation structures for $\mathbf{X}$, given by: $\Sigma = I_p$ (the $p \times p$ identity matrix) or, $\Sigma = \Sigma_\rho$, where the $(i,j)^{th}$ entry of Σ_ρ is given by: $(\Sigma_\rho)_{[i,j]} = \rho^{|i-j|}\ \forall\ 1 \leq i, j \leq p$, with $\rho = 0.2$. For each choice of p and Σ as above, we generated $Y|\mathbf{X}$ based on the standard logistic regression model given by: $Y = 1(\boldsymbol{\beta}_0'\mathbf{X} + \epsilon > 0)$ with $\epsilon \sim$ Logistic $(0,1)$ and $\epsilon \perp\!\!\!\perp (\mathbf{X}, S)$, and throughout, we set $\boldsymbol{\beta}_0 = (\mathbf{1}'_{c_p}, 0.5 * \mathbf{1}'_{c_p}, \mathbf{0}'_{p-2c_p})'$, where $\mathbf{1}_a$ and $\mathbf{0}_a$, for any a, denote $(1, \ldots, 1)'_{a\times 1}$, and $(0, \ldots, 0)'_{a\times 1}$ respectively, and $c_p = \lfloor p^{\frac{1}{2}} \rfloor$ with $\lfloor . \rfloor$ being the floor function. Finally, $S\,|\,\mathbf{X}$ is generated from the standard linear model: $S = \boldsymbol{\alpha}_0'\mathbf{X} + \epsilon^*$ with $\epsilon^* \sim \mathcal{N}_1(0,1)$ and $\epsilon^* \perp\!\!\!\perp (\mathbf{X}, \epsilon)$, where for any choice of (p, Σ, q) as above, α_0 is generated as: $\boldsymbol{\beta}_0 + \boldsymbol{\xi}/(\log N)$ with the entries $\{\boldsymbol{\xi}_{[j]}\}_{j=1}^p$ of $\boldsymbol{\xi}$ being *fixed* realizations from either a Uniform (r_1, r_2) distribution with $r_1 = 2, r_2 = 5$, or from a $\mathcal{N}_1(r_1, r_2^2)$ distribution

with $r_1 = 3, r_2 = 1$. Note that such choices of $\boldsymbol{\alpha}_0$, motivated by our discussions in section 3.4, ensure that while $\boldsymbol{\alpha}_0$ is 'close' to $\boldsymbol{\beta}_0$, yet its deviations of $O(1/\log n)$ from $\boldsymbol{\beta}_0$ are still substantially large enough so as to lead to a non-trivial setting.

For each of the settings above, we replicated the simulations over 500 iterations, and for each such iteration, we obtained our ULASSO estimator, as well as the following supervised estimators based on a labeled data (at both n_{300} and n_{500}): (i) the logistic LASSO estimator (Sup.Log.LASSO) and the logistic MLE (Sup.Log.MLE) through penalized and un-penalized logistic regression of Y on $\mathbf{X}$ respectively, and (ii) the least squares LASSO (Sup.Lin.LASSO) estimator and the OLS estimator (Sup.Lin.OLS) through penalized and un-penalized linear regression of Y on $\mathbf{X}$ respectively. In addition, we also considered $\boldsymbol{\alpha}_0$ as a possible estimator of the $\boldsymbol{\beta}_0$ direction. For the ULASSO estimator, the tuning parameter was selected using the BIC criteria defined in (3.26). For the Sup.Log.LASSO estimator, the tuning parameter was selected using the appropriate BIC criteria based on the logistic loss, while for the Sup.Lin.LASSO estimator, it was selected using the appropriate AIC criteria based on the squared loss, to avoid over-shrinkage. All estimators were further sign-normalized w.r.t. $\boldsymbol{\beta}_0$ as follows: $\widetilde{\boldsymbol{\beta}}' \Sigma \boldsymbol{\beta}_0 \geq 0$ where $\widetilde{\boldsymbol{\beta}}$ denotes any generic estimator of the $\boldsymbol{\beta}_0$ direction.

For all the estimators, we report the empirical mean squared error (Emp. MSE) of the normalized versions of the estimators w.r.t. the normalized version of $\boldsymbol{\beta}_0$ defined as: $\|\widetilde{\boldsymbol{\beta}}/\|\widetilde{\boldsymbol{\beta}}\|_2 - \boldsymbol{\beta}_0/\|\boldsymbol{\beta}_0\|_2\|_2^2$, for any generic estimator $\widetilde{\boldsymbol{\beta}}$, averaged over the 500 iterations. We also report the relative efficiency (RE) of ULASSO w.r.t. all other estimators considered, defined as the inverse ratio of the Emp. MSE of ULASSO to those of the other estimators. In order to compare the prediction/classification performance, we consider the area under the ROC curve (AUC) measure, notably a scale-invariant measure, for all the estimators. For each estimator and each iteration, the AUC was calculated based on an independent validation data of size N, and we report its average (Avg. AUC) over the 500 iterations. As a performance benchmark, we also report the corresponding average AUC (denoted as AUC_{oracle}) obtained by using the true $\boldsymbol{\beta}_0$ direction. Lastly, to compare the variable selection

performance of all the LASSO estimators, we obtained their corresponding true positive rate (TPR) and false positive rate (FPR) w.r.t. $\mathcal{A}(\boldsymbol{\beta}_0)$, defined as: TPR $= |\mathcal{A}(\widetilde{\boldsymbol{\beta}}) \cap \mathcal{A}(\boldsymbol{\beta}_0)|/|\mathcal{A}(\boldsymbol{\beta}_0)|$ and FPR $= |\mathcal{A}(\widetilde{\boldsymbol{\beta}}) \cap \mathcal{A}^c(\boldsymbol{\beta}_0)|/|\mathcal{A}^c(\boldsymbol{\beta}_0)|$ for any generic estimator $\widetilde{\boldsymbol{\beta}}$, and report their average TPR (Avg. TPR) and average FPR (Avg. FPR), averaged over the 500 iterations. Results for all the 16 simulation settings introduced herein are now tabulated through tables 3.1-3.16.

Table 3.1: Results for $p = 20, \Sigma = I_p, q = 0.02$, and $\{\boldsymbol{\xi}_{[j]}\}_{j=1}^{p} \sim \mathcal{N}_1(3, 1)$. Comparison of ULASSO and all supervised estimators, as well as $\boldsymbol{\alpha}_0$, based on Emp. MSE, AUC, TPR and FPR. Shown also are the REs, based on Emp. MSE, of ULASSO w.r.t. all the estimators.

Criteria	ULASSO	Sup.Log.LASSO		Sup.Lin.LASSO		Sup.Log.MLE		Sup.Lin.OLS		$\boldsymbol{\alpha}_0$
		n_{300}	n_{500}	n_{300}	n_{500}	n_{300}	n_{500}	n_{300}	n_{500}	
Emp. MSE	0.018	0.074	0.040	0.122	0.052	0.091	0.054	0.094	0.056	0.100
RE of ULASSO	1.000	4.157	**2.224**	6.799	2.925	5.104	3.026	5.275	3.140	5.599
Avg. AUC	**0.875**	0.863	**0.871**	0.853	0.868	0.860	0.868	0.859	0.867	0.858
AUC_{oracle}	**0.879**	0.879	0.879	0.879	0.879	0.879	0.879	0.879	0.879	0.879
Avg. TPR	**1.000**	0.978	**0.998**	0.867	0.979	–	–	–	–	–
Avg. FPR	**0.002**	0.227	**0.246**	0.035	0.050	–	–	–	–	–

Table 3.2: Results for $p = 20, \Sigma = I_p, q = 0.02$, and $\{\boldsymbol{\xi}_{[j]}\}_{j=1}^{p} \sim$ Uniform $(2, 5)$. Comparison of ULASSO and all supervised estimators, as well as $\boldsymbol{\alpha}_0$, based on Emp. MSE, AUC, TPR and FPR. Shown also are the REs, based on Emp. MSE, of ULASSO w.r.t. all the estimators.

Criteria	ULASSO	Sup.Log.LASSO		Sup.Lin.LASSO		Sup.Log.MLE		Sup.Lin.OLS		$\boldsymbol{\alpha}_0$
		n_{300}	n_{500}	n_{300}	n_{500}	n_{300}	n_{500}	n_{300}	n_{500}	
Emp. MSE	0.012	0.082	0.041	0.134	0.053	0.094	0.054	0.096	0.057	0.124
RE of ULASSO	1.000	6.965	**3.478**	11.321	4.478	7.899	4.574	8.115	4.791	10.460
Avg. AUC	**0.877**	0.862	**0.870**	0.851	0.868	0.859	0.868	0.859	0.867	0.853
AUC_{oracle}	**0.879**	0.879	0.879	0.879	0.879	0.879	0.879	0.879	0.879	0.879
Avg. TPR	**1.000**	0.968	**0.996**	0.848	0.978	–	–	–	–	–
Avg. FPR	**0.000**	0.221	**0.238**	0.034	0.058	–	–	–	–	–

Table 3.3: Results for $p = 20, \Sigma = I_p, q = 0.04$, and $\{\boldsymbol{\xi}_{[j]}\}_{j=1}^{p} \sim \mathcal{N}_1(3,1)$. Comparison of ULASSO and all supervised estimators, as well as $\boldsymbol{\alpha}_0$, based on Emp. MSE, AUC, TPR and FPR. Shown also are the REs, based on Emp. MSE, of ULASSO w.r.t. all the estimators.

Criteria	ULASSO	Sup.Log.LASSO		Sup.Lin.LASSO		Sup.Log.MLE		Sup.Lin.OLS		$\boldsymbol{\alpha}_0$
		n_{300}	n_{500}	n_{300}	n_{500}	n_{300}	n_{500}	n_{300}	n_{500}	
Emp. MSE	0.010	0.080	0.040	0.131	0.056	0.094	0.055	0.097	0.057	0.113
RE of ULASSO	1.000	8.046	**4.027**	13.091	5.572	9.366	5.461	9.693	5.737	11.246
Avg. AUC	**0.877**	0.862	**0.870**	0.851	0.867	0.859	0.867	0.858	0.867	0.855
AUC_{oracle}	**0.879**	0.879	0.879	0.879	0.879	0.879	0.879	0.879	0.879	0.879
Avg. TPR	**1.000**	0.972	**0.997**	0.846	0.973	–	–	–	–	–
Avg. FPR	**0.015**	0.229	**0.234**	0.033	0.053	–	–	–	–	–

Table 3.4: Results for $p = 20, \Sigma = I_p, q = 0.04$, and $\{\boldsymbol{\xi}_{[j]}\}_{j=1}^{p} \sim$ Uniform $(2,5)$. Comparison of ULASSO and all supervised estimators, as well as $\boldsymbol{\alpha}_0$, based on Emp. MSE, AUC, TPR and FPR. Shown also are the REs, based on Emp. MSE, of ULASSO w.r.t. all the estimators.

Criteria	ULASSO	Sup.Log.LASSO		Sup.Lin.LASSO		Sup.Log.MLE		Sup.Lin.OLS		$\boldsymbol{\alpha}_0$
		n_{300}	n_{500}	n_{300}	n_{500}	n_{300}	n_{500}	n_{300}	n_{500}	
Emp. MSE	0.011	0.077	0.042	0.121	0.056	0.093	0.056	0.095	0.059	0.091
RE of ULASSO	1.000	7.232	**3.914**	11.319	5.253	8.692	5.204	8.926	5.501	8.543
Avg. AUC	**0.877**	0.863	**0.870**	0.854	0.867	0.859	0.867	0.859	0.867	0.860
AUC_{oracle}	**0.879**	0.879	0.879	0.879	0.879	0.879	0.879	0.879	0.879	0.879
Avg. TPR	**1.000**	0.976	**0.997**	0.874	0.975	–	–	–	–	–
Avg. FPR	**0.001**	0.234	**0.231**	0.043	0.057	–	–	–	–	–

Table 3.5: Results for $p = 20, \Sigma = \Sigma_\rho, q = 0.02$, and $\{\boldsymbol{\xi}_{[j]}\}_{j=1}^{p} \sim \mathcal{N}_1(3,1)$.. Comparison of ULASSO and all supervised estimators, as well as $\boldsymbol{\alpha}_0$, based on Emp. MSE, AUC, TPR and FPR. Shown also are the REs, based on Emp. MSE, of ULASSO w.r.t. all the estimators.

Criteria	ULASSO	Sup.Log.LASSO		Sup.Lin.LASSO		Sup.Log.MLE		Sup.Lin.OLS		$\boldsymbol{\alpha}_0$
		n_{300}	n_{500}	n_{300}	n_{500}	n_{300}	n_{500}	n_{300}	n_{500}	
Emp. MSE	0.024	0.081	0.043	0.112	0.054	0.115	0.066	0.120	0.071	0.106
RE of ULASSO	1.000	3.358	**1.778**	4.656	2.227	4.776	2.755	4.980	2.944	4.413
Avg. AUC	**0.899**	0.891	**0.897**	0.885	0.895	0.886	0.893	0.885	0.892	0.880
AUC_{oracle}	**0.904**	0.904	0.904	0.904	0.904	0.904	0.904	0.904	0.904	0.904
Avg. TPR	**1.000**	0.977	**0.999**	0.902	0.981	–	–	–	–	–
Avg. FPR	**0.000**	0.200	**0.201**	0.024	0.035	–	–	–	–	–

Table 3.6: Results for $p = 20, \Sigma = \Sigma_\rho, q = 0.02$, and $\{\boldsymbol{\xi}_{[j]}\}_{j=1}^p \sim$ Uniform $(2, 5)$. Comparison of ULASSO and all supervised estimators, as well as $\boldsymbol{\alpha}_0$, based on Emp. MSE, AUC, TPR and FPR. Shown also are the REs, based on Emp. MSE, of ULASSO w.r.t. all the estimators.

Criteria	ULASSO	Sup.Log.LASSO		Sup.Lin.LASSO		Sup.Log.MLE		Sup.Lin.OLS		$\boldsymbol{\alpha}_0$
		n_{300}	n_{500}	n_{300}	n_{500}	n_{300}	n_{500}	n_{300}	n_{500}	
Emp. MSE	0.018	0.078	0.044	0.111	0.054	0.112	0.068	0.118	0.073	0.135
RE of ULASSO	1.000	4.436	**2.477**	6.307	3.040	6.340	3.844	6.689	4.129	7.671
Avg. AUC	**0.901**	0.891	**0.897**	0.885	0.895	0.886	0.893	0.885	0.892	0.873
AUC_{oracle}	**0.904**	0.904	0.904	0.904	0.904	0.904	0.904	0.904	0.904	0.904
Avg. TPR	**1.000**	0.982	**0.998**	0.894	0.984	–	–	–	–	–
Avg. FPR	**0.001**	0.196	**0.209**	0.022	0.034	–	–	–	–	–

Table 3.7: Results for $p = 20, \Sigma = \Sigma_\rho, q = 0.04$, and $\{\boldsymbol{\xi}_{[j]}\}_{j=1}^p \sim \mathcal{N}_1(3, 1)$. Comparison of ULASSO and all supervised estimators, as well as $\boldsymbol{\alpha}_0$, based on Emp. MSE, AUC, TPR and FPR. Shown also are the REs, based on Emp. MSE, of ULASSO w.r.t. all the estimators.

Criteria	ULASSO	Sup.Log.LASSO		Sup.Lin.LASSO		Sup.Log.MLE		Sup.Lin.OLS		$\boldsymbol{\alpha}_0$
		n_{300}	n_{500}	n_{300}	n_{500}	n_{300}	n_{500}	n_{300}	n_{500}	
Emp. MSE	0.031	0.077	0.045	0.111	0.055	0.112	0.068	0.117	0.073	0.113
RE of ULASSO	1.000	2.490	**1.458**	3.589	1.786	3.647	2.225	3.792	2.361	3.670
Avg. AUC	**0.898**	0.891	**0.896**	0.885	0.895	0.886	0.893	0.885	0.892	0.879
AUC_{oracle}	**0.904**	0.904	0.904	0.904	0.904	0.904	0.904	0.904	0.904	0.904
Avg. TPR	**1.000**	0.980	**0.997**	0.906	0.985	–	–	–	–	–
Avg. FPR	**0.095**	0.191	**0.195**	0.023	0.032	–	–	–	–	–

Table 3.8: Results for $p = 20, \Sigma = \Sigma_\rho, q = 0.04$, and $\{\boldsymbol{\xi}_{[j]}\}_{j=1}^p \sim$ Uniform $(2, 5)$. Comparison of ULASSO and all supervised estimators, as well as $\boldsymbol{\alpha}_0$, based on Emp. MSE, AUC, TPR and FPR. Shown also are the REs, based on Emp. MSE, of ULASSO w.r.t. all the estimators.

Criteria	ULASSO	Sup.Log.LASSO		Sup.Lin.LASSO		Sup.Log.MLE		Sup.Lin.OLS		$\boldsymbol{\alpha}_0$
		n_{300}	n_{500}	n_{300}	n_{500}	n_{300}	n_{500}	n_{300}	n_{500}	
Emp. MSE	0.020	0.083	0.045	0.117	0.055	0.115	0.068	0.122	0.073	0.122
RE of ULASSO	1.000	4.201	**2.252**	5.926	2.770	5.816	3.416	6.158	3.679	6.152
Avg. AUC	**0.900**	0.890	**0.896**	0.884	0.895	0.885	0.893	0.884	0.892	0.876
AUC_{oracle}	**0.903**	0.903	0.903	0.903	0.903	0.903	0.903	0.903	0.903	0.903
Avg. TPR	**1.000**	0.974	**0.998**	0.890	0.984	–	–	–	–	–
Avg. FPR	**0.000**	0.200	**0.200**	0.021	0.032	–	–	–	–	–

Table 3.9: Results for $p = 50, \Sigma = I_p, q = 0.02$, and $\{\boldsymbol{\xi}_{[j]}\}_{j=1}^{p} \sim \mathcal{N}_1(3,1)$. Comparison of ULASSO and all supervised estimators, as well as $\boldsymbol{\alpha}_0$, based on Emp. MSE, AUC, TPR and FPR. Shown also are the REs, based on Emp. MSE, of ULASSO w.r.t. all the estimators.

Criteria	ULASSO	Sup.Log.LASSO		Sup.Lin.LASSO		Sup.Log.MLE		Sup.Lin.OLS		$\boldsymbol{\alpha}_0$
		n_{300}	n_{500}	n_{300}	n_{500}	n_{300}	n_{500}	n_{300}	n_{500}	
Emp. MSE	0.022	0.149	0.068	0.353	0.114	0.197	0.107	0.193	0.112	0.148
RE of ULASSO	1.000	6.889	**3.157**	16.357	5.279	9.126	4.972	8.917	5.207	6.875
Avg. AUC	**0.912**	0.882	**0.901**	0.824	0.890	0.870	0.891	0.871	0.890	0.882
AUC_{oracle}	**0.917**	0.917	0.917	0.917	0.917	0.917	0.917	0.917	0.917	0.917
Avg. TPR	**1.000**	0.891	**0.984**	0.612	0.898	–	–	–	–	–
Avg. FPR	**0.001**	0.119	**0.154**	0.012	0.026	–	–	–	–	–

Table 3.10: Results for $p = 50, \Sigma = I_p, q = 0.02$, and $\{\boldsymbol{\xi}_{[j]}\}_{j=1}^{p} \sim$ Uniform $(2,5)$. Comparison of ULASSO and all supervised estimators, as well as $\boldsymbol{\alpha}_0$, based on Emp. MSE, AUC, TPR and FPR. Shown also are the REs, based on Emp. MSE, of ULASSO w.r.t. all the estimators.

Criteria	ULASSO	Sup.Log.LASSO		Sup.Lin.LASSO		Sup.Log.MLE		Sup.Lin.OLS		$\boldsymbol{\alpha}_0$
		n_{300}	n_{500}	n_{300}	n_{500}	n_{300}	n_{500}	n_{300}	n_{500}	
Emp. MSE	0.024	0.145	0.068	0.336	0.111	0.193	0.107	0.188	0.112	0.204
RE of ULASSO	1.000	6.104	**2.838**	14.149	4.667	8.132	4.490	7.918	4.707	8.576
Avg. AUC	**0.911**	0.882	**0.901**	0.831	0.890	0.871	0.891	0.872	0.890	0.868
AUC_{oracle}	**0.917**	0.917	0.917	0.917	0.917	0.917	0.917	0.917	0.917	0.917
Avg. TPR	**1.000**	0.898	**0.985**	0.630	0.902	–	–	–	–	–
Avg. FPR	**0.001**	0.119	**0.149**	0.012	0.027	–	–	–	–	–

Table 3.11: Results for $p = 50, \Sigma = I_p, q = 0.04$, and $\{\boldsymbol{\xi}_{[j]}\}_{j=1}^{p} \sim \mathcal{N}_1(3,1)$. Comparison of ULASSO and all supervised estimators, as well as $\boldsymbol{\alpha}_0$, based on Emp. MSE, AUC, TPR and FPR. Shown also are the REs, based on Emp. MSE, of ULASSO w.r.t. all the estimators.

Criteria	ULASSO	Sup.Log.LASSO		Sup.Lin.LASSO		Sup.Log.MLE		Sup.Lin.OLS		$\boldsymbol{\alpha}_0$
		n_{300}	n_{500}	n_{300}	n_{500}	n_{300}	n_{500}	n_{300}	n_{500}	
Emp. MSE	0.023	0.140	0.067	0.331	0.110	0.195	0.106	0.188	0.111	0.160
RE of ULASSO	1.000	6.067	**2.894**	14.406	4.782	8.480	4.605	8.166	4.834	6.957
Avg. AUC	**0.912**	0.884	**0.901**	0.833	0.891	0.871	0.892	0.872	0.890	0.879
AUC_{oracle}	**0.917**	0.917	0.917	0.917	0.917	0.917	0.917	0.917	0.917	0.917
Avg. TPR	**1.000**	0.904	**0.982**	0.635	0.902	–	–	–	–	–
Avg. FPR	**0.006**	0.117	**0.147**	0.011	0.027	–	–	–	–	–

Table 3.12: Results for $p = 50, \Sigma = I_p, q = 0.04$, and $\{\boldsymbol{\xi}_{[j]}\}_{j=1}^{p} \sim$ Uniform $(2, 5)$. Comparison of ULASSO and all supervised estimators, as well as $\boldsymbol{\alpha}_0$, based on Emp. MSE, AUC, TPR and FPR. Shown also are the REs, based on Emp. MSE, of ULASSO w.r.t. all the estimators.

Criteria	ULASSO	Sup.Log.LASSO		Sup.Lin.LASSO		Sup.Log.MLE		Sup.Lin.OLS		$\boldsymbol{\alpha}_0$
		n_{300}	n_{500}	n_{300}	n_{500}	n_{300}	n_{500}	n_{300}	n_{500}	
Emp. MSE	0.025	0.148	0.066	0.342	0.114	0.198	0.106	0.190	0.112	0.209
RE of ULASSO	1.000	5.947	**2.665**	13.711	4.557	7.930	4.267	7.632	4.510	8.391
Avg. AUC	**0.911**	0.882	**0.901**	0.827	0.890	0.870	0.892	0.872	0.890	0.867
AUC_{oracle}	**0.917**	0.917	0.917	0.917	0.917	0.917	0.917	0.917	0.917	0.917
Avg. TPR	**1.000**	0.894	**0.985**	0.632	0.899	–	–	–	–	–
Avg. FPR	**0.003**	0.113	**0.156**	0.011	0.026	–	–	–	–	–

Table 3.13: Results for $p = 50, \Sigma = \Sigma_\rho, q = 0.02$, and $\{\boldsymbol{\xi}_{[j]}\}_{j=1}^{p} \sim \mathcal{N}_1(3, 1)$. Comparison of ULASSO and all supervised estimators, as well as $\boldsymbol{\alpha}_0$, based on Emp. MSE, AUC, TPR and FPR. Shown also are the REs, based on Emp. MSE, of ULASSO w.r.t. all the estimators.

Criteria	ULASSO	Sup.Log.LASSO		Sup.Lin.LASSO		Sup.Log.MLE		Sup.Lin.OLS		$\boldsymbol{\alpha}_0$
		n_{300}	n_{500}	n_{300}	n_{500}	n_{300}	n_{500}	n_{300}	n_{500}	
Emp. MSE	0.011	0.131	0.068	0.248	0.104	0.250	0.136	0.244	0.148	0.192
RE of ULASSO	1.000	11.634	**5.985**	21.970	9.248	22.164	12.057	21.634	13.151	17.004
Avg. AUC	**0.935**	0.914	**0.926**	0.890	0.918	0.894	0.914	0.895	0.912	0.889
AUC_{oracle}	**0.937**	0.937	0.937	0.937	0.937	0.937	0.937	0.937	0.937	0.937
Avg. TPR	**1.000**	0.931	**0.987**	0.741	0.925	–	–	–	–	–
Avg. FPR	**0.000**	0.101	**0.122**	0.009	0.015	–	–	–	–	–

Table 3.14: Results for $p = 50, \Sigma = \Sigma_\rho, q = 0.02$, and $\{\boldsymbol{\xi}_{[j]}\}_{j=1}^{p} \sim$ Uniform $(2, 5)$. Comparison of ULASSO and all supervised estimators, as well as $\boldsymbol{\alpha}_0$, based on Emp. MSE, AUC, TPR and FPR. Shown also are the REs, based on Emp. MSE, of ULASSO w.r.t. all the estimators.

Criteria	ULASSO	Sup.Log.LASSO		Sup.Lin.LASSO		Sup.Log.MLE		Sup.Lin.OLS		$\boldsymbol{\alpha}_0$
		n_{300}	n_{500}	n_{300}	n_{500}	n_{300}	n_{500}	n_{300}	n_{500}	
Emp. MSE	0.026	0.132	0.066	0.256	0.102	0.253	0.134	0.242	0.147	0.181
RE of ULASSO	1.000	5.140	**2.566**	9.947	3.951	9.817	5.193	9.393	5.686	7.028
Avg. AUC	**0.932**	0.914	**0.926**	0.888	0.919	0.893	0.914	0.895	0.912	0.892
AUC_{oracle}	**0.937**	0.937	0.937	0.937	0.937	0.937	0.937	0.937	0.937	0.937
Avg. TPR	**1.000**	0.933	**0.990**	0.738	0.934	–	–	–	–	–
Avg. FPR	**0.000**	0.108	**0.129**	0.009	0.017	–	–	–	–	–

Table 3.15: Results for $p = 50, \Sigma = \Sigma_\rho, q = 0.04$, and $\{\boldsymbol{\xi}_{[j]}\}_{j=1}^{p} \sim \mathcal{N}_1(3,1)$. Comparison of ULASSO and all supervised estimators, as well as $\boldsymbol{\alpha}_0$, based on Emp. MSE, AUC, TPR and FPR. Shown also are the REs, based on Emp. MSE, of ULASSO w.r.t. all the estimators.

Criteria	ULASSO	Sup.Log.LASSO		Sup.Lin.LASSO		Sup.Log.MLE		Sup.Lin.OLS		$\boldsymbol{\alpha}_0$
		n_{300}	n_{500}	n_{300}	n_{500}	n_{300}	n_{500}	n_{300}	n_{500}	
Emp. MSE	0.026	0.133	0.067	0.256	0.103	0.250	0.134	0.243	0.146	0.165
RE of ULASSO	1.000	5.138	**2.569**	9.896	3.983	9.661	5.186	9.381	5.638	6.377
Avg. AUC	**0.932**	0.914	**0.926**	0.888	0.918	0.894	0.914	0.895	0.912	0.896
AUC_{oracle}	**0.937**	0.937	0.937	0.937	0.937	0.937	0.937	0.937	0.937	0.937
Avg. TPR	**1.000**	0.933	**0.989**	0.728	0.933	–	–	–	–	–
Avg. FPR	**0.016**	0.106	**0.126**	0.006	0.016	–	–	–	–	–

Table 3.16: Results for $p = 50, \Sigma = \Sigma_\rho, q = 0.04$, and $\{\boldsymbol{\xi}_{[j]}\}_{j=1}^{p} \sim$ Uniform $(2,5)$. Comparison of ULASSO and all supervised estimators, as well as $\boldsymbol{\alpha}_0$, based on Emp. MSE, AUC, TPR and FPR. Shown also are the REs, based on Emp. MSE, of ULASSO w.r.t. all the estimators.

Criteria	ULASSO	Sup.Log.LASSO		Sup.Lin.LASSO		Sup.Log.MLE		Sup.Lin.OLS		$\boldsymbol{\alpha}_0$
		n_{300}	n_{500}	n_{300}	n_{500}	n_{300}	n_{500}	n_{300}	n_{500}	
Emp. MSE	0.019	0.135	0.066	0.252	0.100	0.252	0.132	0.241	0.144	0.206
RE of ULASSO	1.000	7.045	**3.423**	13.156	5.226	13.179	6.922	12.595	7.527	10.789
Avg. AUC	**0.934**	0.913	**0.926**	0.890	0.919	0.894	0.915	0.896	0.913	0.885
AUC_{oracle}	**0.937**	0.937	0.937	0.937	0.937	0.937	0.937	0.937	0.937	0.937
Avg. TPR	**1.000**	0.932	**0.990**	0.740	0.934	–	–	–	–	–
Avg. FPR	**0.001**	0.106	**0.129**	0.009	0.016	–	–	–	–	–

Overall, as is evident from all the results presented in tables 3.1-3.16, the performance of the ULASSO estimator seems to be quite satisfactory in its own right w.r.t. all the criteria we have considered, and also seems to be fairly robust to the choice of q as well as the underlying correlation structure of $\mathbf{X}$. Among the supervised estimators considered, the Sup.Log.LASSO estimator seems to have the best performance throughout, which is understandable given that apart from using the true labels, it also exploits the knowledge of the true link as well as the sparsity of $\boldsymbol{\beta}_0$. However, even w.r.t. the Sup.Log.LASSO estimator at both sample sizes n_{300} and n_{500}, the ULASSO seems to be significantly more efficient in almost all cases. Further, its prediction performance, as measured by the AUC, seems to be satisfactorily close to the gold-standard AUC_{oracle} measure, and is in fact, uniformly higher than those achieved

by the supervised estimators over all the settings. The variable selection performance of ULASSO in terms of the TPR and FPR also seems to be near-perfect, especially for the TPR, over all the cases, and is again uniformly superior to those achieved by the supervised estimators. Lastly, the performance of $\boldsymbol{\alpha}_0$, w.r.t. all the criteria considered, is clearly seen to be significantly worse over all cases than those of the ULASSO as well as most of the supervised estimators, thereby indicating that while it is 'close' to $\boldsymbol{\beta}_0$, it is not close enough (and sparse enough) to be considered a reasonable estimator of the $\boldsymbol{\beta}_0$ direction.

3.5.2 Application to EMR Data

We applied our proposed method to an EMR study of rheumatoid arthritis (RA), a systemic autoimmune (AI) disease, conducted at the Partners HealthCare. Further details on this study can be found in Liao et al. (2010, 2013). The study cohort consists of 44014 patients, and the binary outcome of interest in this case was a disease phenotype defined as clinically confirmed diagnosis of RA. The primary goal was to understand and model the disease risk of RA based on several relevant clinical variables, including RA biomarkers, standard medications for RA, as well other relevant AI diseases and/or clinical conditions known to be closely related to RA, rich information for all of which were available through the data for a large number of patients. However, the availability of gold standard outcomes was limited as it required logistically prohibitive manual chart review by the physician. A labeled training data was therefore only available for 500 patients, wherein observations for the gold standard outcome were obtained through manual chart review by two expert rheumatologists. We used this training data to implement standard supervised procedures, while we considered the rest of the massive unlabeled data available and implemented our proposed ULASSO estimator under the surrogate aided unsupervised learning framework considered herein. The surrogate we used for this purpose was a variable called RA-w which corresponds to the total count of ICD9 diagnostic codes for RA for a patient taken at least a week apart. It is natural to expect that when RA-w assumes too high or too low values, the patient is very likely to be

diseased or healthy respectively. Based on the full available data, with RA-w notably being a count variable, a choice of (lower, upper) cut-offs given by: (1 %, 99%) and (2.5 %, 97.5%) for RA-w turned out to correspond to (0, 45) and (0, 70) codes respectively, and based on the training data, the corresponding empirical misclassification errors turned out to be 2/81, and 5/95 respectively, both of which were therefore quite reasonably small.

In order to model the disease risk of RA, we related it to a set of 37 covariates altogether available through the dataset, which included: (i) age, gender, (ii) counts of ICD9 codes for other related AI diseases like psoriatic arthritis (PsA), juvenile rheumatoid arthritis (JRA) and systemic lupus erythometasus (SLE), denoted in our results as PsA-r, JRA-r and SLE-r respectively, (iii) counts of ICD9 codes for PsA, JRA and SLE taken at least a week apart, denoted in our results as PsA-w, JRA-w and SLE-w respectively, (iii) counts of mentions of PsA, JRA and SLE in the physicians' notes extracted via natural language processing (NLP), denoted in our results as PsA-nlp, JRA-nlp and SLE-nlp respectively, (iv) combined counts of normalized ICD9 codes, at least a week apart, for RA, and NLP extracted mentions of RA from physicians' notes, denoted in our results as RA-w-nlp, (v) codified test results and/or NLP extracted mentions of positivity for standard RA biomarkers including rheumatoid factor (RF), anti-cyclic citrullinated polypeptide (anti-CCP) and anti-tumor necrosis factors (anti-TNF) that are routinely checked for RA patients to assess the disease progression, (vi) counts of codified and/or NLP extracted mentions of methotrexate (a frequently used medication for RA), seropositivity, erosion (radiological evidence of bone erosion), as well as several other standard medications and/or relevant clinical conditions that are known to be related to RA, including Enb, Hum, Rem, Ore, Rit, Anak, Sulf, Aza, Plaq, Arava, Pen, Gld, Neo, Facts, PM, DM, PMR, Spond, and other medications (other meds.). A detailed glossary of the abbreviations used above, as well as further explanations regarding the clinical significance of these variables can be found in Liao et al. (2010, 2013). All the count/binary variables were further log-transformed as: $x \to \log(1+x)$, to increase stability of the model fitting. In order to ensure comparability of the point estimates for the regression coefficients

across all the predictors, especially since the estimates would be based on sparsity based estimators, all the covariates were further standardized to have unit variance w.r.t. the full data, and all our results are reported in this standardized scale for the covariates.

Based on these 37 covariates, and the RA-w surrogate, we implemented our ULASSO estimator using the unlabeled data with two choices of q given by: $q = 0.02$ ($n_q = 4375$) and $q = 0.05$ ($n_q = 5040$), and further constructed another ULASSO estimator, denoted in our results as 'combined', wherein we appropriately combined the two estimators for $q = 0.02$ and $q = 0.05$, by averaging their normalized versions. Based on the available labeled data and these 37 covariates, we also implemented the supervised (sup.) logistic LASSO and Adaptive LASSO (ALASSO) estimators. The tuning parameter for ULASSO was selected using the BIC criteria in (3.26), while those for the sup. logistic LASSO and ALASSO estimators were selected using the BIC criteria based on the logistic loss. We also computed the $\boldsymbol{\alpha}_0$ estimator using poisson regression of the surrogate RA-w, a count variable, w.r.t. the 37 covariates based on the full available data. All the estimators obtained were further normalized to have unit L_2 norm. For the ALASSO estimator, we also computed bootstrap based estimates of its standard errors, denoted in our results as boot.sd$_{alasso}$, using 500 bootstrap samples, for reference and also to get a reasonable idea about the significance of the point estimates obtained. Further, in order to examine the prediction/classification performance of all the above estimators, we also computed estimates of their corresponding AUC measures based on the labeled data. For the supervised logistic LASSO and ALASSO estimators, this was computed by additionally using $\mathbb{K}-$fold cross-validation with $\mathbb{K} = 5$, to avoid over-fitted estimates of the AUC. The results for the data analysis are presented in table 3.17.

Overall, the results in table 3.17 seem to be quite satisfactory. First of all, the subset of predictors selected by the various ULASSO estimators, as well as the corresponding point estimates of the regression coefficients for the selected variables, are fairly close to each other, which therefore highlights, based on real data, the robustness of ULASSO to the choice of q as long as it is reasonable. At the same time, the ULASSO for $q = 0.05$, and hence the

Table 3.17: Coordinate-wise comparison of the ULASSO (at $q = 0.02, 0.05$, and their combination) to the supervised logistic LASSO and Adaptive LASSO estimators, as well as to the $\boldsymbol{\alpha}_0$ estimator, for the data example. Shown also are the bootstrap based SE estimates for the Adaptive LASSO estimator, as well as the AUC estimates for all the estimators.

Predictors	ULASSO			Sup. Logistic Estimators			$\boldsymbol{\alpha}_0$
	$q = 0.02$	$q = 0.05$	Combined	LASSO	ALASSO	Boot.sd$_{alasso}$	
Age	0	0	0	0	0	0.123	0.057
Gender	0	0	0	0	0	0.026	-0.012
PsA-r	0	0	0	-0.021	0	0.052	-0.010
JRA-r	0	0	0	-0.078	0	0.064	0.214
SLE-r	0	0	0	0	0	0.076	0.161
PsA-w	0	0	0	0	0	0.035	0.053
JRA-w	0	0	0	0	0	0.089	-0.259
SLE-w	0	0	0	0	0	0.090	-0.090
RA-w-nlp	0.808	0.886	0.847	0.854	0.847	0.144	0.690
PsA-nlp	0	0	0	0	0	0.093	-0.036
JRA-nlp	0	0	0	-0.067	-0.123	0.098	0.016
SLE-nlp	0	0	0	0	0	0.075	-0.094
Methotrexate	0	0.117	0.058	0.211	0.216	0.135	0.124
Anti-TNF	0	0.011	0.005	0.137	0.082	0.103	0.017
Enb	0.126	0.066	0.096	0	0	0.041	0.066
Hum	0	0	0	0	0	0.034	-0.001
Rem	0.043	0	0.022	0	0	0.048	0.010
Ore	0	0	0	0	0	0.034	0.007
Rit	0	0	0	0	0	0.019	-0.036
Anak	0	0	0	0	0	0.025	0.006
Sulf	0	0	0	0	0	0.054	0.060
Aza	0	0	0	0	0	0.042	-0.033
Plaq	0	0	0	0	0	0.016	0.062
Arava	0.252	0.103	0.177	0.015	0	0.062	0.051
Pen	0	0	0	0	0	0.061	-0.002
Gld	0.103	0.033	0.068	0	0	0.051	-0.012
Neo	0	0	0	0	0	0.034	-0.030
Other Meds.	0	0	0	0	0	0.057	0.023
Anti-CCP	0	0	0	0.040	0	0.090	0.007
RF	0	0	0	0.078	0.064	0.097	0.086
Erosion	0.467	0.402	0.435	0.282	0.291	0.138	0.061
Seropositive	0.177	0.053	0.115	0.331	0.354	0.096	0.030
Facts	0.082	0.145	0.114	0	0	0.022	0.559
Spond	0	0	0	0	0	0.029	0.022
PM	0	0	0	0	0	0.008	0.013
DM	0	0	0	0	0	0.006	-0.006
PMR	0	0	0	0	0	0.026	0.008
Est. AUC	**0.945**	**0.943**	**0.945**	**0.950**	**0.950**	–	**0.921**

'combined' ULASSO, does select one or two more clinically relevant variables, including anti-TNF and methotrexate, thereby indicating the potential utility, at least in this case, of considering multiple choices of q for constructing ULASSO, followed by combining the estimators appropriately. Moreover, the subset of predictors selected by the various ULASSO estimators, as well as the corresponding point estimates, all seem to be reasonably close enough, in general, to those obtained for the sup. logistic LASSO/ALASSO estimators. There are indeed a few disparities, in terms of the selected subsets, among the ULASSO and the sup. logistic LASSO/ALASSO estimators. However, considering the standard error estimates for the ALASSO estimator in those disparate coordinates, the estimates are unlikely to be significant, and therefore these disparities are perhaps ignorable. In terms of prediction/classification accuracy, all the ULASSO estimators seem to have quite satisfactory performance in their own right, with estimated AUC measures close to 0.95, and moreover, are nearly similar to those for the sup. logistic LASSO/ALASSO estimators requiring as many as 500 labels. Finally, the performance of the $\boldsymbol{\alpha}_0$ estimator, both in terms of estimation, as well as prediction based on the AUC measure, seems to be substantially worse than those of all the ULASSO as well as the sup. logistic LASSO/ALASSO estimators, thereby indicating its unsuitability as an estimator of the $\boldsymbol{\beta}_0$ direction in this case.

3.6 Discussion

We have considered in this paper a fairly unique surrogate aided unsupervised learning problem for binary outcomes under a SIM set-up, and proposed a penalized estimation procedure for signal recovery under some structured sparsity assumptions. We have provided precise (and fairly sharp) finite sample performance bounds for our estimator establishing its convergence rates, among other implications, as well as presented extensive simulation studies and applications to real data all of which seem to yield fairly satisfactory results. The performance of the estimator also seems to be quite robust to the choice of q, as long as it is chosen to be reasonably small and the corresponding π_q is small enough as well.

This also indicates that the estimator can perhaps be further combined appropriately over multiple choices of q, as shown in our data example, leading to a more stable and efficient estimator. As mentioned earlier, while we have focussed here only on the standard L_1 norm as the choice of our penalty/regularizer for simplicity, other sparsity friendly penalties like weighted versions of the L_1 norm, including the Adaptive LASSO penalty in particular, can also be considered. Moreover, we have focussed here on a setting involving the availability of one surrogate. The proposed procedure can also be extended to settings where we have multiple such surrogates available, each satisfying the desired assumptions, in which case the estimators of the $\boldsymbol{\beta}_0$ direction obtained from each of them (and possibly over several choices of q) can be further combined effectively to give a more stable and efficient estimator.

Lastly, apart from the recovery of $\boldsymbol{\beta}_0$, another key consequence of our sparsity based approach is its capability to perform variable selection. This can be extremely useful in subsequent analyses based on an actual training data with Y observed, wherein only the selected variables may be used, and this can significantly improve the efficiency/accuracy of the final classification rule. To the best of our knowledge, relatively little work has been done on problems of this sort which are highly recent, and our approach as well as the results obtained in this paper are quite novel. The closest connections to this work are some recent but sporadic works in one-bit compressed sensing with 'adversarial' bit flips/corruptions (Laska et al., 2009; Plan and Vershynin, 2013a,b; Chen et al., 2013; Jacques et al., 2013; Li, 2013; Natarajan et al., 2013; Feng et al., 2014), as well as more classical problems considered in the measurement error and misclassification literature (Carroll, 1998; Carroll et al., 2006; Buonaccorsi, 2010). However, in both cases, the problem settings as well as the approach and the necessary assumptions required therein are quite different, and their connections to our approach in this paper are remote at best. A key feature of our problem is that we don't observe Y at all, and instead we use the surrogacy of S to 'synthesize' our outcomes, something which we don't believe has been considered anywhere in the relevant literature.

Appendix A

Proofs of All Results in Chapter 1

A.1 Preliminaries

The following Lemmas A.1-A.3 would be useful in the proofs of the main theorems.

Lemma A.1. *Let $\mathbf{Z} \in \mathbb{R}^l$ be any random vector and $\mathbf{g}(\mathbf{Z}) \in \mathbb{R}^d$ be any measurable function of $\mathbf{Z}$, where l and d are fixed. Let $\mathbb{S}_n = \{\mathbf{Z}_i\}_{i=1}^n \perp\!\!\!\perp \mathbb{S}_m = \{\mathbf{Z}_j\}_{j=1}^m$ be two random samples of n and m i.i.d. observations of $\mathbf{Z}$ respectively. Let $\widehat{\mathbf{g}}_n(.)$ be any estimator of $\mathbf{g}(.)$ based on $\mathbb{S}_n$ such that the random sequence: $\widehat{T}_n \equiv sup_{\mathbf{z}\in\boldsymbol{\chi}}\|\widehat{\mathbf{g}}_n(\mathbf{z})\|$ is $O_p(1)$, where $\boldsymbol{\chi} \subseteq \mathbb{R}^l$ denotes the support of $\mathbf{Z}$. Let $\widehat{\mathbf{G}}_{n,m}$ denote the (double) random sequence: $m^{-1}\sum_{\mathbf{Z}_j\in\mathbb{S}_m}\widehat{\mathbf{g}}_n(\mathbf{Z}_j)$, and let $\overline{\mathbf{G}}_n$ denote the random sequence: $\mathbb{E}_{\mathbb{S}_m}(\widehat{\mathbf{G}}_{n,m}) = \mathbb{E}_{\mathbf{Z}}\{\widehat{\mathbf{g}}_n(\mathbf{Z})\}$, where $\mathbb{E}_{\mathbf{Z}}(.)$ denotes expectation w.r.t. $\mathbf{Z} \in \mathbb{S}_m \perp\!\!\!\perp \mathbb{S}_n$, and all expectations involved are assumed to be finite almost surely (a.s.) $[\mathbb{S}_n]$ $\forall$ n. Then: (a) $\mathbf{G}_{n,m} - \overline{\mathbf{G}}_n = O_p(m^{-\frac{1}{2}})$, and (b) as long as $g(.)$ has finite 2^{nd} moments, $m^{-1}\sum_{\mathbf{Z}_j\in\mathbb{S}_m}\mathbf{g}(\mathbf{Z}_j) - \mathbb{E}_{\mathbf{Z}}\{\mathbf{g}(\mathbf{Z})\} = O_p(m^{-\frac{1}{2}})$.*

The next two lemmas would be useful in the proof of Theorem 1.4. They may also be of more general use in other applications that involve controlling empirical process terms indexed by KS estimators. Suppose that our basic assumption (c) holds, and consider the KS framework introduced in section 1.5. Let $l_{\mathbf{P}_r}(\mathbf{w}) = m_{\mathbf{P}_r}(\mathbf{w})f_{\mathbf{P}_r}(\mathbf{w})$ and $\widetilde{\varphi}^{(\varrho)}_{\mathbf{P}_r}(\mathbf{w}) = (nh^r)^{-1}\sum_{i=1}^n K_h(\mathbf{w}, \mathbf{P}_r'\mathbf{X}_i)Y_i^{\varrho}$, for $\varrho = 0, 1$. Let $\widetilde{f}_{\mathbf{P}_r} = \widetilde{\varphi}^{(0)}_{\mathbf{P}_r}$, $\widetilde{l}_{\mathbf{P}_r} = \widetilde{\varphi}^{(1)}_{\mathbf{P}_r}$, $\widetilde{m}_{\mathbf{P}_r} = \widetilde{l}_{\mathbf{P}_r}/\widetilde{f}_{\mathbf{P}_r}$, $\varphi^{(0)}_{\mathbf{P}_r} = f_{\mathbf{P}_r}$ and $\varphi^{(1)}_{\mathbf{P}_r} = l_{\mathbf{P}_r}$. Let $\varphi^{(\varrho)}(\mathbf{x}; \mathbf{P}_r) = \varphi^{(\varrho)}_{\mathbf{P}_r}(\mathbf{P}_r'\mathbf{x})$ and $\widetilde{\varphi}^{(\varrho)}(\mathbf{x}; \mathbf{P}_r) = \widetilde{\varphi}^{(\varrho)}_{\mathbf{P}_r}(\mathbf{P}_r'\mathbf{x})$ $\forall$ $\varrho = 0, 1$. Let $\widetilde{f} = \widetilde{\varphi}^{(0)}$, $\widetilde{l} = \widetilde{\varphi}^{(1)}$ and $\widetilde{m} = \widetilde{l}/\widetilde{f}$. Now, let $\mathbb{P}_n$ denote the empirical probability measure on $\mathbb{R}^p$

based on $\{\mathbf{X}_i\}_{i=1}^n$, and for any measurable function $\boldsymbol{\gamma}(.)$ (possibly vector valued) of $\mathbf{X}$, let $\mathbb{G}_n^*(\boldsymbol{\gamma}) = n^{\frac{1}{2}} \int \boldsymbol{\gamma}(\mathbf{x})(\mathbb{P}_n - \mathbb{P}_{\mathbf{X}})(d\mathbf{x})$.

Lemma A.2. *Consider the set-up introduced above. For any fixed integer $d \geq 1$, let $\boldsymbol{\lambda}(.)$ be any $\mathbb{R}^d$-valued measurable function of $\mathbf{X}$ that is bounded a.s. $[\mathbb{P}_{\mathbf{X}}]$. Define: $b_n^{(1)} = n^{-\frac{1}{2}}h^{-r} + h^q$ and $a_{n,2} = (\log n)^{\frac{1}{2}}(nh^r)^{-\frac{1}{2}} + h^q$. Assume $b_n^{(1)} = o(1)$ for (A.1) and $n^{\frac{1}{2}}a_{n,2}^2 = o(1)$ for (A.2) below. Then, under Assumption 1.2 (i)-(v), and $\forall\ \varrho \in \{0,1\}$,*

$$\mathbb{G}_n^*[\boldsymbol{\lambda}(.)\{\widetilde{\varphi}^{(\varrho)}(.;\mathbf{P}_r) - \varphi^{(\varrho)}(.;\mathbf{P}_r)\}] = O_p(b_n^{(1)}) = o_p(1), \quad \textit{and} \tag{A.1}$$

$$\mathbb{G}_n^*[\boldsymbol{\lambda}(.)\{\widetilde{m}(.;\mathbf{P}_r) - m(.;\mathbf{P}_r)\}] = O_p(n^{\frac{1}{2}}a_{n,2}^2) = o_p(1). \tag{A.2}$$

Let $\widehat{\varphi}^{(\varrho)}(\mathbf{x};\widehat{\mathbf{P}}_r) = (nh^r)^{-1}\sum_{i=1}^n K_h(\widehat{\mathbf{P}}_r'\mathbf{x}, \widehat{\mathbf{P}}_r'\mathbf{X}_i)Y_i^{\varrho}\ \forall\ \varrho \in \{0,1\}$, where $\widehat{\mathbf{P}}_r$ is as in section 1.4.2 and all other notations are the same as in the set-up of Lemma A.2. Let $\widehat{f}(\mathbf{x};\widehat{\mathbf{P}}_r) = \widehat{\varphi}^{(0)}(\mathbf{x};\widehat{\mathbf{P}}_r)$ and $\widehat{l}(\mathbf{x};\widehat{\mathbf{P}}_r) = \widehat{\varphi}^{(1)}(\mathbf{x};\widehat{\mathbf{P}}_r)$. Then:

Lemma A.3. *Consider the set-up of Lemma A.2. Let $\widehat{\varphi}^{(\varrho)}(\mathbf{x};\widehat{\mathbf{P}}_r)$ be as above, and let $\boldsymbol{\lambda}(.)$ be as in Lemma A.2. Suppose $(\widehat{\mathbf{P}}_r - \mathbf{P}_r) = O_p(\alpha_n)$ for some $\alpha_n = o(1)$. Assume $b_n^{(2)} = o(1)$, where $b_n^{(2)} = \alpha_n + n^{-\frac{1}{2}}\alpha_n h^{-(r+1)} + n^{\frac{1}{2}}\alpha_n^2(h^{-2} + n^{-1}h^{-(r+2)})$. Then, under Assumption 1.2,*

$$\mathbb{G}_n^*[\boldsymbol{\lambda}(.)\{\widehat{\varphi}^{(\varrho)}(.;\widehat{\mathbf{P}}_r) - \widetilde{\varphi}^{(\varrho)}(.;\mathbf{P}_r)\}] = O_p(b_n^{(2)}) = o_p(1)\ \ \forall\ \varrho \in \{0,1\}. \tag{A.3}$$

A.1.1 Proof of Lemma A.1

Firstly, since d is fixed, it suffices to prove the result for any arbitrary scalar coordinate $\widehat{\mathbf{G}}_{n,m}^{(j)} \equiv \widehat{\mathcal{G}}_{n,m}$ (say) and $\overline{\mathbf{G}}_n^{(j)} \equiv \overline{\mathcal{G}}_n$ (say) of $\widehat{\mathbf{G}}_{n,m}$ and $\overline{\mathbf{G}}_n$ respectively, for any $j \in \{1,\ldots,d\}$. For any data $\mathbb{S}$ and $\mathbb{S}^*$, we let $\mathbb{P}_{\mathbb{S}}$ and $\mathbb{P}_{\mathbb{S},\mathbb{S}^*}$ denote the joint probability distributions of the observations in $\mathbb{S}$ and $(\mathbb{S},\mathbb{S}^*)$ respectively, $\mathbb{E}_{\mathbb{S}}(.)$ denote the expectation w.r.t $\mathbb{P}_{\mathbb{S}}$, and $\mathbb{P}_{\mathbb{S}\,|\,\mathbb{S}^*}$ denote the conditional probability distribution of the observations in $\mathbb{S}$ given $\mathbb{S}^*$.

To show that $\widehat{\mathcal{G}}_{n,m} - \overline{\mathcal{G}}_n = O_p(m^{-\frac{1}{2}})$, we first note that since $\mathbb{S}_n \perp\!\!\!\perp \mathbb{S}_m$,

$$\mathbb{P}_{\mathbb{S}_n,\mathbb{S}_m}\left(|\widehat{\mathcal{G}}_{n,m} - \overline{\mathcal{G}}_n| > m^{-\frac{1}{2}}t\right) \;=\; \mathbb{E}_{\mathbb{S}_n}\left\{\mathbb{P}_{\mathbb{S}_m}\left(|\widehat{\mathcal{G}}_{n,m} - \overline{\mathcal{G}}_n| > m^{-\frac{1}{2}}t \,\middle|\, \mathbb{S}_n\right)\right\},$$

for any $t > 0$. Now, conditional on $\mathbb{S}_n$, $\widehat{\mathbf{G}}_{n,m} - \overline{\mathbf{G}}_n$ is a centered average of $\{\widehat{\mathbf{g}}_n(\mathbf{Z}_j)\}_{j=1}^m$ which are i.i.d. and bounded by $\widehat{T}_n < \infty$ a.s. $[\mathbb{P}_{\mathbb{S}_n}]$ $\forall\, n$. Hence, applying Hoeffding's inequality, we have for any n and m,

$$\mathbb{P}_{\mathbb{S}_m}\left(|\widehat{\mathcal{G}}_{n,m} - \overline{\mathcal{G}}_n| > m^{-\frac{1}{2}}t \,\middle|\, \mathbb{S}_n\right) \;\leq\; 2\exp\left(-\frac{2m^2t^2}{4m^2\widehat{T}_n^2}\right) \quad \text{a.s. } [\mathbb{P}_{\mathbb{S}_n}]. \tag{A.4}$$

Now, since $\widehat{T}_n \geq 0$ is $O_p(1)$, we have: for any given $\epsilon > 0$, $\exists\ \delta(\epsilon) > 0$ such that: $\mathbb{P}_{\mathbb{S}_n}\{\widehat{T}_n > \delta(\epsilon)\} \leq \epsilon/4\ \forall\, n$. Let $\mathbb{A}(\epsilon)$ denote the event: $\{\widehat{T}_n > \delta(\epsilon)\}$ and let $\mathbb{A}^c(\epsilon)$ denote its complement. Then, using (A.4), we have: $\forall\, n$ and m,

$$\begin{aligned}
&\mathbb{P}_{\mathbb{S}_n,\mathbb{S}_m}\left(|\widehat{\mathcal{G}}_{n,m} - \overline{\mathcal{G}}_n| > m^{-\frac{1}{2}}t\right) \;\leq\; \mathbb{E}_{\mathbb{S}_n}\left\{2\exp\left(-\frac{2m^2t^2}{4m^2\widehat{T}_n^2}\right)\right\} \\
&= \mathbb{E}_{\mathbb{S}_n}\left\{2\exp\left(-\frac{t^2}{2\widehat{T}_n^2}\right)\right\} = \mathbb{E}_{\mathbb{S}_n}\left[2\exp\left(-\frac{t^2}{2\widehat{T}_n^2}\right)\left\{1_{\mathbb{A}^c(\epsilon)} + 1_{\mathbb{A}(\epsilon)}\right\}\right] \\
&\leq \left[2\exp\left\{-\frac{t^2}{2\delta^2(\epsilon)}\right\}\mathbb{P}_{\mathbb{S}_n}\left\{\mathbb{A}^c(\epsilon)\right\} + 2\,\mathbb{P}_{\mathbb{S}_n}\left\{\mathbb{A}(\epsilon)\right\}\right] \\
&\leq 2\exp\left\{-\frac{t^2}{2\delta^2(\epsilon)}\right\} + \frac{\epsilon}{2} \;\; \leq \frac{\epsilon}{2} + \frac{\epsilon}{2} = \epsilon \text{ (for some suitable choice of } t),
\end{aligned}$$

where the last step follows by choosing $t \equiv t_\epsilon$ to be any large enough t such that $\exp\{-t^2/2\delta^2(\epsilon)\} \leq \epsilon/4$. Such a choice of t_ϵ clearly exists. This establishes the first claim (a) in Lemma A.1. The second claim (b) in Lemma A.1 is a trivial consequence of the Central Limit Theorem (CLT). ∎

A.2 Proof of Theorem 1.1

To show Theorem 1.1, we first note that under Assumption 1.1 (i)-(v), and letting $a_n = (\log n)^{\frac{1}{2}}(nh^p)^{-\frac{1}{2}} + h^q$, the following holds:

$$\sup_{\mathbf{x}\in\mathcal{X}}|\widehat{m}(\mathbf{x}) - m(\mathbf{x})| = O_p(a_n) = \sup_{\mathbf{x}\in\mathcal{X}}|\widehat{f}(\mathbf{x}) - f(\mathbf{x})|. \tag{A.5}$$

(A.5) is a fairly standard result and we only provide a sketch of its proof as follows. Under Assumption 1.1 (ii)-(iii), using Theorem 2 of Hansen (2008), $\sup_{\mathbf{x}\in\mathcal{X}}|\widehat{l}(\mathbf{x}) - \mathbb{E}_{\mathcal{L}}\{\widehat{l}(\mathbf{x})\}| = O_p(a_n^*) = \sup_{\mathbf{x}\in\mathcal{X}}|\widehat{f}(\mathbf{x}) - \mathbb{E}_{\mathcal{L}}\{\widehat{f}(\mathbf{x})\}|$, where $a_n^* = (\log n)^{\frac{1}{2}}(nh^p)^{-\frac{1}{2}}$. Next, using standard arguments based on Taylor series expansions of $l(.)$ and $m(.)$ under their assumed smoothness, and noting that $K(.)$ is a q^{th} order kernel having finite q^{th} moments, we obtain:

$$\sup_{\mathbf{x}\in\mathcal{X}}|\mathbb{E}_{\mathcal{L}}\{\widehat{l}(\mathbf{x})\} - l(\mathbf{x})| = O(h^q) = \sup_{\mathbf{x}\in\mathcal{X}}|\mathbb{E}_{\mathcal{L}}\{\widehat{f}(\mathbf{x})\} - f(\mathbf{x})|.$$

Combining these two results, and the definitions of $m(.)$ and $\widehat{m}(.)$ along with Assumption 1.1 (iv), we have (A.5). Next, note that using (1.4), we have:

$$\begin{aligned}\Gamma_N(\widehat{\boldsymbol{\theta}}_{np} - \boldsymbol{\theta}_0) &= \mathbb{E}_{\mathcal{U}}[N^{-1}\sum_{j=1}^{N}\overrightarrow{\mathbf{X}}_j\{\widehat{m}(\mathbf{X}_j) - \overrightarrow{\mathbf{X}}_j'\boldsymbol{\theta}_0\}] + O_p(N^{-\frac{1}{2}}) \\ &= \mathbb{E}_{\mathbf{X}}[\overrightarrow{\mathbf{X}}\{\widehat{m}(\mathbf{X}) - m(\mathbf{X})\}] + O_p(N^{-\frac{1}{2}}),\end{aligned}$$

where the first step is due to Lemma A.1 (a) with $\sup_{\mathbf{x}\in\mathcal{X}}\|\overrightarrow{\mathbf{x}}\{\widehat{m}(\mathbf{x}) - \overrightarrow{\mathbf{x}}'\boldsymbol{\theta}_0\}\| \leq \sup_{\mathbf{x}\in\mathcal{X}}[\|\overrightarrow{\mathbf{x}}\|\{|\widehat{m}(\mathbf{x}) - m(\mathbf{x})| + |m(\mathbf{x}) - \overrightarrow{\mathbf{x}}'\boldsymbol{\theta}_0|\}] = O_p(1)$ due to (A.5) and the boundedness of $\mathbf{X}$ and $m(.)$, while the last step uses: $\mathbb{E}_{\mathbf{X}}[\overrightarrow{\mathbf{X}}\{m(\mathbf{X}) - \overrightarrow{\mathbf{X}}'\boldsymbol{\theta}_0\}] = \mathbf{0}$ which follows from the definitions of $\boldsymbol{\theta}_0$ and $m(.)$. It then follows further, using $\Gamma_N^{-1} = I_{(p+1)} + O_p(N^{-\frac{1}{2}})$, that

$$n^{\frac{1}{2}}(\widehat{\boldsymbol{\theta}}_{np} - \boldsymbol{\theta}_0) \;=\; n^{\frac{1}{2}}\,\mathbb{E}_{\mathbf{X}}[\overrightarrow{\mathbf{X}}\{\widehat{m}(\mathbf{X}) - m(\mathbf{X})\}] + O_p\left(\frac{n}{N}\right)^{\frac{1}{2}}.$$

Letting $\phi_n(\mathbf{X}) = (nh^p)^{-1}\sum_{i=1}^n K\{(\mathbf{X}-\mathbf{X}_i)/h\}\{Y_i - m(\mathbf{X})\}$, and expanding the first term in the above equation, we now obtain:

$$n^{\frac{1}{2}}\left(\widehat{\boldsymbol{\theta}}_{np} - \boldsymbol{\theta}_0\right) \;=\; \mathbf{T}_{n,1}^{(1)} + \mathbf{T}_{n,1}^{(2)} + O_p\left(\frac{n}{N}\right)^{\frac{1}{2}}, \tag{A.6}$$

where $\mathbf{T}_{n,1}^{(1)} = n^{\frac{1}{2}}\,\mathbb{E}_{\mathbf{X}}\{\overrightarrow{\mathbf{X}}\phi_n(\mathbf{X})/f(\mathbf{X})\}$ and

$$\begin{aligned}
\mathbf{T}_{n,1}^{(2)} &= n^{\frac{1}{2}}\,\mathbb{E}_{\mathbf{X}}\left[\overrightarrow{\mathbf{X}}\phi_n(\mathbf{X})\{\widehat{f}(\mathbf{X})^{-1} - f(\mathbf{X})^{-1}\}\right] \\
&= n^{\frac{1}{2}}\,\mathbb{E}_{\mathbf{X}}[\overrightarrow{\mathbf{X}}\left\{\widehat{m}(\mathbf{X}) - m(\mathbf{X})\right\}\left\{f(\mathbf{X}) - \widehat{f}(\mathbf{X})\right\}/f(\mathbf{X})] \\
&\leq n^{\frac{1}{2}}\,\sup_{\mathbf{x}\in\mathcal{X}}\left\{\|\overrightarrow{\mathbf{x}}\|\,|\widehat{m}(\mathbf{x}) - m(\mathbf{x})|\left|\widehat{f}(\mathbf{x})/f(\mathbf{x}) - 1\right|\right\} \;= O_p\left(n^{\frac{1}{2}}a_n^2\right),
\end{aligned} \tag{A.7}$$

where the last step in (A.7) follows from (A.5), Assumption 1.1 (iv) and the boundedness of $\mathbf{X}$. For $\mathbf{T}_{n,1}^{(1)}$, we have:

$$\begin{aligned}
\mathbf{T}_{n,1}^{(1)} &= n^{\frac{1}{2}}\int_{\mathcal{X}}\overrightarrow{\mathbf{x}}\phi_n(\mathbf{x})d\mathbf{x} = n^{-\frac{1}{2}}\sum_{i=1}^n\int_{\mathcal{X}}\overrightarrow{\mathbf{x}}h^{-p}K_h(\mathbf{x}-\mathbf{X}_i)\left\{Y_i - m(\mathbf{x})\right\}d\mathbf{x} \\
&= n^{\frac{1}{2}}\sum_{i=1}^n n^{-1}\int_{\mathcal{A}_{i,n}}\overrightarrow{(\mathbf{X}_i + h\boldsymbol{\psi}_i)}\,K\left(\boldsymbol{\psi}_i\right)\left\{Y_i - m(\mathbf{X}_i + h\boldsymbol{\psi}_i)\right\}d\boldsymbol{\psi}_i,
\end{aligned} \tag{A.8}$$

where $\boldsymbol{\psi}_i = (\mathbf{x}-\mathbf{X}_i)/h$ and $\mathcal{A}_{i,n} = \{\boldsymbol{\psi}_i \in \mathbb{R}^p : (\mathbf{X}_i + h\boldsymbol{\psi}_i) \in \mathcal{X}\}$. Now, since $K(.)$ is zero outside the bounded set $\mathcal{K}$, the i^{th} integral in (A.8) only runs over $(\mathcal{A}_{i,n} \cap \mathcal{K})$. Further, since $h = o(1)$, using Assumption 1.1 (vi), $\mathcal{A}_{i,n} \supseteq \mathcal{K}$ a.s. $[\mathbb{P}_{\mathcal{L}}]$ or, $(\mathcal{A}_{i,n} \cap \mathcal{K}) = \mathcal{K}$ a.s. $[\mathbb{P}_{\mathcal{L}}]$ $\forall$ $1 \leq i \leq n$ with n large enough. Thus, for large enough n, (A.8) can be written as:

$$\begin{aligned}
\mathbf{T}_{n,1}^{(1)} &= n^{-\frac{1}{2}}\sum_{i=1}^n\int_{\mathcal{K}}\overrightarrow{(\mathbf{X}_i + h\boldsymbol{\psi}_i)}\,K\left(\boldsymbol{\psi}_i\right)\left\{Y_i - m(\mathbf{X}_i + h\boldsymbol{\psi}_i)\right\}d\boldsymbol{\psi}_i \;\text{ a.s. } [\mathbb{P}_{\mathcal{L}}] \\
&= n^{\frac{1}{2}}\sum_{i=1}^n n^{-1}\left[\overrightarrow{\mathbf{X}}_i\left\{Y_i - m(\mathbf{X}_i)\right\} + O_p(h^q)\right] \qquad \text{(A.9)} \\
&= n^{-\frac{1}{2}}\sum_{i=1}^n\overrightarrow{\mathbf{X}}_i\left\{Y_i - m(\mathbf{X}_i)\right\} + O_p\left(n^{\frac{1}{2}}h^q\right), \qquad \text{(A.10)}
\end{aligned}$$

where (A.9), and hence (A.10), follows from standard arguments based on Taylor series expansions of $m(\mathbf{X}_i + h\boldsymbol{\psi}_i)$ around $m(\mathbf{X}_i)$ under the assumed smoothness of $m(.)$, and using the fact that $K(.)$ is a q^{th} order kernel. Combining (A.6), (A.7) and (A.10), and noting that under our assumptions, $(n^{\frac{1}{2}}a_n^2 + n^{\frac{1}{2}}h^q) = O\{n^{\frac{1}{2}}h^q + (\log n)(n^{\frac{1}{2}}h^p)^{-1}\}$, the result of Theorem 1.1 now follows. ∎

A.3 Proof of Theorem 1.2

Let $\Gamma_n = \frac{1}{n}\sum_{i=1}^n \overrightarrow{\mathbf{X}}_i \overrightarrow{\mathbf{X}}_i'$, and

$$\mathbf{T}_n^{(1)} = \frac{1}{n}\sum_{i=1}^n \overrightarrow{\mathbf{X}}_i \left\{Y_i - \mu(\mathbf{X}_i;\mathbf{P}_r)\right\}, \mathbf{T}_{n,\mathbb{K}}^{(2)} = \frac{1}{n}\sum_{k=1}^{\mathbb{K}}\sum_{i\in\mathcal{I}_k} \overrightarrow{\mathbf{X}}_i \widehat{\Delta}_k(\mathbf{X}_i;\mathbf{P}_r,\widehat{\mathbf{P}}_{r,k}).$$

Then, using (1.6)-(1.10), it is straightforward to see that:

$$\mathbb{E}[\overrightarrow{\mathbf{X}}\{Y - \mu(\mathbf{X};\mathbf{P}_r)\}] \equiv \mathbb{E}[\overrightarrow{\mathbf{X}}\{Y - m(\mathbf{X};\mathbf{P}_r) - \overrightarrow{\mathbf{X}}'\boldsymbol{\eta}_{\mathbf{P}_r}\}] = \mathbf{0}, \text{ and} \tag{A.11}$$

$$\Gamma_n\left(\widehat{\boldsymbol{\eta}}_{(\mathbf{P}_r,\mathbb{K})} - \boldsymbol{\eta}_{\mathbf{P}_r}\right) = \mathbf{T}_n^{(1)} - \mathbf{T}_{n,\mathbb{K}}^{(2)}. \tag{A.12}$$

Under (A.11), assumptions (i) and (c), it follows from Lemma A.1 (b) that $\mathbf{T}_n^{(1)} = O_p(n^{-\frac{1}{2}})$. Next, due to assumption (ii) and the boundedness of $\mathbf{X}$,

$$\|\mathbf{T}_{n,\mathbb{K}}^{(2)}\| \leq n^{-1}\sum_{k=1}^{\mathbb{K}}\sum_{i\in\mathcal{I}_k} \sup_{\mathbf{x}\in\mathcal{X}}\{\|\overrightarrow{\mathbf{x}}\|\,|\widehat{\Delta}_k(\mathbf{x};\mathbf{P}_r,\widehat{\mathbf{P}}_{r,k})|\} = O_p(c_{n_{\mathbb{K}}^-}).$$

Finally, under assumptions (c)-(d), we have: $\Gamma_n = I_{(p+1)} + O_p(n^{-\frac{1}{2}})$ using Lemma A.1 (b). Further, since Γ_n is invertible a.s., $\Gamma_n^{-1} = I_{(p+1)} + O_p(n^{-\frac{1}{2}})$. Using all these facts, we then have: $(\widehat{\boldsymbol{\eta}}_{(\mathbf{P}_r,\mathbb{K})} - \boldsymbol{\eta}_{\mathbf{P}_r}) = \Gamma_n^{-1}(\mathbf{T}_n^{(1)} - \mathbf{T}_{n,\mathbb{K}}^{(2)}) = \mathbf{T}_n^{(1)} - \mathbf{T}_{n,\mathbb{K}}^{(2)} + O_p\{n^{-\frac{1}{2}}(n^{-\frac{1}{2}} + c_{n_{\mathbb{K}}^-})\}$. Thus,

$$(\widehat{\boldsymbol{\eta}}_{(\mathbf{P}_r,\mathbb{K})} - \boldsymbol{\eta}_{\mathbf{P}_r}) = \mathbf{T}_n^{(1)} - \mathbf{T}_{n,\mathbb{K}}^{(2)} + O_p(n^{-1} + n^{-\frac{1}{2}}c_{n_{\mathbb{K}}^-}). \tag{A.13}$$

Let $\Gamma_N = N^{-1}\sum_{j=1}^{N}\overrightarrow{\mathbf{X}}_j\overrightarrow{\mathbf{X}}_j'$, $\mathbf{R}_N^{(1)} = N^{-1}\sum_{j=1}^{N}\overrightarrow{\mathbf{X}}_j\{\mu(\mathbf{X}_j;\mathbf{P}_r) - \overrightarrow{\mathbf{X}}_j'\boldsymbol{\theta}_0\}$ and $\widehat{\mathbf{R}}_{N,n}^{(\mathbb{K})} = N^{-1}\sum_{j=1}^{N}\overrightarrow{\mathbf{X}}_j\{\widehat{\mu}(\mathbf{X}_j;\widehat{\mathcal{P}}_{r,\mathbb{K}}) - \mu(\mathbf{X}_j;\mathbf{P}_r)\}$. Then, using (1.9), we have:

$$\Gamma_N(\widehat{\boldsymbol{\theta}}_{(\mathbf{P}_r,\mathbb{K})} - \boldsymbol{\theta}_0) = N^{-1}\sum_{j=1}^{N}\overrightarrow{\mathbf{X}}_j[\widehat{\mu}(\mathbf{X}_j;\widehat{\mathcal{P}}_{r,\mathbb{K}}) - \overrightarrow{\mathbf{X}}_j'\boldsymbol{\theta}_0] = \mathbf{R}_N^{(1)} + \widehat{\mathbf{R}}_{N,n}^{(\mathbb{K})}.$$

Next, using (1.6)-(1.10), we have: $\widehat{\mathbf{R}}_{N,n}^{(\mathbb{K})} = \Gamma_N(\widehat{\boldsymbol{\eta}}_{(\mathbf{P}_r,\mathbb{K})} - \boldsymbol{\eta}_{\mathbf{P}_r}) + \widehat{\mathbf{S}}_{N,n}^{(\mathbb{K})}$, where

$$\widehat{\mathbf{S}}_{N,n}^{(\mathbb{K})} = \mathbb{K}^{-1}\textstyle\sum_{k=1}^{\mathbb{K}}\{N^{-1}\sum_{j=1}^{N}\overrightarrow{\mathbf{X}}_j\widehat{\Delta}_k(\mathbf{X}_j;\mathbf{P}_r,\widehat{\mathbf{P}}_{r,k})\}.$$

Hence, we have: $\Gamma_N(\widehat{\boldsymbol{\theta}}_{(\mathbf{P}_r,\mathbb{K})} - \boldsymbol{\theta}_0) = \Gamma_N(\widehat{\boldsymbol{\eta}}_{(\mathbf{P}_r,\mathbb{K})} - \boldsymbol{\eta}_{\mathbf{P}_r}) + \mathbf{R}_N^{(1)} + \widehat{\mathbf{S}}_{N,n}^{(\mathbb{K})}$.

Now, under assumptions (i)-(ii) and (c)-(d), we have:

$$\text{(I)}\quad \textstyle\sum_{k=1}^{\mathbb{K}}\sup_{\mathbf{x}\in\mathcal{X}}\|\overrightarrow{\mathbf{x}}\widehat{\Delta}_k(\mathbf{x};\mathbf{P}_r,\widehat{\mathbf{P}}_{r,k})\| = O_p(1),$$

so that using Lemma A.1 (a), $\widehat{\mathbf{S}}_{N,n}^{(\mathbb{K})} = \mathbb{K}^{-1}\sum_{k=1}^{\mathbb{K}}\widehat{\mathbf{S}}_{n,k}^{*} + O_p(N^{-\frac{1}{2}})$, where $\widehat{\mathbf{S}}_{n,k}^{*} = \mathbb{E}_{\mathbf{X}}\{\overrightarrow{\mathbf{X}}\widehat{\Delta}_k(\mathbf{X};\mathbf{P}_r,\widehat{\mathbf{P}}_{r,k})\}\ \forall\ 1\le k\le\mathbb{K}$;

$$\text{(II)}\quad \mathbf{R}_N^{(1)} = \mathbb{E}[\overrightarrow{\mathbf{X}}\{\mu(\mathbf{X};\mathbf{P}_r) - \overrightarrow{\mathbf{X}}'\boldsymbol{\theta}_0\}] + O_p(N^{-\frac{1}{2}}) = O_p(N^{-\frac{1}{2}})$$

from Lemma A.1 (b) and $\mathbb{E}[\overrightarrow{\mathbf{X}}\{\mu(\mathbf{X};\mathbf{P}_r) - \overrightarrow{\mathbf{X}}'\boldsymbol{\theta}_0\}] = \mathbf{0}$ due to (A.11) and 1.1; and lastly, (III) $\Gamma_N^{-1} = I_{(p+1)} + O_p(N^{-\frac{1}{2}})$. It then follows from (I)-(III) that

$$\widehat{\boldsymbol{\theta}}_{(\mathbf{P}_r,\mathbb{K})} - \boldsymbol{\theta}_0 = (\widehat{\boldsymbol{\eta}}_{(\mathbf{P}_r,\mathbb{K})} - \boldsymbol{\eta}_{\mathbf{P}_r}) + \mathbb{K}^{-1}\sum_{k=1}^{\mathbb{K}}\widehat{\mathbf{S}}_{n,k}^{*} + O_p(N^{-\frac{1}{2}}). \tag{A.14}$$

Using (A.13) and (1.11) in (A.14), we then have:

$$\widehat{\boldsymbol{\theta}}_{(\mathbf{P}_r,\mathbb{K})} - \boldsymbol{\theta}_0 = n^{-1}\sum_{i=1}^{n}\boldsymbol{\psi}(\mathbf{Z}_i,\mathbf{P}_r) - \mathbb{K}^{-1}\sum_{k=1}^{\mathbb{K}}\Big\{n_{\mathbb{K}}^{-1}\sum_{i\in\mathcal{I}_k}\widehat{\mathbf{G}}_k(\mathbf{X}_i)\Big\} + O_p(b_{n,\mathbb{K}}),$$

where $b_{n,\mathbb{K}} = n^{-1} + n^{-\frac{1}{2}} c_{n_{\mathbb{K}}^-} + N^{-\frac{1}{2}}$. It follows, as claimed in (1.12), that

$$n^{\frac{1}{2}}(\widehat{\boldsymbol{\theta}}_{(\mathbf{P}_r,\mathbb{K})} - \boldsymbol{\theta}_0) = n^{-\frac{1}{2}} \sum_{i=1}^{n} \boldsymbol{\psi}(\mathbf{Z}_i, \mathbf{P}_r) - \mathbb{G}_{n,\mathbb{K}} + O_p(c^*_{n,\mathbb{K}}) \quad \blacksquare \tag{A.15}$$

We next show that $\mathbb{G}_{n,\mathbb{K}} = O_p(c_{n_{\mathbb{K}}^-})$ for any fixed $\mathbb{K} \geq 2$. To this end, let $\mathbb{T}_k^{(n)} = (n_{\mathbb{K}})^{-\frac{1}{2}} \sum_{i \in \mathcal{I}_k} \widehat{\mathbf{G}}_k(\mathbf{X}_i)$, $\widehat{D}_k = \sup_{\mathbf{x} \in \mathcal{X}} |\widehat{\Delta}_k(\mathbf{x}; \mathbf{P}_r, \widehat{\mathbf{P}}_{r,k})|$ and $C = \sup_{\mathbf{x} \in \mathcal{X}} \|\overrightarrow{\mathbf{x}}\| < \infty$. For any subset $\mathcal{A} \subseteq \mathcal{L}$, let $\mathbb{P}_{\mathcal{A}}$ denote the joint distribution of the observations in $\mathcal{A}$, and let $\mathbb{E}_{\mathcal{A}}(.)$ denote expectation w.r.t. $\mathbb{P}_{\mathcal{A}}$. By definition, $\mathbb{G}_{n,\mathbb{K}} = \mathbb{K}^{-\frac{1}{2}} \sum_{k=1}^{\mathbb{K}} \mathbb{T}_k^{(n)} = O_p(c_{n_{\mathbb{K}}^-})$ if and only if given any $\epsilon > 0$, $\exists\ M_\epsilon > 0$ such that $\mathbb{P}\left(\|\mathbb{G}_{n,\mathbb{K}}\| > M_\epsilon c_{n_{\mathbb{K}}^-}\right) \leq \epsilon\ \forall\ n$. Note that for any $M > 0$,

$$\begin{aligned}
\mathbb{P}\left(\|\mathbb{G}_{n,\mathbb{K}}\| > M c_{n_{\mathbb{K}}^-}\right) &\leq \mathbb{P}\left(\mathbb{K}^{-\frac{1}{2}} \sum_{k=1}^{\mathbb{K}} \|\mathbb{T}_k^{(n)}\| > M c_{n_{\mathbb{K}}^-}\right) \\
&\leq \sum_{k=1}^{\mathbb{K}} \mathbb{P}\left(\mathbb{K}^{-\frac{1}{2}} \|\mathbb{T}_k^{(n)}\| > \frac{M c_{n_{\mathbb{K}}^-}}{\mathbb{K}}\right) \leq \sum_{k=1}^{\mathbb{K}} \sum_{l=1}^{p+1} \mathbb{P}\left\{|\mathbb{T}_{k[l]}^{(n)}| > \frac{M c_{n_{\mathbb{K}}^-}}{\mathbb{K}^{\frac{1}{2}}(p+1)^{\frac{1}{2}}}\right\} \\
&= \sum_{k=1}^{\mathbb{K}} \sum_{l=1}^{p+1} \mathbb{E}_{\mathcal{L}_k^-}\left[\mathbb{P}_{\mathcal{L}_k}\left\{|\mathbb{T}_{k[l]}^{(n)}| > \frac{M c_{n_{\mathbb{K}}^-}}{\mathbb{K}^{\frac{1}{2}}(p+1)^{\frac{1}{2}}} \,\middle|\, \mathcal{L}_k^-\right\}\right],
\end{aligned} \tag{A.16}$$

where the steps follow from repeated use of Bonferroni's inequality and other standard arguments. Now, conditional on $\mathcal{L}_k^-$ ($\perp\!\!\!\perp \mathcal{L}_k$, with $\mathbb{K} \geq 2$), $n_{\mathbb{K}}^{\frac{1}{2}} \mathbb{T}_k^{(n)}$ is a centered sum of the i.i.d. random vectors $\{\overrightarrow{\mathbf{X}}_i \widehat{\Delta}_k(\mathbf{X}_i; \mathbf{P}_r, \widehat{\mathbf{P}}_{r,k})\}_{i \in \mathcal{I}_k}$ which, due to assumption (ii) and the compactness of $\mathcal{X}$, are bounded by: $C\widehat{D}_k < \infty$ a.s. $[\mathbb{P}_{\mathcal{L}_k^-}]\ \forall\ k, n$. Hence, applying Hoeffding's inequality to $\mathbb{T}_{k[l]}^{(n)}\ \forall\ l$, we have:

$$\begin{aligned}
&\mathbb{P}_{\mathcal{L}_k}\left\{|\mathbb{T}_{k[l]}^{(n)}| > \frac{M c_{n_{\mathbb{K}}^-}}{\mathbb{K}^{\frac{1}{2}}(p+1)^{\frac{1}{2}}} \,\middle|\, \mathcal{L}_k^-\right\} \leq 2 \exp\left\{-\frac{M^2 c^2_{n_{\mathbb{K}}^-}}{2(p+1)\mathbb{K}C^2\widehat{D}_k^2}\right\} \\
&\text{a.s. } [\mathbb{P}_{\mathcal{L}_k^-}]\ \forall\ n;\ \text{for each } k \in \{1, ..., \mathbb{K}\} \text{ and } \forall\ 1 \leq l \leq (p+1).
\end{aligned} \tag{A.17}$$

Now, since $\widehat{D}_k = O_p(c_{n_{\mathbb{K}}^-})$, $(c_{n_{\mathbb{K}}^-}/\widehat{D}_k) \geq 0$ is stochastically bounded away from 0. Thus, $\forall\ k$, and for any given $\epsilon > 0$, $\exists\ \delta(k, \epsilon) > 0$ (independent of n) such that: $\mathbb{P}_{\mathcal{L}_k^-}\{(c_{n_{\mathbb{K}}^-}/\widehat{D}_k) \leq$

$\delta(k,\epsilon)\} \leq \epsilon^* \ \forall\ n$, where $\epsilon^* = \epsilon/\{4\mathbb{K}(p+1)\} > 0$. Let $\widetilde{\delta}(\mathbb{K},\epsilon) = \min\{\delta(k,\epsilon) : k = 1, ..., \mathbb{K}\} > 0$ (as $\mathbb{K}$ is fixed). Let $\mathbb{A}(k,\epsilon)$ denote the event: $\{(c_{n_{\mathbb{K}}^-}/\widehat{D}_k) \leq \widetilde{\delta}(\mathbb{K},\epsilon)\}$, and let $\mathbb{A}^c(k,\epsilon)$ be its complement. Then, $\mathbb{P}_{\mathcal{L}_k^-}\{\mathbb{A}(k,\epsilon)\} \leq \epsilon^*$, while on $\mathbb{A}^c(k,\epsilon), (c_{n_{\mathbb{K}}^-}/\widehat{D}_k) > \widetilde{\delta}(\mathbb{K},\epsilon)$. Thus, the bound in (A.17) is dominated by: $2\ \exp[-M^2\widetilde{\delta}^2(\mathbb{K},\epsilon)/\{2(p+1)\mathbb{K}C^2\}]$ on $\mathbb{A}^c(k,\epsilon)$, and trivially by 2 on $\mathbb{A}(k,\epsilon)\ \forall\ k$. Plugging the bound of (A.17) into (A.16) and using all these facts, we then have:

$$
\begin{aligned}
\mathbb{P}\left(\|\mathbb{G}_{n,\mathbb{K}}\| > M c_{n_{\mathbb{K}}^-}\right) \ &\leq\ \sum_{k=1}^{\mathbb{K}}\sum_{l=1}^{p+1}\mathbb{E}_{\mathcal{L}_k^-}\left[2\exp\left\{-\frac{M^2 c^2_{n_{\mathbb{K}}^-}}{2(p+1)\mathbb{K}C^2\widehat{D}_k^2}\right\}\right] \\
&=\ \sum_{k=1}^{\mathbb{K}}\sum_{l=1}^{p+1}\mathbb{E}_{\mathcal{L}_k^-}\left[2\exp\left\{-\frac{M^2 c^2_{n_{\mathbb{K}}^-}}{2(p+1)\mathbb{K}C^2\widehat{D}_k^2}\right\}\left\{1_{\mathbb{A}^c(k,\epsilon)} + 1_{\mathbb{A}(k,\epsilon)}\right\}\right] \\
&\leq\ \sum_{k=1}^{\mathbb{K}}\sum_{l=1}^{p+1}\left[2\exp\left\{-\frac{M^2\widetilde{\delta}^2(\mathbb{K},\epsilon)}{2(p+1)\mathbb{K}C^2}\right\}\mathbb{P}_{\mathcal{L}_k^-}\{\mathbb{A}^c(k,\epsilon)\} + 2\,\mathbb{P}_{\mathcal{L}_k^-}\{\mathbb{A}(k,\epsilon)\}\right] \\
&\leq\ 2\mathbb{K}(p+1)\left[\exp\left\{-\frac{M^2\widetilde{\delta}^2(\mathbb{K},\epsilon)}{2(p+1)\mathbb{K}C^2}\right\} + \epsilon^*\right] \\
&\leq\ \frac{\epsilon}{2} + \frac{\epsilon}{2} = \epsilon\ \ \text{(with some suitable choice } M_\epsilon \text{ for } M), \qquad \text{(A.18)}
\end{aligned}
$$

where the last step follows from noting the definition of ϵ^* and choosing M_ϵ to be any M large enough such that $4\exp[-\ M^2\widetilde{\delta}^2(\mathbb{K},\epsilon)/\{2(p+1)\mathbb{K}C^2\}] \leq \epsilon/\{\mathbb{K}(p+1)\}$. Thus, (A.18) shows $\mathbb{G}_{n,\mathbb{K}} = O_p(c_{n_{\mathbb{K}}^-})$ for any fixed $\mathbb{K} \geq 2$. This further establishes (1.13) and all its associated implications. The proof of Theorem 1.2 is now complete. ∎

A.4 Proof of Theorem 1.3

We first note that for $a_{n,2} = (\log n)^{\frac{1}{2}}(nh^r)^{-\frac{1}{2}} + h^q$,

$$
\sup_{\mathbf{w}\in\mathcal{X}_{\mathbf{P}_r}}|\widetilde{\varphi}^{(\varrho)}_{\mathbf{P}_r}(\mathbf{w}) - \varphi^{(\varrho)}_{\mathbf{P}_r}(\mathbf{w})| = O_p(a_{n,2}), \quad \forall\ \varrho \in \{0,1\}. \qquad \text{(A.19)}
$$

To see this, note that under Assumption 1.2 (ii)-(iii), Theorem 2 of Hansen (2008) applies, and we have for $d_n = (\log n)^{\frac{1}{2}}(nh^r)^{-\frac{1}{2}}$,

$$\sup_{\mathbf{w}\in\mathcal{X}_{\mathbf{P}_r}} |\widetilde{\varphi}^{(\varrho)}_{\mathbf{P}_r}(\mathbf{w}) - \mathbb{E}_{\mathcal{L}}\{\widetilde{\varphi}^{(\varrho)}_{\mathbf{P}_r}(\mathbf{w})\}| = O_p(d_n) \quad \forall\ \varrho \in \{0,1\}.$$

Next, using standard arguments based on a q^{th} order Taylor series expansion of $\varphi^{(\varrho)}_{\mathbf{P}_r}(.)$ and noting that $K(.)$ is a q^{th} order kernel, we obtain:

$$\sup_{\mathbf{w}\in\mathcal{X}_{\mathbf{P}_r}} |\mathbb{E}_{\mathcal{L}}\{\widetilde{\varphi}^{(\varrho)}_{\mathbf{P}_r}(\mathbf{w})\} - \varphi^{(\varrho)}_{\mathbf{P}_r}(\mathbf{w})| = O(h^q) \quad \forall\ \varrho \in \{0,1\}.$$

Combining these two results gives (A.19). Further,

$$\begin{aligned}
&\sup_{\mathbf{x}\in\mathcal{X}} |\widetilde{m}(\mathbf{x};\mathbf{P}_r) - m(\mathbf{x};\mathbf{P}_r)| \;=\; \sup_{\mathbf{w}\in\mathcal{X}_{\mathbf{P}_r}} |\widetilde{m}_{\mathbf{P}_r}(\mathbf{w}) - m_{\mathbf{P}_r}(\mathbf{w})| \\
&\leq\; \sup_{\mathbf{w}\in\mathcal{X}_{\mathbf{P}_r}} \left|\frac{\widetilde{l}_{\mathbf{P}_r}(\mathbf{w}) - l_{\mathbf{P}_r}(\mathbf{w})}{\widetilde{f}_{\mathbf{P}_r}(\mathbf{w})}\right| + \sup_{\mathbf{w}\in\mathcal{X}_{\mathbf{P}_r}} \left\{\left|\frac{|l_{\mathbf{P}_r}(\mathbf{w})|}{f_{\mathbf{P}_r}(\mathbf{w})} - \frac{|l_{\mathbf{P}_r}(\mathbf{w})|}{\widetilde{f}_{\mathbf{P}_r}(\mathbf{w})}\right|\right\} \\
&=\; O_p(a_{n,2}),
\end{aligned} \tag{A.20}$$

where the last step follows from repeated use of (A.19) and Assumption 1.2 (iii)-(iv). Next, we aim to bound $\sup_{\mathbf{x}\in\mathcal{X}}|\widehat{\varphi}^{(\varrho)}(\mathbf{x};\widehat{\mathbf{P}}_r) - \widetilde{\varphi}^{(\varrho)}(\mathbf{x};\mathbf{P}_r)|$ to account for the potential estimation error of $\widehat{\mathbf{P}}_r$. Using a first order Taylor series expansion of $K(.)$ under Assumption 1.2 (vi), we have: $\forall\ \varrho \in \{0,1\}$,

$$\begin{aligned}
\widehat{\varphi}^{(\varrho)}(\mathbf{x};\widehat{\mathbf{P}}_r) - \widetilde{\varphi}^{(\varrho)}(\mathbf{x};\mathbf{P}_r) \;&=\; \frac{1}{nh^r}\sum_{i=1}^{n} \boldsymbol{\nabla} K'(\mathbf{w}_{i,\mathbf{x}})(\widehat{\mathbf{P}}_r - \mathbf{P}_r)' \left(\frac{\mathbf{x}-\mathbf{X}_i}{h}\right) Y_i^{\varrho} \\
&=\; \text{trace}\left\{(\widehat{\mathbf{P}}_r - \mathbf{P}_r)'\widehat{\mathbf{M}}^{(1)}_{n,\varrho,\mathbf{x}}\right\} + \text{trace}\left\{(\widehat{\mathbf{P}}_r - \mathbf{P}_r)'\widehat{\mathbf{M}}^{(2)}_{n,\varrho,\mathbf{x}}\right\},
\end{aligned} \tag{A.21}$$

where

$$\widehat{\mathbf{M}}^{(1)}_{n,\varrho,\mathbf{x}} = \frac{1}{nh^{r+1}}\sum_{i=1}^{n}(\mathbf{x}-\mathbf{X}_i)\left\{\boldsymbol{\nabla} K'\left(\frac{\mathbf{P}'_r\mathbf{x} - \mathbf{P}'_r\mathbf{X}_i}{h}\right)\right\} Y_i^{\varrho} \qquad \text{and}$$

$$\widehat{\mathbf{M}}^{(2)}_{n,\varrho,\mathbf{x}} = \frac{1}{nh^{r+1}} \sum_{i=1}^{n} (\mathbf{x} - \mathbf{X}_i) \left\{ \boldsymbol{\nabla} K'(\mathbf{w}_{i,\mathbf{x}}) - \boldsymbol{\nabla} K' \left(\frac{\mathbf{P}'_r \mathbf{x} - \mathbf{P}'_r \mathbf{X}_i}{h} \right) \right\} Y_i^{\varrho},$$

with $\mathbf{w}_{i,\mathbf{x}} \in \mathbb{R}^r$ being 'intermediate' points satisfying: $\|\mathbf{w}_{i,\mathbf{x}} - \mathbf{P}'_r(\mathbf{x} - \mathbf{X}_i)h^{-1}\| \leq \|\widehat{\mathbf{P}}'_r(\mathbf{x} - \mathbf{X}_i)h^{-1} - \mathbf{P}'_r(\mathbf{x} - \mathbf{X}_i)h^{-1}\| \leq O_p(\alpha_n h^{-1})$. The last bound, based on $(\widehat{\mathbf{P}}_r - \mathbf{P}_r) = O_p(\alpha_n)$ and the compactness of $\mathcal{X}$, is uniform in $(i, \mathbf{x})$. For any matrix $\mathbf{A} = [a_{\text{ij}}]$, let $\|\mathbf{A}\|_{\max}$ denote the max-norm of $\mathbf{A}$, and $|\mathbf{A}|$ denote the matrix $[|a_{\text{ij}}|]$. Now, Assumption 1.2 (viii) implies: $\|\boldsymbol{\nabla} K(\mathbf{w}_1) - \boldsymbol{\nabla} K(\mathbf{w}_2)\| \leq B\|\mathbf{w}_1 - \mathbf{w}_2\| \ \forall \ \mathbf{w}_1, \mathbf{w}_2 \in \mathbb{R}^r$, for some constant $B < \infty$. Then using the above arguments, we note that $\forall \ \varrho \in \{0, 1\}$, $\| \sup_{\mathbf{x} \in \mathcal{X}} |\widehat{\mathbf{M}}^{(2)}_{n,\varrho,\mathbf{x}}| \|_{\max}$ is bounded by:

$$\begin{aligned}
&\sup_{\mathbf{x} \in \mathcal{X}} \left\{ \frac{B}{nh^{r+1}} \sum_{i=1}^{n} \|\mathbf{x} - \mathbf{X}_i\| \left\| \mathbf{w}_{i,\mathbf{x}} - \frac{\mathbf{P}'_r \mathbf{x} - \mathbf{P}'_r \mathbf{X}_i}{h} \right\| |Y_i^{\varrho}| \right\} \\
\leq \ &\sup_{\mathbf{x} \in \mathcal{X}} \left\{ \frac{B}{nh^{r+1}} \sum_{i=1}^{n} \|\mathbf{x} - \mathbf{X}_i\| \left\| \frac{(\widehat{\mathbf{P}}_r - \mathbf{P}_r)'(\mathbf{x} - \mathbf{X}_i)}{h} \right\| |Y_i^{\varrho}| \right\} \\
\leq &\sup_{\mathbf{x} \in \mathcal{X}, \mathbf{X} \in \mathcal{X}} \left\{ \|\mathbf{x} - \mathbf{X}\| \|(\widehat{\mathbf{P}}_r - \mathbf{P}_r)'(\mathbf{x} - \mathbf{X})\| \right\} \frac{B}{nh^{r+2}} \sum_{i=1}^{n} |Y_i^{\varrho}| \ = O_p \left(\frac{\alpha_n}{h^{r+2}} \right).
\end{aligned}$$

The first two steps above use the triangle inequality, the Lipschitz continuity of $\boldsymbol{\nabla} K(.)$ and the definition of $\mathbf{w}_{i,\mathbf{x}}$, while the next two use the compactness of $\mathcal{X}$, the uniform bound obtained in the last paragraph, the Law of Large Numbers (LLN), and that $(\widehat{\mathbf{P}}_r - \mathbf{P}_r) = O_p(\alpha_n)$. Thus, we have:

$$\sup_{\mathbf{x} \in \mathcal{X}} \left| \text{trace} \left\{ (\widehat{\mathbf{P}}_r - \mathbf{P}_r)' \widehat{\mathbf{M}}^{(2)}_{n,\varrho,\mathbf{x}} \right\} \right| = O_p \left(\frac{\alpha_n^2}{h^{r+2}} \right) \quad \forall \ \varrho \in \{0, 1\}. \tag{A.22}$$

Now for bounding $\widehat{\mathbf{M}}^{(1)}_{n,\varrho,\mathbf{x}}$, let us first write it as:

$$\begin{aligned}
&\widehat{\mathbf{M}}^{(1)}_{n,\varrho,\mathbf{x}} = \widehat{\mathbf{M}}^{(1,1)}_{n,\varrho,\mathbf{x}} - \widehat{\mathbf{M}}^{(1,2)}_{n,\varrho,\mathbf{x}}, \ \text{where} \\
&\widehat{\mathbf{M}}^{(1,1)}_{n,\varrho,\mathbf{x}} = (nh^{r+1})^{-1} \sum_{i=1}^{n} \mathbf{x} \boldsymbol{\nabla} K' \{ \mathbf{P}'_r(\mathbf{x} - \mathbf{X}_i)/h \} Y_i^{\varrho} \ \text{and} \\
&\widehat{\mathbf{M}}^{(1,2)}_{n,\varrho,\mathbf{x}} = (nh^{r+1})^{-1} \sum_{i=1}^{n} \mathbf{X}_i \boldsymbol{\nabla} K' \{ \mathbf{P}'_r(\mathbf{x} - \mathbf{X}_i)/h \} Y_i^{\varrho} \ \forall \ \varrho \in \{0, 1\}.
\end{aligned}$$

Then, under Assumption 1.2 (iii), (vi) and (vii), using Theorem 2 of Hansen (2008) along with the compactness of $\mathcal{X}$, we have: for each $s \in \{1,2\}$ and $\varrho \in \{0,1\}$,

$$\left\| \sup_{\mathbf{x}\in\mathcal{X}} \left| \widehat{\mathbf{M}}^{(1,s)}_{n,\varrho,\mathbf{x}} - \mathbb{E}_{\mathcal{L}}\left(\widehat{\mathbf{M}}^{(1,s)}_{n,\varrho,\mathbf{x}}\right) \right| \right\|_{\max} \leq O_p\left(\frac{\log n}{nh^{r+2}}\right)^{\frac{1}{2}}. \tag{A.23}$$

Now, $\forall\ \varrho \in \{0,1\}$, let $\nu^{(\varrho)}(\mathbf{w}) = \mathbb{E}\{Y^{\varrho} \mid \mathbf{X}_{\mathbf{P}_r} = \mathbf{w}\} f_{\mathbf{P}_r}(\mathbf{w})$ and $\boldsymbol{\xi}^{(\varrho)}(\mathbf{w}) = \mathbb{E}\{\mathbf{X}Y^{\varrho} \mid \mathbf{X}_{\mathbf{P}_r} = \mathbf{w}\} f_{\mathbf{P}_r}(\mathbf{w})$. Further, let $\{\boldsymbol{\nabla}\nu^{(\varrho)}(\mathbf{w})\}_{r\times 1}$ and $\{\boldsymbol{\nabla}\boldsymbol{\xi}^{(\varrho)}(\mathbf{w})\}_{p\times r}$ denote their respective first order derivatives. Then, $\forall\ \varrho \in \{0,1\}$, we have:

$$\begin{aligned}
& \left\| \sup_{\mathbf{x}\in\mathcal{X}} \left| \mathbb{E}_{\mathcal{L}}\left(\widehat{\mathbf{M}}^{(1,1)}_{n,\varrho,\mathbf{x}}\right) \right| \right\|_{\max} \\
= & \left\| \sup_{\mathbf{x}\in\mathcal{X}} \left| \frac{\mathbf{x}}{h^{r+1}} \int \nu^{(\varrho)}(\mathbf{w})\ \boldsymbol{\nabla} K'\left(\frac{\mathbf{P}'_r\mathbf{x} - \mathbf{w}}{h}\right) d\mathbf{w} \right| \right\|_{\max} \\
= & \left\| \sup_{\mathbf{x}\in\mathcal{X}} \left| \mathbf{x} \int \boldsymbol{\nabla}\nu^{(\varrho)'}\left(\mathbf{P}'_r\mathbf{x} + h\boldsymbol{\psi}\right)\ K(\boldsymbol{\psi}) d\boldsymbol{\psi} \right| \right\|_{\max} = O(1), \quad \text{and}
\end{aligned} \tag{A.24}$$

$$\begin{aligned}
& \left\| \sup_{\mathbf{x}\in\mathcal{X}} \left| \mathbb{E}_{\mathcal{L}}\left(\widehat{\mathbf{M}}^{(1,2)}_{n,\varrho,\mathbf{x}}\right) \right| \right\|_{\max} \\
= & \left\| \sup_{\mathbf{x}\in\mathcal{X}} \left| h^{-(r+1)} \int \boldsymbol{\xi}^{(\varrho)}(\mathbf{w})\ \boldsymbol{\nabla} K'\left(\frac{\mathbf{P}'_r\mathbf{x} - \mathbf{w}}{h}\right) d\mathbf{w} \right| \right\|_{\max} \\
= & \left\| \sup_{\mathbf{x}\in\mathcal{X}} \left| \mathbf{D}(\mathbf{x}) \int \boldsymbol{\nabla}\boldsymbol{\xi}^{(\varrho)}\left(\mathbf{P}'_r\mathbf{x} + h\boldsymbol{\psi}\right)\ K(\boldsymbol{\psi}) d\boldsymbol{\psi} \right| \right\|_{\max} = O(1),
\end{aligned} \tag{A.25}$$

where, $\forall\ \mathbf{x} \in \mathcal{X}$, $\mathbf{D}(\mathbf{x})$ denotes the $p \times p$ diagonal matrix: $\text{diag}(\mathbf{x}_{[1]}, \ldots, \mathbf{x}_{[p]})$. In both (A.24) and (A.25), the first step follows from definition, the second from standard arguments based on integration by parts (applied coordinate-wise) and change of variable, while the last one is due to compactness of $\mathcal{X}$ and a medley of the conditions in Assumption 1.2 namely, boundedness and integrability of $K(.)$ and $\boldsymbol{\nabla}K(.)$, (iii) and (v) for (A.24) so that $\boldsymbol{\nabla}\nu^{(\varrho)}(.)$ is bounded on $\mathcal{X}_{\mathbf{P}_r}$, and (ix) for (A.25). It now follows that for each $\varrho \in \{0,1\}$,

$$\left\| \sup_{\mathbf{x}\in\mathcal{X}} \left| \mathbb{E}_{\mathcal{L}}\left(\widehat{\mathbf{M}}^{(1)}_{n,\varrho,\mathbf{x}}\right) \right| \right\|_{\max} = O(1). \tag{A.26}$$

Letting $d_n^* = (\log n)^{\frac{1}{2}}(nh^{r+2})^{-\frac{1}{2}}$, we now have from (A.23) and (A.26):

$$\sup_{\mathbf{x}\in\mathcal{X}} \left|\text{trace}\left\{(\widehat{\mathbf{P}}_r' - \mathbf{P}_r')\widehat{\mathbf{M}}^{(1)}_{n,\varrho,\mathbf{x}}\right\}\right| = O_p\left(\alpha_n d_n^* + \alpha_n\right) \quad \forall\ \varrho \in \{0,1\}. \tag{A.27}$$

Applying (A.27) and (A.22) to (A.21) using the triangle inequality, we have $\forall\ \varrho$,

$$\sup_{\mathbf{x}\in\mathcal{X}} |\widehat{\varphi}^{(\varrho)}(\mathbf{x};\widehat{\mathbf{P}}_r) - \widetilde{\varphi}^{(\varrho)}(\mathbf{x};\mathbf{P}_r)| = O_p\left\{\frac{\alpha_n^2}{h^{r+2}} + \alpha_n\frac{(\log\ n)^{\frac{1}{2}}}{(nh^{r+2})^{\frac{1}{2}}} + \alpha_n\right\}. \tag{A.28}$$

Finally, note that $\widehat{m}(\mathbf{x};\widehat{\mathbf{P}}_r) = \widehat{l}(\mathbf{x};\widehat{\mathbf{P}}_r)/\widehat{f}(\mathbf{x};\widehat{\mathbf{P}}_r) = \widehat{\varphi}^{(1)}(\mathbf{x};\widehat{\mathbf{P}}_r)/\widehat{\varphi}^{(0)}(\mathbf{x};\widehat{\mathbf{P}}_r)$. Repeated use of (A.28), along with (A.20) and Assumption 1.2 (iii)-(iv), leads to:

$$\begin{aligned}
&\sup_{\mathbf{x}\in\mathcal{X}} \left|\widehat{m}(\mathbf{x};\widehat{\mathbf{P}}_r) - m(\mathbf{x};\mathbf{P}_r)\right| \\
&\leq \sup_{\mathbf{x}\in\mathcal{X}} \left|\widehat{m}(\mathbf{x};\widehat{\mathbf{P}}_r) - \widetilde{m}(\mathbf{x};\mathbf{P}_r)\right| + \sup_{\mathbf{x}\in\mathcal{X}} |\widetilde{m}(\mathbf{x};\mathbf{P}_r) - m(\mathbf{x};\mathbf{P}_r)| \\
&\leq \sup_{\mathbf{x}\in\mathcal{X}} \left\{\left|\frac{\widehat{l}(\mathbf{x};\widehat{\mathbf{P}}_r) - \widetilde{l}(\mathbf{x};\mathbf{P}_r)}{\widehat{f}(\mathbf{x};\widehat{\mathbf{P}}_r)}\right| + \left|\frac{\widetilde{l}(\mathbf{x};\mathbf{P}_r)}{\widetilde{f}(\mathbf{x};\mathbf{P}_r)} - \frac{\widetilde{l}(\mathbf{x};\mathbf{P}_r)}{\widehat{f}(\mathbf{x};\widehat{\mathbf{P}}_r)}\right|\right\} + O_p(a_{n,2}) \\
&\leq O_p\left\{\frac{\alpha_n^2}{h^{r+2}} + \alpha_n\frac{(\log\ n)^{\frac{1}{2}}}{(nh^{r+2})^{\frac{1}{2}}} + \alpha_n\right\} + O_p(a_{n,2}) \quad = O_p(a_{n,1} + a_{n,2}).
\end{aligned} \tag{A.29}$$

The proof of Theorem 1.3 is now complete. ∎

A.5 Proofs of Lemmas A.2-A.3 and Theorem 1.4

A.5.1 Proof of Lemma A.2

First note that for each $\varrho \in \{0,1\}$,

$$\int \widetilde{\varphi}^{(\varrho)}(\mathbf{x};\mathbf{P}_r)\mathbb{P}_n(d\mathbf{x}) = n^{-2}\sum_{i_1=1}^{n}\sum_{i_2=1}^{n}\mathbf{H}^{(n,\varrho)}_{i_1,i_2}$$

is a V-statistic, where $\mathbf{H}^{(n,\varrho)}_{i_1,i_2} = h^{-r}\boldsymbol{\lambda}(\mathbf{X}_{i_1})Y_{i_2}^{\varrho}K\{\mathbf{P}_r'(\mathbf{X}_{i_1} - \mathbf{X}_{i_2})/h\}$. Using the V-statistic projection result given in Lemma 8.4 of Newey and McFadden (1994), it then follows that

for each $\varrho \in \{0,1\}$,

$$\begin{aligned} & \mathbb{G}_n^* \left\{ \boldsymbol{\lambda}(.)[\widetilde{\varphi}^{(\varrho)}(.;\mathbf{P}_r) - \mathbb{E}_{\mathcal{L}}\{\widetilde{\varphi}^{(\varrho)}(.;\mathbf{P}_r)\}] \right\} \\ = \;& n^{-\frac{1}{2}} O_p \left[\mathbb{E}(\|\mathbf{H}_{i_1,i_1}^{(n,\varrho)}\|) + \{\mathbb{E}(\|\mathbf{H}_{i_1,i_2}^{(n,\varrho)}\|^2)\}^{\frac{1}{2}} \right] = O_p\left(n^{-\frac{1}{2}} h^{-r}\right), \end{aligned} \tag{A.30}$$

The last step follows from $K(.)$ and $\boldsymbol{\lambda}(.)$ being bounded and Y^{ϱ} having finite 2^{nd} moments. Now, observe that $n^{\frac{1}{2}}\mathbb{G}_n^*\left\{\boldsymbol{\lambda}(.)[\mathbb{E}_{\mathcal{L}}\{\widetilde{\varphi}_\star^{(\varrho)}(.;\mathbf{P}_r)\}]\right\}$ is a centered sum of i.i.d. random vectors bounded by:

$$D_{n,\varrho} = \sup_{\mathbf{x}\in\mathcal{X}} \left\{ \|\boldsymbol{\lambda}(\mathbf{x})\| \; |\mathbb{E}_{\mathcal{L}}\{\widetilde{\varphi}_\star^{(\varrho)}(\mathbf{x};\mathbf{P}_r)\}| \right\} = O(h^q) \quad \forall\; \varrho \in \{0,1\},$$

where throughout, for any estimator $\tilde{\xi}(.)$ with population limit $\xi(.)$, we use the notation $\tilde{\xi}_\star(.)$ to denote its centered version given by: $\tilde{\xi}_\star(.) = \tilde{\xi}(.) - \xi(.)$. Here, $D_{n,\varrho} = O(h^q)$ since $\boldsymbol{\lambda}(.)$ is bounded and $\sup_{\mathbf{x}\in\mathcal{X}}|\mathbb{E}_{\mathcal{L}}\{\widetilde{\varphi}_\star(\mathbf{x};\mathbf{P}_r)\}| = \sup_{\mathbf{w}\in\mathcal{X}_{\mathbf{P}_r}}|\mathbb{E}_{\mathcal{L}}\{\widetilde{\varphi}_{\mathbf{P}_r}^{(\varrho)}(\mathbf{w})\} - \varphi_{\mathbf{P}_r}^{(\varrho)}(\mathbf{w})| = O(h^q)$, as argued while proving (A.19). Hence, $\exists$ a constant $\kappa_\varrho > 0$ such that $h^q/D_{n,\varrho} \geq \kappa_\varrho \;\forall\; n$. Then, using Hoeffding's Inequality, we have: $\forall\; n$, given any $\epsilon > 0$ and any $M = M(\epsilon)$ large enough,

$$\begin{aligned} & \sum_{l=1}^{d} \mathbb{P}\left[\left| \mathbb{G}_n^*\left\{ \boldsymbol{\lambda}_{[l]}(.)[\mathbb{E}_{\mathcal{L}}\{\widetilde{\varphi}_\star^{(\varrho)}(.;\mathbf{P}_r)\}] \right\} \right| > \frac{Mh^q}{d^{\frac{1}{2}}} \right] \leq 2d \exp\left(-\frac{M^2 h^{2q}}{2dD_{n,\varrho}^2} \right) \Rightarrow \\ & \mathbb{P}\left[\left\| \mathbb{G}_n^*\left\{ \boldsymbol{\lambda}(.)[\mathbb{E}_{\mathcal{L}}\{\widetilde{\varphi}_\star^{(\varrho)}(.;\mathbf{P}_r)\}] \right\} \right\| > Mh^q \right] \leq 2d \exp\left(-\frac{M^2\kappa_\varrho^2}{2d} \right) \leq \epsilon \;\; \Rightarrow \\ & \mathbb{G}_n^*\left\{ \boldsymbol{\lambda}(.)[\mathbb{E}_{\mathcal{L}}\{\widetilde{\varphi}_\star^{(\varrho)}(.;\mathbf{P}_r)\}] \right\} = O_p(h^q) \quad \forall\; \varrho \in \{0,1\}. \end{aligned} \tag{A.31}$$

Combining (A.30) and (A.31) using the linearity of $\mathbb{G}_n^*(.)$, we then have (A.1). ∎

Next, to show (A.2), let $f(\mathbf{x};\mathbf{P}_r) = \varphi^{(0)}(\mathbf{x};\mathbf{P}_r)$ and $l(\mathbf{x};\mathbf{P}_r) = \varphi^{(1)}(\mathbf{x};\mathbf{P}_r)$. Then, we write

$$\mathbb{G}_n^*[\boldsymbol{\lambda}(.)\{\widetilde{m}_\star(.;\mathbf{P}_r)\}] \;=\; \mathbb{G}_n^*[\boldsymbol{\lambda}(.)\{\widetilde{\mathbf{T}}_{n,\mathbf{P}_r}^{(1)}(.) - \widetilde{\mathbf{T}}_{n,\mathbf{P}_r}^{(2)}(.) - \widetilde{\mathbf{T}}_{n,\mathbf{P}_r}^{(3)}(.) + \widetilde{\mathbf{T}}_{n,\mathbf{P}_r}^{(4)}(.)\}],$$

where

$$\widetilde{\mathbf{T}}^{(1)}_{n,\mathbf{P}_r}(\mathbf{x}) = \frac{\widetilde{l}_\star(\mathbf{x};\mathbf{P}_r)}{f(\mathbf{x};\mathbf{P}_r)}, \qquad \widetilde{\mathbf{T}}^{(2)}_{n,\mathbf{P}_r}(\mathbf{x}) = \frac{\widetilde{f}_\star(\mathbf{x};\mathbf{P}_r)l(\mathbf{x};\mathbf{P}_r)}{f(\mathbf{x};\mathbf{P}_r)^2},$$
$$\widetilde{\mathbf{T}}^{(3)}_{n,\mathbf{P}_r}(\mathbf{x}) = \frac{\widetilde{l}_\star(\mathbf{x};\mathbf{P}_r)\widetilde{f}_\star(\mathbf{x};\mathbf{P}_r)}{\widetilde{f}(\mathbf{x};\mathbf{P}_r)f(\mathbf{x};\mathbf{P}_r)}, \text{ and } \widetilde{\mathbf{T}}^{(4)}_{n,\mathbf{P}_r}(\mathbf{x}) = \frac{l(\mathbf{x};\mathbf{P}_r)\widetilde{f}_\star(\mathbf{x};\mathbf{P}_r)^2}{\widetilde{f}(\mathbf{x};\mathbf{P}_r)f(\mathbf{x};\mathbf{P}_r)^2}. \tag{A.32}$$

Since $\boldsymbol{\lambda}^{(1)}_{\mathbf{P}_r}(\mathbf{x}) \equiv \boldsymbol{\lambda}(\mathbf{x})f(\mathbf{x};\mathbf{P}_r)^{-1}$ and $\boldsymbol{\lambda}^{(2)}_{\mathbf{P}_r}(\mathbf{x}) \equiv \boldsymbol{\lambda}(\mathbf{x})l(\mathbf{x};\mathbf{P}_r)f(\mathbf{x};\mathbf{P}_r)^{-2}$ are bounded a.s. $[\mathbb{P}_{\mathbf{X}}]$ due to Assumption 1.2 (iii)-(iv) and the boundedness of $\boldsymbol{\lambda}(.)$, using these as choices of '$\boldsymbol{\lambda}(.)$' in (A.1), we have:

$$\mathbb{G}^*_n\{\boldsymbol{\lambda}^{(1)}_{\mathbf{P}_r}(.)\widetilde{l}_\star(.;\mathbf{P}_r)\} = \mathbb{G}^*_n\{\boldsymbol{\lambda}(.)\widetilde{\mathbf{T}}^{(1)}_{n,\mathbf{P}_r}(.)\} = O_p(b^{(1)}_n),$$
$$\mathbb{G}^*_n\{\boldsymbol{\lambda}^{(2)}_{\mathbf{P}_r}(.)\widetilde{f}_\star(.;\mathbf{P}_r)\} = \mathbb{G}^*_n\{\boldsymbol{\lambda}(.)\widetilde{\mathbf{T}}^{(2)}_{n,\mathbf{P}_r}(.)\} = O_p(b^{(1)}_n).$$

Further, for each $s \in \{3,4\}$, $\sup_{\mathbf{x}\in\mathcal{X}} \|\widetilde{\mathbf{T}}^{(s)}_{n,\mathbf{P}_r}(\mathbf{x})\| \leq O_p(a^2_{n,2})$ which follows from repeated use of (A.19) along with Assumption 1.2 (iii)-(iv). Consequently, with $\boldsymbol{\lambda}(.)$ bounded a.s. $[\mathbb{P}_{\mathbf{X}}]$, for each $s \in \{3,4\}$, $\mathbb{G}^*_n\{\boldsymbol{\lambda}(.)\widetilde{\mathbf{T}}^{(s)}_{n,\mathbf{P}_r}(.)\}$ is bounded by: $O_p(n^{\frac{1}{2}}a^2_{n,2})$. Combining all these results using the linearity of $\mathbb{G}^*_n(.)$, we finally obtain: $\mathbb{G}^*_n\{\boldsymbol{\lambda}(.)\widetilde{m}_\star(.;\mathbf{P}_r)\} = O_p(b^{(1)}_n + n^{\frac{1}{2}}a^2_{n,2})$ $= O_p(n^{\frac{1}{2}}a^2_{n,2})$, thus leading to (A.2). The proof of the lemma is now complete. ∎

A.5.2 Proof of Lemma A.3

Throughout this proof, all additional notations introduced, if not explicitly defined, are understood to have been adopted from the proof of Theorem 1.3 in section IV. Now, using (A.21), $\widehat{\varphi}^{(\varrho)}(\mathbf{x};\widehat{\mathbf{P}}_r) - \widetilde{\varphi}^{(\varrho)}(\mathbf{x};\mathbf{P}_r) = \text{trace}\{(\widehat{\mathbf{P}}'_r - \mathbf{P}'_r)\widehat{\mathbf{M}}^{(1)}_{n,\varrho,\mathbf{x}}\} + \text{trace}\{(\widehat{\mathbf{P}}'_r - \mathbf{P}'_r)\widehat{\mathbf{M}}^{(2)}_{n,\varrho,\mathbf{x}}\}$, and $\widehat{\mathbf{M}}^{(1)}_{n,\varrho,\mathbf{x}} = \widehat{\mathbf{M}}^{(1,1)}_{n,\varrho,\mathbf{x}} - \widehat{\mathbf{M}}^{(1,2)}_{n,\varrho,\mathbf{x}}$ as defined in section IV. Thus,

$$\mathbb{G}^*_n[\boldsymbol{\lambda}(.)\{\widehat{\varphi}^{(\varrho)}(.;\widehat{\mathbf{P}}_r) - \widetilde{\varphi}^{(\varrho)}(.;\mathbf{P}_r)\}] = \mathbb{G}^*_n\left\{\widehat{\boldsymbol{\zeta}}^{(1,1)}_{n,\varrho,\boldsymbol{\lambda}}(.) - \widehat{\boldsymbol{\zeta}}^{(1,2)}_{n,\varrho,\boldsymbol{\lambda}}(.) + \widehat{\boldsymbol{\zeta}}^{(2)}_{n,\varrho,\boldsymbol{\lambda}}(.)\right\},$$

where $\forall\ (\omega) \in \{(1,1),(1,2),(2)\}$, $\varrho \in \{0,1\}$, and $\mathbf{x} \in \mathcal{X}$,

$$\widehat{\boldsymbol{\zeta}}^{(\omega)}_{n,\varrho,\boldsymbol{\lambda}}(\mathbf{x}) = \boldsymbol{\lambda}(\mathbf{x})\ \text{trace}\left\{(\widehat{\mathbf{P}}'_r - \mathbf{P}'_r)\widehat{\mathbf{M}}^{(\omega)}_{n,\varrho,\mathbf{x}}\right\}. \tag{A.33}$$

Then, $\forall\ s \in \{1,2\}$ and $l \in \{1,\ldots,d\}$, each element of

$$\int \boldsymbol{\lambda}_{[l]}(\mathbf{x})\widehat{\mathbf{M}}^{(1,s)}_{n,\varrho,\mathbf{x}}\mathbb{P}_n(d\mathbf{x}) = n^{-2}\sum_{i_1=1}^{n}\sum_{i_2=1}^{n}\mathbb{H}^{(n,s)}_{l,\varrho}(i_1,i_2)$$

is a V-statistic, where

$$\mathbb{H}^{(n,s)}_{l,\varrho}(i_1,i_2) = h^{-(r+1)}\boldsymbol{\lambda}_{[l]}(\mathbf{X}_{i_1})Y^{\varrho}_{i_2}\mathbf{U}^{(s)}(i_1,i_2)\boldsymbol{\nabla}K'\{\mathbf{P}'_r(\mathbf{X}_{i_1}-\mathbf{X}_{i_2})/h\}$$

with $\mathbf{U}^{(1)}(i_1,i_2) = \mathbf{X}_{i_1}$ and $\mathbf{U}^{(2)}(i_1,i_2) = \mathbf{X}_{i_2}$. Hence, similar to the proof of (A.30), using Lemma 8.4 of Newey and McFadden (1994) with $\mathcal{X}$ compact, $\boldsymbol{\nabla}K(.)$ and $\boldsymbol{\lambda}(.)$ bounded, and Y^{ϱ} having finite 2^{nd} moments, we have: for each $l \in \{1,\ldots,d\}$, $s \in \{1,2\}$ and $\varrho \in \{0,1\}$,

$$\left\|\mathbb{G}^*_n\left[\boldsymbol{\lambda}_{[l]}(.)\widehat{\mathbf{M}}^{(1,s)}_{n,\varrho,(.)} - \mathbb{E}_{\mathcal{L}}\left\{\boldsymbol{\lambda}_{[l]}(.)\widehat{\mathbf{M}}^{(1,s)}_{n,\varrho,(.)}\right\}\right]\right\|_{\max} = O_p\left(n^{-\frac{1}{2}}h^{-(r+1)}\right).$$

It then follows from $(\widehat{\mathbf{P}}_r - \mathbf{P}_r) = O_p(\alpha_n)$ that for each s and ϱ,

$$\mathbb{G}^*_n\left[\widehat{\boldsymbol{\zeta}}^{(1,s)}_{n,\varrho,\boldsymbol{\lambda}}(.) - \mathbb{E}_{\mathcal{L}}\left\{\widehat{\boldsymbol{\zeta}}^{(1,s)}_{n,\varrho,\boldsymbol{\lambda}}(.)\right\}\right] = O_p\left(\alpha_n n^{-\frac{1}{2}}h^{-(r+1)}\right). \tag{A.34}$$

Next, for any given l, s and ϱ, each element of $n^{\frac{1}{2}}\mathbb{G}^*_n[\mathbb{E}_{\mathcal{L}}\{\boldsymbol{\lambda}_{[l]}(.)\widehat{\mathbf{M}}^{(1,s)}_{n,\varrho,(.)}\}]$ is a centered sum of i.i.d. random variables which are bounded by:

$$\left\|\sup_{\mathbf{x}\in\mathcal{X}}\left\{\|\boldsymbol{\lambda}(\mathbf{x})\|\ |\mathbb{E}_{\mathcal{L}}(\widehat{\mathbf{M}}^{(1,s)}_{n,\varrho,\mathbf{x}})|\right\}\right\|_{\max} = O(1),$$

where the order follows from (A.24), (A.25) and the boundedness of $\boldsymbol{\lambda}(.)$. Hence, similar to the proof of (A.31), using Hoeffding's inequality and that $(\widehat{\mathbf{P}}_r - \mathbf{P}_r) = O_p(\alpha_n)$, we have:

$\forall\, l \in \{1,\ldots,d\}$, $s \in \{1,2\}$ and $\varrho \in \{0,1\}$,

$$\left\|\mathbb{G}_n^*\left[\mathbb{E}_{\mathcal{L}}\left\{\boldsymbol{\lambda}_{[l]}(.)\widehat{\mathbf{M}}_{n,\varrho,(.)}^{(1,s)}\right\}\right]\right\|_{\max} = O_p(1) \Rightarrow \mathbb{G}_n^*\left[\mathbb{E}_{\mathcal{L}}\left\{\widehat{\boldsymbol{\zeta}}_{n,\varrho,\boldsymbol{\lambda}}^{(1,s)}(.)\right\}\right] = O_p(\alpha_n). \tag{A.35}$$

For any matrix $\mathbf{A}$, let us denote by $\mathbf{A}_{[a,b]}$ the $(a,b)^{th}$ element of $\mathbf{A}$. Now, to control $\mathbb{G}_n^*\{\widehat{\boldsymbol{\zeta}}_{n,\varrho,\boldsymbol{\lambda}}^{(2)}(.)\}$ in (A.33), note that $\|\mathbb{G}_n^*\{\widehat{\boldsymbol{\zeta}}_{n,\varrho,\boldsymbol{\lambda}}^{(2)}(.)\}\|$ is bounded by:

$$\begin{aligned}
&n^{\frac{1}{2}} \sup_{\mathbf{x}\in\mathcal{X}} \|\boldsymbol{\lambda}(\mathbf{x})\| \sum_{a,b} \int \left|(\widehat{\mathbf{P}}_r' - \mathbf{P}_r')_{[b,a]} \left(\widehat{\mathbf{M}}_{n,\varrho,\mathbf{x}}^{(2)}\right)_{[a,b]}\right| (\mathbb{P}_n + \mathbb{P}_{\mathbf{X}})(d\mathbf{x}) \\
&\leq n^{\frac{1}{2}} rp \sup_{\mathbf{x}\in\mathcal{X},\mathbf{X}\in\mathcal{X}} \{\|\boldsymbol{\lambda}(\mathbf{x})\|\,\|\mathbf{x}-\mathbf{X}\|\} \left\|\widehat{\mathbf{P}}_r - \mathbf{P}_r\right\|_{\max} \widehat{\mathbb{Z}}_n^{\varrho *} \\
&\leq O_p\left(n^{\frac{1}{2}}\alpha_n\right) \widehat{\mathbb{Z}}_n^{\varrho *},
\end{aligned} \tag{A.36}$$

where the last step follows from $(\widehat{\mathbf{P}}_r - \mathbf{P}_r) = O_p(\alpha_n)$ and the boundedness of $\mathcal{X}$ and $\boldsymbol{\lambda}(.)$, and $\widehat{\mathbb{Z}}_n^{\varrho *} = \int \widehat{\mathbb{Z}}_n^{(\varrho)}(\mathbf{x})\,(\mathbb{P}_n + \mathbb{P}_{\mathbf{X}})(d\mathbf{x})$ with

$$\widehat{\mathbb{Z}}_n^{(\varrho)}(\mathbf{x}) = n^{-1} \sum_{i=1}^{n} \frac{|Y_i^{\varrho}|}{h^{r+1}} \left\|\boldsymbol{\nabla} K(\mathbf{w}_{i,\mathbf{x}}) - \boldsymbol{\nabla} K\left\{\frac{\mathbf{P}_r'(\mathbf{x}-\mathbf{X}_i)}{h}\right\}\right\|.$$

Now, $\|\mathbf{w}_{i,\mathbf{x}} - \mathbf{P}_r'(\mathbf{x}-\mathbf{X}_i)h^{-1}\| \leq \|(\widehat{\mathbf{P}}_r - \mathbf{P}_r)'(\mathbf{x}-\mathbf{X}_i)h^{-1}\| \leq O_p(\alpha_n h^{-1})$ uniformly in $(i,\mathbf{x})$, as noted while proving (A.22). Further, with L^*, as defined in Assumption 1.2 (vii), let $\mathbb{A}_n$ denote the event: $\{\|(\widehat{\mathbf{P}}_r - \mathbf{P}_r)'(\mathbf{x}-\mathbf{X}_i)h^{-1}\| \leq L^* \;\; \forall\; \mathbf{x} \in \mathcal{X},\; i = 1,..,n\}$. Then, with $(\widehat{\mathbf{P}}_r - \mathbf{P}_r) = O_p(\alpha_n)$, $\mathcal{X}$ compact and $\alpha_n h^{-1} = o(1)$ since $n^{\frac{1}{2}}\alpha_n^2 h^{-2} = o(1)$ as assumed, it follows that $\mathbb{P}(\mathbb{A}_n) \to 1$. Using these along with Assumption 1.2 (vii) and the function $\phi(.)$ defined therein, we have: on $\mathbb{A}_n$ with $\mathbb{P}(\mathbb{A}_n) \to 1$,

$$\begin{aligned}
\widehat{\mathbb{Z}}_n^{(\varrho)}(\mathbf{x}) &\leq \sum_{i=1}^{n} \frac{|Y_i^{\varrho}|}{nh^{r+1}} \left\|\frac{(\widehat{\mathbf{P}}_r - \mathbf{P}_r)'(\mathbf{x}-\mathbf{X}_i)}{h}\right\| \phi\left\{\frac{\mathbf{P}_r'(\mathbf{x}-\mathbf{X}_i)}{h}\right\} \\
&\leq \sqrt{rp} \sup_{\mathbf{x}\in\mathcal{X},\mathbf{X}\in\mathcal{X}} \|\mathbf{x}-\mathbf{X}\| \left\|\widehat{\mathbf{P}}_r - \mathbf{P}_r\right\|_{\max} \sum_{i=1}^{n} \frac{|Y_i^{\varrho}|}{nh^{r+2}}\, \phi\left\{\frac{\mathbf{P}_r'(\mathbf{x}-\mathbf{X}_i)}{h}\right\}.
\end{aligned}$$

Thus, $\widehat{\mathbb{Z}}_n^{\varrho *} \leq O_p\left(\alpha_n \widetilde{\mathbb{Z}}_n^{\varrho *}\right)$, where $\widetilde{\mathbb{Z}}_n^{\varrho *} = \int \widetilde{\mathbb{Z}}_n^{(\varrho)}(\mathbf{x})(\mathbb{P}_n + \mathbb{P}_{\mathbf{X}})(d\mathbf{x})$,

$$\widetilde{\mathbb{Z}}_n^{(\varrho)}(\mathbf{x}) = n^{-1}\sum_{i=1}^{n} \widetilde{\mathbb{Z}}_n^{(\varrho)}(\mathbf{x}; \mathbf{Z}_i), \text{ and } \widetilde{\mathbb{Z}}_n^{(\varrho)}(\mathbf{x}; \mathbf{Z}) = \frac{|Y^{\varrho}|}{h^{r+2}}\phi\left\{\frac{\mathbf{P}_r'(\mathbf{x} - \mathbf{X})}{h}\right\}.$$

Let $\mathbf{Z}^0 \equiv (Y^0, \mathbf{X}^{0\prime})' \sim \mathbb{P}_{\mathbf{Z}}$ be generated independent of $\mathcal{L}$, and define:

$$\widetilde{\mathbb{U}}_{n,\varrho}^{(1)} = n^{-1}\sum_{i=1}^{n} \mathbb{E}_{\mathbf{X}^0}\{\widetilde{\mathbb{Z}}_n^{(\varrho)}(\mathbf{X}^0; \mathbf{Z}_i)\}, \quad \widetilde{\mathbb{U}}_{n,\varrho}^{(2)} = n^{-1}\sum_{i=1}^{n} \mathbb{E}_{\mathbf{Z}^0}\{\widetilde{\mathbb{Z}}_n^{(\varrho)}(\mathbf{X}_i; \mathbf{Z}^0)\},$$
$$\widetilde{\mathbb{U}}_{n,\varrho}^{(1,1)} = \mathbb{E}\{\widetilde{\mathbb{Z}}_n^{(\varrho)}(\mathbf{X}^0; \mathbf{Z}^0)\}, \quad \text{and} \qquad \widetilde{\mathbb{V}}_{n,\varrho}^{(k)} = \mathbb{E}\{\widetilde{\mathbb{Z}}_n^{(\varrho)}(\mathbf{X}^0; \mathbf{Z})^k\} \text{ for } k = 1, 2.$$

Then, first note that: $\int \widetilde{\mathbb{Z}}_n^{(\varrho)}(\mathbf{x})\mathbb{P}_{\mathbf{X}}(d\mathbf{x}) = \widetilde{\mathbb{U}}_{n,\varrho}^{(1)}$. Further, since

$$\int \widetilde{\mathbb{Z}}_n^{(\varrho)}(\mathbf{x})\mathbb{P}_n(d\mathbf{x}) = n^{-2}\sum_{i_1=1}^{n}\sum_{i_2=1}^{n} \widetilde{\mathbb{Z}}_n^{(\varrho)}(\mathbf{X}_{i_1}; \mathbf{Z}_{i_2})$$

is a V-statistic, we have:

$$\int \widetilde{\mathbb{Z}}_n^{(\varrho)}(\mathbf{x})\mathbb{P}_n(d\mathbf{x}) = \widetilde{\mathbb{U}}_{n,\varrho}^{(1)} + \widetilde{\mathbb{U}}_{n,\varrho}^{(2)} - \widetilde{\mathbb{V}}_{n,\varrho}^{(1)} + O_p\{n^{-1}\widetilde{\mathbb{U}}_{n,\varrho}^{(1,1)} + n^{-1}(\widetilde{\mathbb{V}}_{n,\varrho}^{(2)})^{\frac{1}{2}}\}$$

using Lemma 8.4 of Newey and McFadden (1994). Then, with all notations as above, we have:

$$n^{-1}\widetilde{\mathbb{U}}_{n,\varrho}^{(1,1)} + n^{-1}(\widetilde{\mathbb{V}}_{n,\varrho}^{(2)})^{\frac{1}{2}} \leq O_p\left(n^{-1}h^{-(r+2)}\right), \tag{A.37}$$

$$\begin{aligned} \text{and} \qquad \widetilde{\mathbb{U}}_{n,\varrho}^{(1)} &= \frac{1}{nh^{r+2}}\sum_{i=1}^{n} |Y_i^{\varrho}| \int_{\mathcal{X}_{\mathbf{P}_r}} \phi\left(\frac{\mathbf{w} - \mathbf{P}_r'\mathbf{X}_i}{h}\right) f_{\mathbf{P}_r}(\mathbf{w})d\mathbf{w} \\ &\leq \frac{B_{\mathbf{P}_r}}{nh^2}\sum_{i=1}^{n}\left\{|Y_i^{\varrho}| \int_{A_{\mathbf{X}_i}^n} \phi(\boldsymbol{\psi}_i)d\boldsymbol{\psi}_i\right\}, \\ &\leq \frac{B_{\mathbf{P}_r}}{h^2}\left\{\int_{\mathbb{R}^r} \phi(\boldsymbol{\psi})d\boldsymbol{\psi}\right\}\left\{n^{-1}\sum_{i=1}^{n} |Y_i^{\varrho}|\right\} \leq O_p\left(h^{-2}\right), \end{aligned} \tag{A.38}$$

where $\boldsymbol{\psi}_i = h^{-1}(\mathbf{w} - \mathbf{P}_r'\mathbf{X}_i)\ \forall\ i$, $A_{\mathbf{x}}^n = \{\boldsymbol{\psi} : (\mathbf{P}_r'\mathbf{x} + h\boldsymbol{\psi}) \in \mathcal{X}_{\mathbf{P}_r}\}\ \forall\ \mathbf{x} \in \mathcal{X}$, and $B_{\mathbf{P}_r} =$

$\sup_{\mathbf{w}\in\mathcal{X}_{\mathbf{P}_r}} f_{\mathbf{P}_r}(\mathbf{w}) < \infty$. The error rate in (A.37) follows since $\phi(.)$ is bounded and Y^{ϱ} has finite 2^{nd} moments, while that of $\widetilde{\mathbb{U}}^{(1)}_{n,\varrho}$ follows from Assumption 1.2 (iii), integrability of $\phi(.)$, and LLN applied to the sequence $\{Y_i^{\varrho}\}_{i=1}^n$ having finite 2^{nd} moments. Now, note that $\widetilde{\mathbb{U}}^{(2)}_{n,\varrho} - \widetilde{\mathbb{V}}^{(1)}_{n,\varrho}$ is a centered average of $[\mathbb{E}_{\mathbf{Z}^0}\{\widetilde{\mathbb{Z}}^{(\varrho)}_n(\mathbf{X}_i;\mathbf{Z}^0)\}]_{i=1}^n$ which are i.i.d. and bounded by:

$$\sup_{\mathbf{x}\in\mathcal{X}} \mathbb{E}_{\mathbf{Z}}\{\widetilde{\mathbb{Z}}^{(\varrho)}_n(\mathbf{x};\mathbf{Z})\} = \sup_{\mathbf{x}\in\mathcal{X}} \frac{1}{h^{r+2}} \int_{\mathcal{X}_{\mathbf{P}_r}} \phi\left(\frac{\mathbf{P}_r'\mathbf{x}-\mathbf{w}}{h}\right) \overline{m}^{(\varrho)}_{\mathbf{P}_r}(\mathbf{w}) f_{\mathbf{P}_r}(\mathbf{w}) d\mathbf{w},$$

where $\overline{m}^{(\varrho)}_{\mathbf{P}_r}(\mathbf{w}) = \mathbb{E}(|Y|^{\varrho} \mid \mathbf{X}_{\mathbf{P}_r} = \mathbf{w})\ \forall\ \varrho \in \{0,1\}$ and $\mathbf{w} \in \mathcal{X}_{\mathbf{P}_r}$. Using the integrability of $\phi(.)$, we then have:

$$\begin{aligned}
&\sup_{\mathbf{x}\in\mathcal{X}} \mathbb{E}_{\mathbf{Z}}\{\widetilde{\mathbb{Z}}^{(\varrho)}_n(\mathbf{x};\mathbf{Z})\} \ \leq\ \sup_{\mathbf{x}\in\mathcal{X}} \frac{C^{(\varrho)}_{\mathbf{P}_r}}{h^{r+2}} \int_{\mathcal{X}_{\mathbf{P}_r}} \phi\left(\frac{\mathbf{P}_r'\mathbf{x}-\mathbf{w}}{h}\right) d\mathbf{w} \\
\leq \sup_{\mathbf{x}\in\mathcal{X}} \frac{C^{(\varrho)}_{\mathbf{P}_r}}{h^2} \int_{A^n_{\mathbf{x}}} \phi(-\boldsymbol{\psi})\ d\boldsymbol{\psi} &\leq \frac{C^{(\varrho)}_{\mathbf{P}_r}}{h^2} \left\{\int_{\mathbb{R}^r} \phi(\boldsymbol{\psi}) d\boldsymbol{\psi}\right\} \ =\ O\left(h^{-2}\right),
\end{aligned}$$

where $C^{(\varrho)}_{\mathbf{P}_r} = \sup_{\mathbf{w}\in\mathcal{X}_{\mathbf{P}_r}} \overline{m}^{(\varrho)}_{\mathbf{P}_r}(\mathbf{w}) f_{\mathbf{P}_r}(\mathbf{w}) < \infty$ due to Assumption 1.2 (iii), and $A^n_{\mathbf{x}} = \{\boldsymbol{\psi} : (\mathbf{P}_r'\mathbf{x} + h\boldsymbol{\psi}) \in \mathcal{X}_{\mathbf{P}_r}\}$, as before. It then follows, similar to the proof of (A.31), from a simple application of Hoeffding's inequality that

$$\widetilde{\mathbb{U}}^{(2)}_{n,\varrho} - \widetilde{\mathbb{V}}^{(1)}_{n,\varrho} = O_p\left(n^{-\frac{1}{2}} h^{-2}\right). \tag{A.39}$$

Using (A.37)-(A.39), we finally have: $\widetilde{\mathbb{Z}}^{\varrho *}_n = O_p(h^{-2} + n^{-1}h^{-(r+2)})$. Hence,

$$\widehat{\mathbb{Z}}^{\varrho *}_n = \int \widehat{\mathbb{Z}}^{(\varrho)}_n(\mathbf{x})\, (\mathbb{P}_n + \mathbb{P}_{\mathbf{X}})(d\mathbf{x}) \leq O_p\left(\alpha_n \widetilde{\mathbb{Z}}^{\varrho *}_n\right) = O_p\left(\frac{\alpha_n}{h^2} + \frac{\alpha_n}{nh^{r+2}}\right), \tag{A.40}$$

$$\text{and} \quad \left\|\mathbb{G}^*_n\left\{\widehat{\boldsymbol{\zeta}}^{(2)}_{n,\varrho,\boldsymbol{\lambda}}(.)\right\}\right\| \leq O_p\left(\frac{n^{\frac{1}{2}}\alpha_n^2}{h^2} + \frac{n^{\frac{1}{2}}\alpha_n^2}{nh^{r+2}}\right) \ \forall\ \varrho \in \{0,1\}, \tag{A.41}$$

where the final bound in (A.41) follows from (A.36). The desired result in (A.3) now follows by applying (A.34), (A.35) and (A.41) to (A.33) using the linearity of $\mathbb{G}^*_n(.)$. The proof of the lemma is now complete. (Note that conditions (i), (iv) and (viii) in Assumption 1.2 were

actually not used in this proof). ∎

A.5.3 Proof of Theorem 1.4

Finally, to establish the result of Theorem 1.4, let $\boldsymbol{\lambda}_0(\mathbf{x}) = \overrightarrow{\mathbf{x}}$ which is measurable and bounded on $\mathcal{X}$. Further, with $\mathbb{G}_n^*(.)$ as defined in appendix A.1, note that $\mathbb{G}_{n,\mathbb{K}}$ for $\mathbb{K} = 1$ is given by:

$$\mathbb{G}_{n,\mathbb{K}} = \mathbb{G}_n^*\{\boldsymbol{\lambda}_0(.)\widetilde{m}_\star(.;\mathbf{P}_r)\} + \mathbb{G}_n^*[\boldsymbol{\lambda}_0(.)\{\widehat{m}(.;\widehat{\mathbf{P}}_r) - \widetilde{m}(.;\mathbf{P}_r)\}], \tag{A.42}$$

due to linearity of $\mathbb{G}_n^*(.)$. Now, using Lemma A.2, we have:

$$\mathbb{G}_n^*\{\boldsymbol{\lambda}_0(.)\widetilde{m}_\star(.;\mathbf{P}_r)\} = O_p(n^{\frac{1}{2}}a_{n,2}^2) = O_p(a_{n,2}^*). \tag{A.43}$$

The second term $\mathbb{G}_n^*[\boldsymbol{\lambda}_0(.)\{\widehat{m}(.;\widehat{\mathbf{P}}_r) - \widetilde{m}(.;\mathbf{P}_r)\}]$ in (A.42) can be written as:

$$\begin{aligned}&\mathbb{G}_n^*[\boldsymbol{\lambda}_0(.)\{\widehat{\mathbf{T}}_{n,\mathbf{P}_r}^{(1)}(.) - \widehat{\mathbf{T}}_{n,\mathbf{P}_r}^{(2)}(.) - \widehat{\mathbf{T}}_{n,\mathbf{P}_r}^{(3)}(.) + \widehat{\mathbf{T}}_{n,\mathbf{P}_r}^{(4)}(.)\}]\\ &= O_p\left(b_n^{(2)} + n^{\frac{1}{2}}a_{n,1}^2 + n^{\frac{1}{2}}a_{n,1}a_{n,2}\right) = O_p\left(a_{n,1}^*\right),\end{aligned} \tag{A.44}$$

where with slight abuse of notation,

$$\begin{aligned}\widehat{\mathbf{T}}_{n,\mathbf{P}_r}^{(1)}(\mathbf{x}) &= \frac{\widehat{a} - \widetilde{a}}{b}, \quad \widehat{\mathbf{T}}_{n,\mathbf{P}_r}^{(2)}(\mathbf{x}) = \frac{a(\widehat{b} - \widetilde{b})}{b^2},\\ \widehat{\mathbf{T}}_{n,\mathbf{P}_r}^{(3)}(\mathbf{x}) &= \frac{(\widehat{a} - \widetilde{a})(\widetilde{b} - b)}{b\,\widetilde{b}} + \frac{(\widehat{a} - \widetilde{a})(\widehat{b} - \widetilde{b})}{\widetilde{b}\,\widehat{b}}, \quad \text{and}\\ \widehat{\mathbf{T}}_{n,\mathbf{P}_r}^{(4)}(\mathbf{x}) &= \frac{\widetilde{a}(\widehat{b} - \widetilde{b})^2}{\widehat{b}\,b^2} - \frac{(\widetilde{a} - a)(\widehat{b} - \widetilde{b})}{b^2} + \frac{a(\widehat{b} - \widetilde{b})(\widetilde{b} - b)(b + \widetilde{b})}{(b\,\widetilde{b})(b\,\widehat{b})},\end{aligned}$$

with $(a, b) = \{l(\mathbf{x};\mathbf{P}_r), f(\mathbf{x};\mathbf{P}_r)\}$, $(\widetilde{a}, \widetilde{b}) = \{\widetilde{l}(\mathbf{x};\mathbf{P}_r), \widetilde{f}(\mathbf{x};\mathbf{P}_r)\}$ and $(\widehat{a}, \widehat{b}) = \{\widehat{l}(\mathbf{x};\widehat{\mathbf{P}}_r), \widehat{f}(\mathbf{x};\widehat{\mathbf{P}}_r)\}\}$.

For (A.44), the starting expansion is due to a linearization similar to (A.32), while the final rate is due to the following: note that $\boldsymbol{\lambda}_{0,\mathbf{P}_r}^{(1)}(.) \equiv b^{-1}\boldsymbol{\lambda}_0(.)$ and $\boldsymbol{\lambda}_{0,\mathbf{P}_r}^{(2)}(.) \equiv ab^{-2}\boldsymbol{\lambda}_0(.)$ are both bounded a.s. $[\mathbb{P}_{\mathbf{X}}]$ due to Assumption 1.2 (iii)-(iv) and the boundedness of $\boldsymbol{\lambda}_0(.)$.

Hence, using these as choices of '$\boldsymbol{\lambda}(.)$' in Lemma A.3, we have: $\mathbb{G}_n^*\{(\widehat{a}-\widetilde{a})\boldsymbol{\lambda}_{0,\mathbf{P}_r}^{(1)}(.)\} = \mathbb{G}_n^*[\boldsymbol{\lambda}_0(.)\{\widehat{\mathbf{T}}_{n,\mathbf{P}_r}^{(1)}(.)\}] = O_p(b_n^{(2)})$ and $\mathbb{G}_n^*\{(\widehat{b}-\widetilde{b})\boldsymbol{\lambda}_{0,\mathbf{P}_r}^{(2)}(.)\} = \mathbb{G}_n^*[\boldsymbol{\lambda}_0(.)\{\widehat{\mathbf{T}}_{n,\mathbf{P}_r}^{(2)}(.)\}] = O_p(b_n^{(2)})$ respectively. Further, note that for each $s \in \{3,4\}$, $\sup_{\mathbf{x}\in\mathcal{X}}\|\widehat{\mathbf{T}}_{n,\mathbf{P}_r}^{(s)}(\mathbf{x})\| \leq O_p(a_{n,1}^2 + a_{n,1}a_{n,2})$ which follows from repeated use of (A.19), (A.28) along with Assumption 1.2 (iii)-(iv). Consequently, with $\boldsymbol{\lambda}_0(\mathbf{x})$ bounded a.s. $[\mathbb{P}_{\mathbf{X}}]$, for each $s \in \{3,4\}$, $\mathbb{G}_n^*[\boldsymbol{\lambda}_0(.)\{\widehat{\mathbf{T}}_{n,\mathbf{P}_r}^{(s)}(.)\}]$ is bounded by: $O_p(n^{\frac{1}{2}}a_{n,1}^2 + n^{\frac{1}{2}}a_{n,1}a_{n,2})$. Combining all these results using the linearity of $\mathbb{G}_n^*(.)$ and noting that with $a_{n,2}^* = o(1)$, $(b_n^{(2)} + n^{\frac{1}{2}}a_{n,1}^2 + n^{\frac{1}{2}}a_{n,1}a_{n,2}) = O(a_{n,1}^*)$, (A.44) now follows and, along with (A.43) and (A.42), implies: $\mathbb{G}_{n,\mathbb{K}} = O_p(a_{n,1}^* + a_{n,2}^*)$ as claimed in Theorem 1.4. Lastly, using this in (1.12), the expansion in (1.17) and its associated implications follow. The proof of Theorem 1.4 is now complete. ∎

Appendix B

Proofs of All Results in Chapter 3

B.1 Preliminaries

We first state a few preliminary lemmas that would be useful in the proofs of theorem 3.3-3.4.

Lemma B.1. (Properties of sub-gaussian distributions) *Let Z be any random variable with $\mathbb{E}(Z) = 0$, and suppose that Z follows a sub-gaussian distribution with parameter σ^2, to be denoted as $Z \sim \mathbb{SG}(\sigma^2)$, for some $\sigma \geq 0$, so that $\mathbb{E}\{exp(aZ)\} \leq \exp(\sigma^2 a^2/2)\ \forall\ a \in \mathbb{R}$. Then,*

(i) For any $\epsilon > 0$, $\mathbb{P}(Z > \epsilon) \leq exp\left\{-\frac{\epsilon^2}{2\sigma^2}\right\}$, and $\mathbb{P}(|Z| > \epsilon) \leq 2\ exp\left\{-\frac{\epsilon^2}{2\sigma^2}\right\}$.

(ii) For any $b \in \mathbb{R}$, $bZ \sim \mathbb{SG}(b^2\sigma^2)$. Further, for any $Z_1 \sim \mathbb{SG}(\sigma_1^2)$ and $Z_2 \sim \mathbb{SG}(\sigma_2^2)$, with Z_1 and Z_2 not necessarily independent, $(Z_1 + Z_2) \sim \mathbb{SG}\{(\sigma_1 + \sigma_2)^2\}$. If Z_1 and Z_2 are additionally independent, then $(Z_1 + Z_2) \sim \mathbb{SG}(\sigma_1^2 + \sigma_2^2)$, with an improved parameter.

(iii) For each positive integer $m \geq 2$, $\mathbb{E}(|Z|^m) \leq 2(\sqrt{2}\sigma)^m \Gamma(m/2 + 1)$, where $\Gamma(.)$ denotes the standard gamma function given by: $\Gamma(t) = \int_0^\infty x^{t-1}e^{-x}\ dx\ \ \forall\ t \geq 0$.

(iv) For any collection $\{Z_j\}_{j=1}^m$ of sub-gaussian random variables (not necessarily independent) with parameter σ^2, $\mathbb{P}\left(\max_{1\leq j\leq m} |Z_j| > \epsilon\right) \leq 2\ exp\left\{-\frac{\epsilon^2}{2\sigma^2} + log\, m\right\}$, for any $\epsilon > 0$.

(v) A random vector $\mathbf{Z} \in \mathbb{R}^d$ for any d, with $\mathbb{E}(\mathbf{Z}) = \mathbf{0}$, is said to follow a sub-gaussian distribution with parameter σ^2, denoted as $\mathbf{Z} \sim \mathbb{SG}(\sigma^2)$, for some $\sigma \geq 0$, if $\forall\ \mathbf{t} \in \mathbb{R}^d$, the random variable $\mathbf{t}'\mathbf{Z} \sim \mathbb{SG}\{\sigma^2(t)\}$, for some $\sigma(t) \geq 0$ such that $\sigma^2(t) \leq \sigma^2\|\mathbf{t}\|_2^2$. For

any collection $\{\mathbf{Z}_j\}_{j=1}^m$ *of random vectors (not necessarily independent) in* $\mathbb{R}^d$ *such that* $\mathbf{Z}_j \sim \mathbb{SG}(\sigma^2)\ \forall\ j$, $\mathbb{P}\left(\max_{1\leq j\leq m} \|\mathbf{Z}_j\|_\infty > \epsilon\right) \leq 2\ exp\left\{-\frac{\epsilon^2}{2\sigma^2} + log\,(md)\right\}$, *for any* $\epsilon > 0$.

Lemma B.2. (Sub-gaussian properties for binary variables) *Let* $Z \in \{0,1\}$ *be a binary random variable with* $\mathbb{E}(Z) \equiv \mathbb{P}(Z=1) = p \in [0,1]$. *Let* $\widetilde{Z} = (Z-p)$ *denote the corresponding centered version of* Z. *Then,* $\widetilde{Z} \sim \mathbb{SG}(\widetilde{p}^{\,2})$, *where* $\widetilde{p} > 0$ *is given by:* $\widetilde{p} = 0$ *if* $p \in \{0,1\}$, $\widetilde{p} = 1/2$ *if* $p = 1/2$, *and* $\widetilde{p} = [(p-1/2)/log\,\{p/(1-p)\}]^{1/2}$ *if* $p \notin \{0,1,1/2\}$.

Lemma B.3. (Bernstein's inequality) *Let* $\{\mathbb{Z}_1, \ldots, Z_n\}$ *denote any collection of* n *independent (not necessarily i.i.d.) random variables* $\in \mathbb{R}$, *such that* $\mathbb{E}(Z_i) = 0\ \forall\ 1 \leq i \leq n$. *Suppose* $\exists$ *constants* $\sigma \geq 0$ *and* $K \geq 0$, *such that* $n^{-1}\sum_{i=1}^n \mathbb{E}(|Z_i|^m \leq (m!/2)K^{m-2}\sigma^2$, *for each positive integer* $m \geq 2$. *Then, the following concentration bound holds:*

$$\mathbb{P}\left(\frac{1}{n}\left|\sum_{i=1}^n Z_i\right| \geq \sqrt{2}\sigma\epsilon + K\epsilon^2\right)\ \leq\ 2\ exp\,(-n\epsilon^2) \quad for\ any\ \epsilon > 0.$$

Lemma B.4. (Useful bounds for the standard normal density and c.d.f.) *Let* $\phi(.)$ *and* $\Phi(.)$ *respectively denote the density and the c.d.f. of the standard* $\mathcal{N}_1(0,1)$ *distribution. Further, let* $\overline{\Phi}(t) = \{1-\Phi(t)\} \equiv \Phi(-t)\ \forall\ t \in \mathbb{R}$. *Then the following bounds hold: for any* $t > 0$,

$$\frac{t}{1+t^2}\phi(t)\ \leq\ \overline{\Phi}(t)\ \leq\ \frac{\phi(t)}{t},\ \ and\ \ \overline{\Phi}(t)\ \leq\ exp\left(-\frac{t^2}{2}\right).$$

Lemma B.5. (Properties of the truncated normal distribution) *Let* $Z \sim \mathcal{N}_1(0,\sigma^2)$ *distribution for some* $\sigma > 0$, *and let* $\phi(.)$ *and* $\Phi(.)$ *respectively denote the density and the c.d.f. of the standard* $\mathcal{N}_1(0,1)$ *distribution. For any* a, b *such that* $-\infty \leq a < b \leq \infty$, *consider the truncated random variable:* $Z_{a,b} \equiv (Z \mid a \leq Z \leq b)$. *Let* $\overline{a} = (a/\sigma)$ *and* $\overline{b} = (b/\sigma)$. *Then,*

(i) For any a, b *as above,* $Z_{a,b}$ *satisfies the following distributional properties:*

$$\mathbb{E}(Z \mid a \leq Z \leq b)\ =\ \sigma\frac{\phi(\overline{a}) - \phi(\overline{b})}{\Phi(\overline{b}) - \Phi(\overline{a})},\ \ \mathbb{E}(Z^2 \mid a \leq Z \leq b)\ =\ \sigma^2\left\{1 + \frac{\overline{a}\phi(\overline{a}) - \overline{b}\phi(\overline{b})}{\Phi(\overline{b}) - \Phi(\overline{a})}\right\},$$

$$\textit{and } \; MGF_{Z_{a,b}}(t) \equiv \mathbb{E}\left(e^{tZ} \mid a \leq Z \leq b\right) \; = \; e^{\sigma^2 t^2/2} \left\{ \frac{\Phi(\overline{b} - \sigma t) - \Phi(\overline{a} - \sigma t)}{\Phi(\overline{b}) - \Phi(\overline{a})} \right\} \; \forall \, t \in \mathbb{R}.$$

(ii) For any $q \in (0,1]$, let z_q and $\overline{z}_q$ respectively denote the $(q/2)^{th}$ and $(1-q/2)^{th}$ quantiles of the standard $\mathcal{N}_1(0,1)$ distribution, so that $-z_q = \overline{z}_q \geq 0$. Consider the function: $f_q(t) = \frac{1}{2\Phi(z_q)}\{\Phi(z_q + t) + \Phi(z_q - t)\}$ $\forall \, t \in \mathbb{R}$. Then, $f_q(.)$ satisfies: for any $t \in \mathbb{R}$,

$$f_q(t) \; \leq \; exp\left(t^2 \overline{z}_q^2\right) \; \forall \, q \in (0, 1/2], \;\; while \;\; f_q(t) \; \leq \; 2 \; trivially \; \forall \, q \in (1/2, 1].$$

Lemma B.1 is a collection of several well-known properties of sub-gaussian distributions, and proofs and/or discussions of these results (or equivalent versions) can be found in several relevant references, including Vershynin (2010) for instance. Lemma B.2 explicitly characterizes the sub-gaussian properties of (centered) binary random variables, and its proof can be found in Buldygin and Moskvichova (2013). Lemma B.3 is one of many versions of the well-known Bernstein's inequality, and this particular version has been adopted from Van de Geer and Lederer (2013). Lemma B.4 provides some useful and fairly well known bounds involving the standard normal c.d.f. and density, and their mentions and/or discussions can be found in Düembgen (2010), Chiani et al. (2003) and the references cited therein. Lastly, lemma B.5 provides some useful distributional properties of truncated normal distributions. For the results in (i), proofs and/or mentions of them (or much more general versions) can be found in a combination of references including Tallis (1961), Johnson et al. (1994), Horrace (2004, 2005) and Burkardt (2014). Result (ii) in lemma B.5 is a fairly straightforward conclusion, and can be obtained, for instance, through direct numerical verification. We skip the details here for the sake of brevity, and leave them to the interested reader to verify.

B.2 Proof of Theorem 3.1

We first note that owing to our model based assumptions (3.4)-(3.5), and the linearity condition (3.8) in assumption 3.2, the following identities hold: $\forall\ \mathbf{v} \in \mathbb{R}^p$,

$$\mathbb{E}(\mathbf{v}'\mathbf{X} \mid \boldsymbol{\alpha}_0'\mathbf{X}, \boldsymbol{\beta}_0'\mathbf{X}, \epsilon, \epsilon^*) = \mathbb{E}(\mathbf{v}'\mathbf{X} \mid \boldsymbol{\alpha}_0'\mathbf{X}, \boldsymbol{\beta}_0'\mathbf{X}) = c_{\mathbf{v}} + a_{\mathbf{v}}(\boldsymbol{\alpha}_0'\mathbf{X}) + b_{\mathbf{v}}(\boldsymbol{\beta}_0'\mathbf{X}), \tag{B.1}$$

$$\text{and } \mathbf{v}'\boldsymbol{\mu}_q \equiv \mathbb{E}_q(\mathbf{v}'\mathbf{X}) = \mathbb{E}_q\{\mathbb{E}_q(\mathbf{v}'\mathbf{X} \mid \boldsymbol{\alpha}_0'\mathbf{X}, \boldsymbol{\beta}_0'\mathbf{X}, \epsilon, \epsilon^*)\} = \mathbb{E}_q\{\mathbb{E}(\mathbf{v}'\mathbf{X} \mid \boldsymbol{\alpha}_0'\mathbf{X}, \boldsymbol{\beta}_0'\mathbf{X}, \epsilon, \epsilon^*)\}$$

$$= \mathbb{E}_q\{c_{\mathbf{v}} + a_{\mathbf{v}}(\boldsymbol{\alpha}_0'\mathbf{X}) + b_{\mathbf{v}}(\boldsymbol{\beta}_0'\mathbf{X})\} = c_{\mathbf{v}} + a_{\mathbf{v}}(\boldsymbol{\alpha}_0'\boldsymbol{\mu}_q) + b_{\mathbf{v}}(\boldsymbol{\beta}_0'\boldsymbol{\mu}_q), \tag{B.2}$$

where, for the second equality in obtaining (B.2), we have used the fact that owing to (3.5), S is completely determined by the conditioning variables $\{\boldsymbol{\beta}_0'\mathbf{X}, \boldsymbol{\alpha}_0'\mathbf{X}, \epsilon, \epsilon^*\}$, so that the term $\mathbb{E}_q(.\mid \boldsymbol{\beta}_0'\mathbf{X}, \boldsymbol{\alpha}_0'\mathbf{X}, \epsilon, \epsilon^*)$ inside can be replaced by $\mathbb{E}(.\mid \boldsymbol{\beta}_0'\mathbf{X}, \boldsymbol{\alpha}_0'\mathbf{X}, \epsilon, \epsilon^*)$. Note further that for all the steps in obtaining (B.2), it is implicitly understood, as would be the case henceforth, that the values assumed by the conditioning variables $\{\boldsymbol{\beta}_0'\mathbf{X}, \boldsymbol{\alpha}_0'\mathbf{X}, \epsilon, \epsilon^*\}$ are such that the underlying restriction $\{S \in \mathcal{I}_q\}$ is indeed feasible so that $\mathbb{E}_q(.|\boldsymbol{\beta}_0'\mathbf{X}, \boldsymbol{\alpha}_0'\mathbf{X}, \epsilon, \epsilon^*)$ is well-defined.

Next, note that that expected squared loss function $\mathbb{L}_q(\mathbf{v})$ satisfies: $\forall\ \mathbf{v} \in \mathbb{R}^p$,

$$\begin{aligned}
\mathbb{L}_q(\mathbf{v}) &\equiv \mathbb{E}_q[\{Y - p_q - \mathbf{v}'(\mathbf{X} - \boldsymbol{\mu}_q)\}^2] \\
&= \mathbb{E}_q(\mathbb{E}_q[\{Y - p_q - \mathbf{v}'(\mathbf{X} - \boldsymbol{\mu}_q)\}^2 \mid \boldsymbol{\beta}_0'\mathbf{X}, \boldsymbol{\alpha}_0'\mathbf{X}, \epsilon, \epsilon^*]) \\
&= \mathbb{E}_q(\mathbb{E}[\{Y - p_q - \mathbf{v}'(\mathbf{X} - \boldsymbol{\mu}_q)\}^2 \mid \boldsymbol{\beta}_0'\mathbf{X}, \boldsymbol{\alpha}_0'\mathbf{X}, \epsilon, \epsilon^*]) \\
&\geq \mathbb{E}_q\left\{\left(\mathbb{E}[\{Y - p_q - \mathbf{v}'(\mathbf{X} - \boldsymbol{\mu}_q)\} \mid \boldsymbol{\beta}_0'\mathbf{X}, \boldsymbol{\alpha}_0'\mathbf{X}, \epsilon, \epsilon^*]\right)^2\right\} \\
&= \mathbb{E}_q\left[\left\{\mathbb{E}(Y \mid \boldsymbol{\beta}_0'\mathbf{X}, \boldsymbol{\alpha}_0'\mathbf{X}, \epsilon, \epsilon^*) - p_q - \mathbb{E}(\mathbf{v}'\mathbf{X} \mid \boldsymbol{\beta}_0'\mathbf{X}, \boldsymbol{\alpha}_0'\mathbf{X}, \epsilon, \epsilon^*) + \mathbf{v}'\boldsymbol{\mu}_q\right\}^2\right] \\
&= \mathbb{E}_q\left[\left\{Y - p_q - a_{\mathbf{v}}(\boldsymbol{\alpha}_0'\mathbf{X}) - b_{\mathbf{v}}(\boldsymbol{\beta}_0'\mathbf{X}) + a_{\mathbf{v}}(\boldsymbol{\alpha}_0'\boldsymbol{\mu}_q) + b_{\mathbf{v}}(\boldsymbol{\beta}_0'\boldsymbol{\mu}_q)\right\}^2\right] \\
&= \mathbb{E}_q\left[\left\{Y - p_q - (a_{\mathbf{v}}\boldsymbol{\alpha}_0 + b_{\mathbf{v}}\boldsymbol{\beta}_0)'(\mathbf{X} - \boldsymbol{\mu}_q)\right\}^2\right] \equiv \mathbb{L}_q(a_{\mathbf{v}}\boldsymbol{\alpha}_0 + b_{\mathbf{v}}\boldsymbol{\beta}_0).
\end{aligned} \tag{B.3}$$

The second equality in obtaining (B.3) follows from arguments similar to those mentioned earlier while obtaining (B.2). The subsequent inequality follows from (conditional) Jensen's

inequality, while in the penultimate step we have used (B.1)-(B.2), as well as the fact that owing to (3.4), Y is completely determined (hence constant) by the conditioning variables $\{\boldsymbol{\beta}_0'\mathbf{X}, \boldsymbol{\alpha}_0'\mathbf{X}, \epsilon, \epsilon^*\}$. Thus, (B.3) now shows that the value of $\mathbb{L}_q(.)$ at every $\mathbf{v} \in \mathbb{R}^p$ is bounded below by its value at a corresponding point of the form $(a_{\mathbf{v}}\boldsymbol{\alpha}_0 + b_{\mathbf{v}}\boldsymbol{\beta}_0) \in \mathbb{R}^p$. In particular, this also applies to $\mathbf{v} = \overline{\boldsymbol{\beta}}_q$, which however is the unique minimizer of $\mathbb{L}_q(\mathbf{v})$ over $\mathbf{v} \in \mathbb{R}^p$. Hence, $\overline{\boldsymbol{\beta}}_q$ must be of the form $(a_{\mathbf{v}}\boldsymbol{\alpha}_0 + b_{\mathbf{v}}\boldsymbol{\beta}_0)$ with $\mathbf{v} = \overline{\boldsymbol{\beta}}_q$. This establishes (3.10). ∎

To show (3.11), we first note that owing to (3.5) and (3.8)-(3.9), we have: $\forall\ \mathbf{v} \in \mathbb{R}^p$,

$$\mathbb{E}(\mathbf{v}'\mathbf{X} \mid \boldsymbol{\alpha}_0'\mathbf{X}, \epsilon^*) = \mathbb{E}(\mathbf{v}'\mathbf{X} \mid \boldsymbol{\alpha}_0'\mathbf{X}) = (c_{\mathbf{v}} + b_{\mathbf{v}}\overline{c}) + (a_{\mathbf{v}} + b_{\mathbf{v}}\overline{a})(\boldsymbol{\alpha}_0'\mathbf{X}), \text{ and} \tag{B.4}$$

$$\begin{aligned}\mathbf{v}'\boldsymbol{\mu}_q \equiv \mathbb{E}_q(\mathbf{v}'\mathbf{X}) &= \mathbb{E}_q\{\mathbb{E}_q(\mathbf{v}'\mathbf{X} \mid \boldsymbol{\alpha}_0'\mathbf{X}, \epsilon^*)\} = \mathbb{E}_q\{\mathbb{E}(\mathbf{v}'\mathbf{X} \mid \boldsymbol{\alpha}_0'\mathbf{X}, \epsilon^*)\} \\ &= \mathbb{E}_q\{(c_{\mathbf{v}} + b_{\mathbf{v}}\overline{c}) + (a_{\mathbf{v}} + b_{\mathbf{v}}\overline{a})(\boldsymbol{\alpha}_0'\mathbf{X})\} = (c_{\mathbf{v}} + b_{\mathbf{v}}\overline{c}) + (a_{\mathbf{v}} + b_{\mathbf{v}}\overline{a})(\boldsymbol{\alpha}_0'\boldsymbol{\mu}_q),\end{aligned} \tag{B.5}$$

where, for the second equality in obtaining (B.5), we have used the fact that owing to (3.5), S is completely determined by the conditioning variables $\{\boldsymbol{\alpha}_0'\mathbf{X}, \epsilon^*\}$, so that the term $\mathbb{E}_q(. \mid \boldsymbol{\beta}_0'\mathbf{X}, \boldsymbol{\alpha}_0'\mathbf{X}, \epsilon, \epsilon^*)$ inside can be replaced by $\mathbb{E}(. \mid \boldsymbol{\beta}_0'\mathbf{X}, \boldsymbol{\alpha}_0'\mathbf{X}, \epsilon, \epsilon^*)$.

Next, note that that expected squared loss function $\mathbb{L}_q^*(\mathbf{v})$ satisfies: $\forall\ \mathbf{v} \in \mathbb{R}^p$,

$$\begin{aligned}\mathbb{L}_q^*(\mathbf{v}) &\equiv \mathbb{E}_q[\{Y_q^* - p_q^* - \mathbf{v}'(\mathbf{X} - \boldsymbol{\mu}_q)\}^2] \\ &= \mathbb{E}_q(\mathbb{E}_q[\{Y_q^* - p_q^* - \mathbf{v}'(\mathbf{X} - \boldsymbol{\mu}_q)\}^2 \mid \boldsymbol{\alpha}_0'\mathbf{X}, \epsilon^*]) \\ &= \mathbb{E}_q(\mathbb{E}[\{Y_q^* - p_q^* - \mathbf{v}'(\mathbf{X} - \boldsymbol{\mu}_q)\}^2 \mid \boldsymbol{\alpha}_0'\mathbf{X}, \epsilon^*]) \\ &\geq \mathbb{E}_q\left\{\left(\mathbb{E}[\{Y_q^* - p_q^* - \mathbf{v}'(\mathbf{X} - \boldsymbol{\mu}_q)\} \mid \boldsymbol{\alpha}_0'\mathbf{X}, \epsilon^*]\right)^2\right\} \\ &= \mathbb{E}_q\left[\left\{\mathbb{E}(Y_q^* \mid \boldsymbol{\alpha}_0'\mathbf{X}, \epsilon^*) - p_q^* - \mathbb{E}(\mathbf{v}'\mathbf{X} \mid \boldsymbol{\alpha}_0'\mathbf{X}, \epsilon^*) + \mathbf{v}'\boldsymbol{\mu}_q\right\}^2\right] \\ &= \mathbb{E}_q\left[\left\{Y_q^* - p_q^* - (a_{\mathbf{v}} + b_{\mathbf{v}}\overline{a})(\boldsymbol{\alpha}_0'\mathbf{X}) + (a_{\mathbf{v}} + b_{\mathbf{v}}\overline{a})(\boldsymbol{\alpha}_0'\boldsymbol{\mu}_q)\right\}^2\right] \\ &= \mathbb{E}_q\left[\left\{Y_q^* - p_q^* - (a_{\mathbf{v}} + b_{\mathbf{v}}\overline{a})\boldsymbol{\alpha}_0'(\mathbf{X} - \boldsymbol{\mu}_q)\right\}^2\right] \equiv \mathbb{L}_q^*\{(a_{\mathbf{v}} + b_{\mathbf{v}}\overline{a})\boldsymbol{\alpha}_0\}.\end{aligned} \tag{B.6}$$

The second equality in obtaining (B.6) follows from arguments similar to those mentioned earlier while obtaining (B.5). The subsequent inequality follows from (conditional) Jensen's

inequality, while in the penultimate step we have used (B.4)-(B.5), as well as the fact that owing to (3.5) and the very definition of Y_q^*, Y_q^* is completely determined (hence constant) by the conditioning variables $\{\boldsymbol{\alpha}_0'\mathbf{X}, \epsilon^*\}$. Thus, (B.6) now shows that the value of $\mathbb{L}_q^*(.)$ at every $\mathbf{v} \in \mathbb{R}^p$ is bounded below by its value at a corresponding point of the form $(a_{\mathbf{v}} + b_{\mathbf{v}}\overline{a})\boldsymbol{\alpha}_0 \in \mathbb{R}^p$. In particular, this also applies to $\mathbf{v} = \overline{\boldsymbol{\alpha}}_q$, which however is the unique minimizer of $\mathbb{L}_q^*(\mathbf{v})$ over $\mathbf{v} \in \mathbb{R}^p$. Hence, $\overline{\boldsymbol{\alpha}}_q$ must be of the form $(a_{\mathbf{v}} + b_{\mathbf{v}}\overline{a})\boldsymbol{\alpha}_0$ with $\mathbf{v} = \overline{\boldsymbol{\alpha}}_q$. We have therefore established (3.11), as required. The proof of theorem 3.1 is now complete. ∎

B.3 Proof of Theorem 3.2

The proof of this result, at least the initial part of it, relies substantially on a useful result from Negahban et al. (2012). We will therefore try to adopt some of their basic notations and terminology at the beginning of this proof in order to facilitate the use of that result.

Let $\mathcal{R}(\mathbf{u}) = \|\mathbf{u}\|_1 \; \forall \; \mathbf{u} \in \mathbb{R}^p$, and let $\mathcal{R}^*(\mathbf{u}) \equiv \sup_{\mathbf{v} \in \mathbb{R}^p \setminus \{\mathbf{0}\}} \{\mathbf{u}'\mathbf{v}/\mathcal{R}(\mathbf{v})\} \; \forall \; \mathbf{u} \in \mathbb{R}^p$ denote the 'dual norm' for $\mathcal{R}(.)$. Further, for any subspace $\mathcal{M} \subseteq \mathbb{R}^p$, let $\Psi(\mathcal{M}) \equiv \sup_{\mathbf{v} \in \mathcal{M} \setminus \{\mathbf{0}\}} \{\mathcal{R}(\mathbf{u})/\|\mathbf{u}\|_2\}$ denote the 'subspace compatibility constant' for $\mathcal{M}$ w.r.t. the norm $\mathcal{R}(.)$. Then, with $\mathcal{J}, \mathcal{M}_{\mathcal{J}}$ and $\mathcal{M}_{\mathcal{J}}^{\perp}$ as defined at the beginning of section 3.3.1, it is not difficult to show that: (i) $\mathcal{R}(.)$ is *decomposable* w.r.t. the orthogonal subspace pair $(\mathcal{M}_{\mathcal{J}}, \mathcal{M}_{\mathcal{J}}^{\perp})$ for any $\mathcal{J} \subseteq \{1, \ldots, p\}$, in the sense that $\mathcal{R}(\mathbf{u} + \mathbf{v}) = \mathcal{R}(\mathbf{u}) + \mathcal{R}(\mathbf{v}) \; \forall \; \mathbf{u} \in \mathcal{M}_{\mathcal{J}}, \mathbf{v} \in \mathcal{M}_{\mathcal{J}}^{\perp}$; (ii) $\mathcal{R}^*(\mathbf{u}) = \|\mathbf{u}\|_{\infty} \; \forall \; \mathbf{u} \in \mathbb{R}^p$; and (iii) with $\mathcal{J} = \mathcal{A}(\mathbf{v})$ for any $\mathbf{v} \in \mathbb{R}^p$, $\Psi^2(\mathcal{M}_{\mathcal{J}}) = s_{\mathbf{v}}$. We refer the interested reader to Negahban et al. (2012) for further discussions and/or proofs of these facts.

Then, owing to the decomposability of $\mathcal{R}(.)$ over $(\mathcal{M}_{\mathcal{J}}, \mathcal{M}_{\mathcal{J}}^{\perp})$ with $\mathcal{J}$ chosen to be $\mathcal{A}(\boldsymbol{\beta}_0)$, and under our restricted strong convexity assumption 3.3 regarding $\mathcal{L}_{n_q}(\mathcal{Z}_{n_q}^*; \boldsymbol{\beta})$ at $\boldsymbol{\beta} = \overline{\boldsymbol{\beta}}_q$, we have, using Theorem 1 of Negahban et al. (2012), that: for any given $\mathcal{Z}_{n_q}^*$ and $\lambda \geq 4\|\mathbb{T}_{n_q}\|_{\infty}$,

$$\left\|\widehat{\boldsymbol{\beta}}_{n_q}(\lambda; \mathcal{Z}_{n_q}^*) - \overline{\boldsymbol{\beta}}_q\right\|_2^2 \;\leq\; 9 s_{\boldsymbol{\beta}_0} \frac{\lambda^2}{\kappa_q^2} + 4\,|a_q|\,\frac{\lambda}{\kappa_q}\left\|\Pi_{\boldsymbol{\beta}_0}^c(\boldsymbol{\alpha}_0)\right\|_1, \tag{B.7}$$

where, while applying the result from Negahban et al. (2012), we have chosen the parameter $\boldsymbol{\theta}^*$, in their notation, as $\boldsymbol{\theta}^* = \overline{\boldsymbol{\beta}}_q$, and also used $2\mathcal{R}^*[\boldsymbol{\nabla}\{\mathcal{L}_{n_q}(\mathbb{Z}^*_{n_q}; \overline{\boldsymbol{\beta}}_q)\}] \equiv 4\|\mathbb{T}_{n_q}\|_\infty$, and $P^{\perp}_{\mathcal{A}(\boldsymbol{\beta}_0)}(\overline{\boldsymbol{\beta}}_q) = \Pi_{\mathcal{A}^c(\boldsymbol{\beta}_0)}(\overline{\boldsymbol{\beta}}_q) \equiv \Pi^c_{\boldsymbol{\beta}_0}(\overline{\boldsymbol{\beta}}_q) = \Pi^c_{\boldsymbol{\beta}_0}(a_q\boldsymbol{\alpha}_0)$, so that $\mathcal{R}\{P^{\perp}_{\mathcal{A}(\boldsymbol{\beta}_0)}(\overline{\boldsymbol{\beta}}_q)\} = |a_q|\|\Pi^c_{\boldsymbol{\beta}_0}(\boldsymbol{\alpha}_0)\|_1$.

It needs to be mentioned that the result from Negahban et al. (2012) used here is quite a powerful one, since it provides highly flexible and general bounds for penalized M-estimators based on loss functions satisfying some restricted strong convexity, like the one we assume in 3.3, and regularizers based on norms that are 'decomposable' over orthogonal subspace pairs, as shown to hold for the L_1 norm for subspace pairs like: $(\mathcal{M}_{\mathcal{J}}, \mathcal{M}^{\perp}_{\mathcal{J}})$ for any $\mathcal{J}$. The bounds hold for any such subspace pair which, in our case, we choose to be $\{\mathcal{M}_{\mathcal{A}(\boldsymbol{\beta}_0)}, \mathcal{M}^{\perp}_{\mathcal{A}(\boldsymbol{\beta}_0)}\}$. More importantly, the result provides deviation bounds of the estimator w.r.t. *any* point that can be reasonably viewed as a possible 'target', and *not* necessarily the exact parameter that minimizes the expected loss. Only the lower bound for λ needs to be appropriately defined for each such 'target'. Of course, the deviation bound depends directly on λ, and is only useful if the (random) lower bound, defined by this 'target', for λ can be bounded above w.h.p. by a sequence converging fast enough to 0. In our case, owing to theorem 3.1, $\overline{\boldsymbol{\alpha}}_q$, the minimizer of $\mathbb{L}^*_q(.)$, is really the 'official' target parameter. However, even with $\overline{\boldsymbol{\beta}}_q$ as the 'target', the corresponding deviation bound for $\widehat{\boldsymbol{\beta}}_{n_q}(\lambda)$ may still satisfy reasonable rates, since the lower bound $4\|\mathbb{T}_{n_q}\|_\infty$ for λ, defined through $\overline{\boldsymbol{\beta}}_q$, may still have a fast enough convergence rate w.h.p. owing to the definition of $\overline{\boldsymbol{\beta}}_q$ and the surrogacy assumption. In fact, this is what we precisely show in theorem 3.3. Nevertheless, the entire point of this digression was to provide some helpful perspectives regarding the nuances underlying our result and its proof, and also elaborate to some extent on the general usefulness of the tools used here.

Coming back to the proof, we next note that under assumption 3.4 conditions 1-2, with the λ chosen as above in (B.7) being further assumed to be $(\boldsymbol{\beta}_0, \boldsymbol{\alpha}_0, q)$-admissible, $\exists$ some realization $z^*_{n_q}$ (not necessarily the observed one) of $\mathcal{Z}^*_{n_q}$ such that the corresponding estimator $\widehat{\boldsymbol{\beta}}_{n_q}(\lambda; z^*_{n_q})$ based on $z^*_{n_q}$ and the given choice of λ, satisfies the property: $\widehat{\boldsymbol{\beta}}_{n_q[j]}(z^*_{n_q}; \lambda) = 0$

for some $j \in \mathcal{A}^c(\boldsymbol{\beta}_0) \cap \mathcal{A}(\boldsymbol{\alpha}_0)$. Noting that the bound in (B.7) is deterministic and applies to any realization of $\mathcal{Z}^*_{n_q}$, including $z^*_{n_q}$ in particular, we then have:

$$
\begin{aligned}
&|a_q|^2 \, |\boldsymbol{\alpha}_{0[j]}|^2 \;\equiv\; \left\{\widehat{\boldsymbol{\beta}}_{n_q[j]}(\lambda; z^*_{n_q}) - \overline{\boldsymbol{\beta}}_{q[j]}\right\}^2 \;\leq\; \left\|\widehat{\boldsymbol{\beta}}_{n_q}(\lambda; z^*_{n_q}) - \overline{\boldsymbol{\beta}}_q\right\|_2^2 \\
&\leq 9 s_{\boldsymbol{\beta}_0} \frac{\lambda^2}{\kappa_q^2} + 4\,|a_q| \, \frac{\lambda}{\kappa_q} \left\|\Pi^c_{\boldsymbol{\beta}_0}(\boldsymbol{\alpha}_0)\right\|_1 \qquad \text{(using B.7)}, \qquad \text{and therefore,} \\
&|a_q| \;\leq\; \frac{\lambda}{\kappa_q \left|\boldsymbol{\alpha}_{0[j]}\right|^2} \left[2 \left\|\Pi^c_{\boldsymbol{\beta}_0}(\boldsymbol{\alpha}_0)\right\|_1 + \left\{4 \left\|\Pi^c_{\boldsymbol{\beta}_0}(\boldsymbol{\alpha}_0)\right\|_1^2 + 9 s_{\boldsymbol{\beta}_0} \left|\boldsymbol{\alpha}_{0[j]}\right|^2\right\}^{\frac{1}{2}}\right] \\
&\leq\; \frac{\lambda}{\kappa_q C_{\min}(\boldsymbol{\alpha}_0, \boldsymbol{\beta}_0)^2} \left\{4 \left\|\Pi^c_{\boldsymbol{\beta}_0}(\boldsymbol{\alpha}_0)\right\|_1 + 3 s_{\boldsymbol{\beta}_0}^{\frac{1}{2}} C_{\max}(\boldsymbol{\alpha}_0, \boldsymbol{\beta}_0)\right\} \;\equiv\; \frac{\lambda}{\kappa_q} \overline{d}(\boldsymbol{\alpha}_0, \boldsymbol{\beta}_0), \quad \text{(B.8)}
\end{aligned}
$$

where $\overline{d}(\boldsymbol{\alpha}_0, \boldsymbol{\beta}_0)$ is as defined in theorem 3.2, and the preliminary bound on $|a_q|$ in the second last step follows from noting that the previous step leads to a quadratic inequality in $|a_q|$ and therefore, some straightforward algebra involving standard theory of quadratic inequalities leads to this bound. Finally, to obtain our desired result, we now note that:

$$
\begin{aligned}
&\left\|\widehat{\boldsymbol{\beta}}_{n_q}(\lambda; \mathcal{Z}^*_{n_q}) - b_q \beta_0\right\|_2 \;=\; \left\|\widehat{\boldsymbol{\beta}}_{n_q}(\lambda; \mathcal{Z}^*_{n_q}) - \overline{\boldsymbol{\beta}}_q + a_q \boldsymbol{\alpha}_0\right\|_2 \\
&\leq\; \left\|\widehat{\boldsymbol{\beta}}_{n_q}(\lambda; \mathcal{Z}^*_{n_q}) - \overline{\boldsymbol{\beta}}_q\right\|_2 + |a_q| \, \|\boldsymbol{\alpha}_0\|_2 \\
&\leq\; \left\{9 s_{\boldsymbol{\beta}_0} \frac{\lambda^2}{\kappa_q^2} + 4\,|a_q| \, \frac{\lambda}{\kappa_q} \left\|\Pi^c_{\boldsymbol{\beta}_0}(\boldsymbol{\alpha}_0)\right\|_1\right\}^{\frac{1}{2}} + |a_q| \, \|\boldsymbol{\alpha}_0\|_2 \\
&\leq\; \frac{\lambda}{\kappa_q} \left\{9 s_{\boldsymbol{\beta}_0} + 4 \left\|\Pi^c_{\boldsymbol{\beta}_0}(\boldsymbol{\alpha}_0)\right\|_1 \overline{d}(\boldsymbol{\alpha}_0, \boldsymbol{\beta}_0)\right\}^{\frac{1}{2}} + \frac{\lambda}{\kappa_q} \overline{d}(\boldsymbol{\alpha}_0, \boldsymbol{\beta}_0) \, \|\boldsymbol{\alpha}_0\|_2 \\
&\equiv\; \frac{\lambda}{\kappa_q} \left[\left\{9 s_{\boldsymbol{\beta}_0} + d_1(\boldsymbol{\alpha}_0, \boldsymbol{\beta}_0)\right\}^{\frac{1}{2}} + d_2(\boldsymbol{\alpha}_0, \boldsymbol{\beta}_0)\right], \qquad \text{(B.9)}
\end{aligned}
$$

where the bounds follow due to (B.7)-(B.8). The proof of theorem 3.2 is now complete. ∎

B.4 Proof of Theorem 3.3

With $\mathbf{X} \sim \mathbb{SG}_q(\sigma_q^2)$, we first note that $(\mathbf{X}_i - \overline{\mathbf{X}}_{n_q}) \sim \mathbb{SG}_q(\overline{\sigma}_q^2) \; \forall \; 1 \leq i \leq n_q$, owing to lemma B.1 (ii) and (v), where $\overline{\sigma}_q^2 = \sigma_q^2(1 - n_q^{-1}) \leq \sigma_q^2$. Further, let us define: $\mathbb{Z}^*_q = |Y - Y^*_q| \in \{0, 1\}$, and define $\mathbb{Z}^*_{q,i} \; \forall \; 1 \leq i \leq n_q$ accordingly. Then, $(\mathbb{Z}^*_q \mid S \in \mathcal{I}_q)$ is a binary variable with

$\mathbb{P}_q(\mathbb{Z}_q^* = 1) = \pi_q$. Hence, using lemma B.2, $(\mathbb{Z}_q^* - \pi_q) \sim \mathbb{SG}_q(\widetilde{\pi}_q^2)$, and further, using lemma B.1 (ii), $n^{-1}(\sum_{i=1}^{n_q} \mathbb{Z}_{q,i}^*) \sim \mathbb{SG}_q(\widetilde{\pi}_q^2/n_q)$. Using lemma B.1 (v) and (i), we now have: $\forall\ \epsilon_1, \epsilon_2 > 0$,

$$\mathbb{P}_q\left(\max_{1\le i\le n_q} \left\|\mathbf{X}_i - \overline{\mathbf{X}}_{n_q}\right\|_\infty > \epsilon_1\right) \ \le\ 2\exp\left\{-\frac{\epsilon_1^2}{2\sigma_q^2} + \log\left(n_q p\right)\right\}, \text{ and} \tag{B.10}$$

$$\mathbb{P}_q\left\{\frac{1}{n_q}\sum_{i=1}^{n_q}\mathbb{Z}_{q,i}^* > (\pi_q + \epsilon_2)\right\} \ \le\ \exp\left(-\frac{n_q\epsilon_2^2}{2\widetilde{\pi}_q^2}\right). \tag{B.11}$$

Using (B.10)-(B.11), and noting the definition of $\mathbb{T}_{n_q}^{(1)}$ in (3.18), we then have: $\forall\ \epsilon_1, \epsilon_2 > 0$,

$$\begin{aligned}
&\mathbb{P}_q\left\{\left\|\mathbb{T}_{n_q}^{(1)}\right\|_\infty > \epsilon_1(\pi_q + \epsilon_2)\right\} \ \equiv\ \mathbb{P}_q\left\{\left\|\frac{1}{n_q}\sum_{i=1}^{n_q}(\mathbf{X}_i - \overline{\mathbf{X}}_{n_q})(Y_{q,i}^* - Y_i)\right\|_\infty > \epsilon_1(\pi_q + \epsilon_2)\right\} \\
&\le\ \mathbb{P}_q\left\{\left(\max_{1\le i\le n_q}\left\|\mathbf{X}_i - \overline{\mathbf{X}}_{n_q}\right\|_\infty\right)\left(\frac{1}{n_q}\sum_{i=1}^{n_q}\mathbb{Z}_{q,i}^*\right) > \epsilon_1(\pi_q + \epsilon_2)\right\} \\
&\le\ \mathbb{P}_q\left(\max_{1\le i\le n_q}\left\|\mathbf{X}_i - \overline{\mathbf{X}}_{n_q}\right\|_\infty > \epsilon_1\right) + \mathbb{P}_q\left\{\frac{1}{n_q}\sum_{i=1}^{n_q}\mathbb{Z}_{q,i}^* > (\pi_q + \epsilon_2)\right\} \\
&\le\ 2\exp\left\{-\frac{\epsilon_1^2}{2\sigma_q^2} + \log\left(n_q p\right)\right\} + \exp\left(-\frac{n_q\epsilon_2^2}{2\widetilde{\pi}_q^2}\right).
\end{aligned} \tag{B.12}$$

(B.12) therefore establishes the first of the three bounds in (3.24). To obtain the other two, Let us first define: $\widetilde{\mathbf{X}}_q = (\mathbf{X} - \boldsymbol{\mu}_q)$, $\widetilde{Y}_q = (Y - p_q)$ and $\widetilde{\mathbb{Z}}_q = (\widetilde{Y}_q - \overline{\boldsymbol{\beta}}_q'\widetilde{\mathbf{X}}_q)$. Then, we first note that $\widetilde{\mathbf{X}}_q \sim \mathbb{SG}_q(\sigma_q^2)$ by assumption, $\widetilde{Y}_q \sim \mathbb{SG}_q(\widetilde{p}_q^2)$ owing to lemma B.2, and $\widetilde{\mathbb{Z}}_q \sim \mathbb{SG}_q(\gamma_q^2)$ owing to lemma B.1 (ii) and (v), where $\gamma_q^2 = (\widetilde{p}_q + \sigma_q\|\overline{\boldsymbol{\beta}}_q\|_2)^2$ is as defined in theorem 3.3. Hence, applying lemma B.1 (iii) to $\widetilde{\mathbb{Z}}_q$ and $\widetilde{\mathbf{X}}_q$, we then have: $\forall\ 1 \le j \le p$, and any integer $m \ge 2$,

$$\begin{aligned}
\mathbb{E}_q\left(|\widetilde{\mathbf{X}}_{q[j]}\widetilde{\mathbb{Z}}_q|^m\right) \ &\le\ \left\{\mathbb{E}_q\left(|\widetilde{\mathbf{X}}_{q[j]}|\right)^{2m}\right\}^{\frac{1}{2}}\left\{\mathbb{E}_q\left(|\widetilde{\mathbb{Z}}_q|^{2m}\right)\right\}^{\frac{1}{2}} \\
&\le\ \left\{2\left(\sqrt{2}\sigma_q\right)^{2m}\Gamma(m+1)\right\}^{\frac{1}{2}}\left\{2\left(\sqrt{2}\gamma_q\right)^{2m}\Gamma(m+1)\right\}^{\frac{1}{2}} \\
&= 2\{\Gamma(m+1)\}(2\sigma_q\gamma_q)^m \ =\ \frac{m!}{2}(2\sigma_q\gamma_q)^{m-2}(4\sigma_q\gamma_q)^2,
\end{aligned}$$

where the first inequality follows from Holder's inequality, and the rest are due to lemma B.1 (iii) applied to $\widetilde{\mathbb{Z}}_q$ and $\widetilde{\mathbf{X}}_{q[j]}$ for each j. Next, note that owing to the very definition of $\overline{\boldsymbol{\beta}}_q$, $\mathbb{E}_q(\widetilde{\mathbf{X}}_q\widetilde{\mathbb{Z}}_q) = \mathbf{0}$ and further, the above bound ensures that the random variable $\widetilde{\mathbf{X}}_{q[j]}\widetilde{\mathbb{Z}}_q$, for each j satisfies the moment conditions required in lemma B.3 with $\sigma \equiv 4\sigma_q\gamma_q$ and $K \equiv 2\sigma_q\gamma_q$. Hence applying lemma B.3, and noting the definition of $\mathbb{T}^{(2)}_{n_q,1}$ in (3.19), we have: for any $\epsilon_3 > 0$,

$$\begin{aligned}
\mathbb{P}_q\left\{\left\|\mathbb{T}^{(2)}_{n_q,1}\right\|_\infty > 2\sigma_q\gamma_q(2\sqrt{2}\epsilon_3+\epsilon_3^2)\right\} &\equiv \mathbb{P}_q\left\{\left\|\frac{1}{n_q}\sum_{i=1}^{n_q}\widetilde{\mathbf{X}}_{q,i}\widetilde{\mathbb{Z}}_{q,i}\right\|_\infty > 2\sigma_q\gamma_q(2\sqrt{2}\epsilon_3+\epsilon_3^2)\right\} \\
&\leq \sum_{j=1}^{p}\mathbb{P}_q\left\{\frac{1}{n_q}\left|\sum_{i=1}^{n_q}\widetilde{\mathbf{X}}_{q,i[j]}\widetilde{\mathbb{Z}}_{q,i[j]}\right| > 2\sigma_q\gamma_q(2\sqrt{2}\epsilon_3+\epsilon_3^2)\right\} \\
&\leq 2p\exp\left(-n_q\epsilon_3^2\right) \equiv 2\exp\left(-n_q\epsilon_3^2+\log p\right), \qquad \text{(B.13)}
\end{aligned}$$

where the first inequality follows from a straightforward application of the union bound, and the next one follows from the use of lemma B.3. (B.13) therefore establishes the second bound in (3.24). To establish the third and final bound in (3.24), we first note that using lemma B.2 (ii), $(\overline{\mathbf{X}}_{n_q} - \boldsymbol{\mu}_q) \sim \mathbb{SG}_q(\sigma_q^2/n_q)$, and $n_q^{-1}\sum_{i=1}^{n_q}\widetilde{\mathbb{Z}}_{q,i} \sim \mathbb{SG}_q(\gamma_q^2/n_q)$. Using lemma B.1 (v) and (i), and noting that $\mathbb{E}_q(\widetilde{\mathbb{Z}}_q) = 0$, we now have: $\forall\ \epsilon_4, \epsilon_5 > 0$,

$$\begin{aligned}
\mathbb{P}_q\left(\left\|\overline{\mathbf{X}}_{n_q} - \boldsymbol{\mu}_q\right\|_\infty > \epsilon_4\right) &\leq 2\exp\left\{-\frac{n_q\epsilon_4^2}{2\sigma_q^2}+\log p\right\}, \text{ and} \qquad \text{(B.14)} \\
\mathbb{P}_q\left(\frac{1}{n_q}\left|\sum_{i=1}^{n_q}\widetilde{\mathbb{Z}}_{q,i}\right| > \epsilon_5\right) &\leq \exp\left(-\frac{n_q\epsilon_5^2}{2\gamma_q^2}\right). \qquad \text{(B.15)}
\end{aligned}$$

Using (B.14)-(B.15), and noting the definition of $\mathbb{T}^{(2)}_{n_q,2}$ in (3.20), we then have: $\forall\ \epsilon_4, \epsilon_5 > 0$,

$$\begin{aligned}
\mathbb{P}_q\left(\left\|\mathbb{T}^{(2)}_{n_q,2}\right\|_\infty > \epsilon_4\epsilon_5\right) &\equiv \mathbb{P}_q\left\{\left\|\left(\overline{\mathbf{X}}_{n_q} - \boldsymbol{\mu}_q\right)\left(\frac{1}{n_q}\sum_{i=1}^{n_q}\widetilde{\mathbb{Z}}_{q,i}\right)\right\|_\infty > \epsilon_4\epsilon_5\right\} \\
&\leq \mathbb{P}_q\left\{\left\|\overline{\mathbf{X}}_{n_q} - \boldsymbol{\mu}_q\right\|_\infty\left(\frac{1}{n_q}\left|\sum_{i=1}^{n_q}\widetilde{\mathbb{Z}}_{q,i}\right|\right) > \epsilon_4\epsilon_5\right\}
\end{aligned}$$

$$\leq \mathbb{P}_q\left(\left\|\overline{\mathbf{X}}_{n_q} - \boldsymbol{\mu}_q\right\|_\infty > \epsilon_4\right) + \mathbb{P}_q\left\{\frac{1}{n_q}\left|\sum_{i=1}^{n_q}\widetilde{\mathbb{Z}}_{q,i}\right| > \epsilon_5\right\}$$
$$\leq 2\exp\left\{-\frac{n_q\epsilon_4^2}{2\sigma_q^2} + \log p\right\} + 2\exp\left(-\frac{n_q\epsilon_5^2}{2\gamma_q^2}\right). \quad \text{(B.16)}$$

(B.16) therefore establishes the third and final bound in (3.24). Lastly, the claim (3.25) in theorem 3.3 follows from noting the representations (3.17)-(3.20) of $\mathbb{T}_{n_q}$ in terms of $\mathbb{T}^{(1)}_{n_q}$, $\mathbb{T}^{(2)}_{n_q,1}$ and $\mathbb{T}^{(2)}_{n_q,2}$, and straightforward use of (B.12), (B.13) and (B.16) through appropriate choices of $\{\epsilon_1, \ldots, \epsilon_5\}$ in terms of universal constants $\{c_1, \ldots c_6\}$ as follows: $\epsilon_1 = \{2\log(n_q^{c_1}p^{c_2})\}^{\frac{1}{2}}\sigma_q$, for any $c_1, c_2 > 0$ such that $\max(c_1, c_2) > 1$; $\epsilon_2 = \{c_3(1-2\pi_q)/n_q\}^{\frac{1}{2}}$ for any $c_3 > 0$, and also noting the definition of $\widetilde{\pi}_q$ when $\pi_q < 1/2$ (as assumed for this part); $\epsilon_3 = \{c_4(\log p)/n_q\}^{\frac{1}{2}}$ for any $c_4 > 1$; and lastly, $\epsilon_4 = \{2c_5(\log p)/n_q\}^{\frac{1}{2}}\sigma_q$ and $\epsilon_5 = \{2c_6(\log p)/n_q\}^{\frac{1}{2}}\gamma_q$, for any $c_5 > 1$ and any $c_6 > 0$ respectively. The proof of theorem 3.3 is now complete. ∎

B.5 Proof of Theorem 3.4

First of all, the results for $\{\mathbb{E}(S|S \leq \delta_q), \mathbb{E}(S^2|S \leq \delta_q)\}$ and $\{\mathbb{E}(S|S \geq \overline{\delta}_q), \mathbb{E}(S^2|S \geq \overline{\delta}_q)\}$, are straightforward implications of lemma B.5 (i) with the choices of a, b as: $\{a = -\infty, b = \delta_q\}$ and $\{a = \overline{\delta}_q, b = \infty\}$ respectively. Further, the results for $\mathbb{E}_q(S)$ and $\mathrm{Var}_q(S)$ follow from noting that: $\mathbb{E}_q(S) = 1/2\{\mathbb{E}(S\,|\,S \leq \delta_q) + \mathbb{E}(S\,|\,S \geq \overline{\delta}_q)\}$, and $1/2\{\mathbb{E}(S^2\,|\,S \leq \delta_q) + \mathbb{E}(S^2\,|\,S \geq \overline{\delta}_q)\} = \mathbb{E}_q(S^2) \equiv \mathrm{Var}_q(S)$ since $\mathbb{E}_q(S) = 0$. Next, for the corresponding results regarding $\mathbf{X}$, we first note that under the assumed set-up, $\mathbf{X} \mid S$ follows a linear model given by: $\mathbf{X} = \boldsymbol{\gamma}_0 S + \boldsymbol{\epsilon}$ with $\boldsymbol{\epsilon} \sim \mathcal{N}_p(\mathbf{0}, \Gamma)$ and $\boldsymbol{\epsilon} \perp\!\!\!\perp S$, where $\boldsymbol{\gamma}_0$ and Γ are as defined therein. Using this relation and the results already proved, the results for $\{\mathbb{E}(\mathbf{X}\,|\,S \leq \delta_q), \mathbb{E}(\mathbf{X}\mathbf{X}'\,|\,S \leq \delta_q)\}$ and $\{\mathbb{E}(\mathbf{X}\,|\,S \geq \overline{\delta}_q), \mathbb{E}(\mathbf{X}\mathbf{X}'\,|\,S \geq \overline{\delta}_q)\}$ now follow immediately. Further, the results for $\mathbb{E}_q(\mathbf{X})$ and $\mathrm{Var}_q(\mathbf{X})$ follow from noting that: $\mathbb{E}_q(\mathbf{X}) = 1/2\{\mathbb{E}(\mathbf{X} \mid S \leq \delta_q) + \mathbb{E}(\mathbf{X} \mid S \geq \overline{\delta}_q)\}$, and $1/2\{\mathbb{E}(\mathbf{X}\mathbf{X}' \mid S \leq \delta_q) + \mathbb{E}(\mathbf{X}\mathbf{X}' \mid S \geq \overline{\delta}_q)\} = \mathbb{E}_q(\mathbf{X}\mathbf{X}') \equiv \mathrm{Var}_q(\mathbf{X})$ since $\mathbb{E}_q(\mathbf{X}) = 0$. This completes the proof of all the results mentioned in (i) and (ii) in theorem 3.4. ∎

Next, we note that $\forall\, t \in \mathbb{R}$, $\mathrm{MGF}_{S,q}(t) \equiv \mathbb{E}_q(e^{tS}) = 1/2\{\mathbb{E}(e^{tS}\,|\,S \leq \delta_q) + \mathbb{E}(e^{tS}\,|\,S \geq \overline{\delta}_q)\}$,

and the result in (iii) for $\text{MGF}_{S,q}(t)$ now follows directly from lemma B.5 (i) using the choices of a, b as: $\{a = -\infty, b = \delta_q\}$ and $\{a = \overline{\delta}_q, b = \infty\}$ respectively. For $\text{MGF}_{\mathbf{X},q}(\mathbf{t})$, we note that $\forall\ \mathbf{t} \in \mathbb{R}^p$, $\text{MGF}_{\mathbf{X},q}(\mathbf{t}) \equiv \mathbb{E}_q(e^{\mathbf{t}'\mathbf{X}}) = \mathbb{E}_q[\exp\{(\mathbf{t}'\boldsymbol{\gamma}_0)S + \mathbf{t}'\boldsymbol{\epsilon}\}] = \{\mathbb{E}(e^{\mathbf{t}'\boldsymbol{\epsilon}})\}\text{MGF}_{S,q}(\mathbf{t}'\boldsymbol{\gamma}_0)$ where, in the last step, we use $\boldsymbol{\epsilon} \perp\!\!\!\perp S$. The result now follows from using the result for $\text{MGF}_{S,q}(.)$, and the standard expression for the m.g.f. of $\boldsymbol{\epsilon} \sim \mathcal{N}_p(\mathbf{0}, \Gamma)$, as well as using the fact that $\Gamma = (\Sigma - \sigma_S^2\boldsymbol{\gamma}_0\boldsymbol{\gamma}_0')$. This completes the proof of all results in (iii). Further, all the bounds in (iv) for $\text{MGF}_{S,q}(.)$ are straightforward implications of lemma B.5 (ii), and so are the bounds for $\text{MGF}_{\mathbf{X},q}(.)$ in (iv), where we additionally use standard the inequalities: $\mathbf{t}'\Sigma\mathbf{t} \leq \lambda_{\max}(\Sigma)\|\mathbf{t}\|_2^2$ and $|\mathbf{t}'\boldsymbol{\gamma}_0|^2 \leq \|\mathbf{t}\|_2^2\|\boldsymbol{\gamma}_0\|_2^2\ \forall\ \mathbf{t} \in \mathbb{R}^p$. This therefore completes the proof of all the m.g.f. related results mentioned in (iii) and (iv) in theorem 3.4. ∎

Next, for the bounds on $\overline{\delta}_q$ in result (vi), the upper bound is a straightforward consequence of the second inequality provided in lemma B.4, and noting that: $q/2 = \overline{\Phi}(\overline{z}_q)$ and $\overline{\delta}_q = \sigma_S\overline{z}_q$. The lower bound follows from the lower bound given in the first inequality in lemma B.4. The restriction $q \geq 0.0002$ in the statement of the lower bound result in (vi) is needed to bound the quantity $\{(1 + \overline{z}_q^2)/\overline{z}_q\}$ that inevitably comes up while using the inequality from the lemma. In particular, this restriction implies that $\{(1 + \overline{z}_q^2)/\overline{z}_q\} \leq 5\sqrt{2/\pi}$ and ensures the final bound stated in the result. This completes the proof of result (vi) in theorem 3.4. ∎

Finally, to show the results in (v) regarding π_q^-, π_q^+ and π_q, we first note that:

$$\begin{aligned}
\pi_q^- \ \equiv\ \mathbb{P}(Y = 1 \mid S \leq \delta_q) \ &= \ \mathbb{E}\{\mathbb{P}(Y = 1 \mid \mathbf{X}, S) \mid S \leq \delta_q\} \\
&= \ \mathbb{E}\{\mathbb{P}(Y = 1 \mid \mathbf{X}) \mid S \leq \delta_q\} \ = \ \mathbb{E}\{\psi(\boldsymbol{\beta}_0'\mathbf{X}) \mid S \leq \delta_q\} \\
&\leq \ \mathbb{E}\left\{e^{\boldsymbol{\beta}_0'\mathbf{X}} \mid S \leq \delta_q\right\} \ = \ e^{\frac{1}{2}\eta_0^2}\frac{\Phi(-\overline{z}_q - \sigma_S\boldsymbol{\beta}_0'\boldsymbol{\gamma}_0)}{\Phi(-\overline{z}_q)}, \qquad \text{(B.17)} \\
\pi_q^+ \ \equiv\ \mathbb{P}(Y = 0 \mid S \geq \overline{\delta}_q) \ &= \ \mathbb{E}\{\mathbb{P}(Y = 0 \mid \mathbf{X}, S) \mid S \geq \overline{\delta}_q\} \\
&= \ \mathbb{E}\{\mathbb{P}(Y = 0 \mid \mathbf{X}) \mid S \geq \overline{\delta}_q\} \ = \ \mathbb{E}\{\psi(-\boldsymbol{\beta}_0'\mathbf{X}) \mid S \geq \overline{\delta}_q\} \\
&\leq \ \mathbb{E}\left\{e^{-\boldsymbol{\beta}_0'\mathbf{X}} \mid S \geq \overline{\delta}_q\right\} \ = \ \frac{\Phi(-\overline{z}_q - \sigma_S\boldsymbol{\beta}_0'\boldsymbol{\gamma}_0)}{\Phi(-\overline{z}_q)}e^{\eta_0^2/2}, \qquad \text{(B.18)}
\end{aligned}$$

and therefore, we have:

$$\pi_q \equiv \frac{1}{2}(\pi_q^- + \pi_q^+) \leq e^{\frac{1}{2}\eta_0^2}\frac{\Phi(-\overline{z}_q - \sigma_S \boldsymbol{\beta}_0'\boldsymbol{\gamma}_0)}{\Phi(-\overline{z}_q)}. \tag{B.19}$$

For both (B.17) and (B.18), the second steps use the fact that $(Y \perp\!\!\!\perp S) \mid \mathbf{X}$, and the final bounds follow, similar to the earlier proofs for the results in (iii), from straightforward uses of the results regarding m.g.f.s of truncated normal distributions given in lemma B.5, as well as use of the relationship between $\mathbf{X}$ and S given by: $\mathbf{X} = \boldsymbol{\gamma}_0 S + \boldsymbol{\epsilon}$ with $\boldsymbol{\epsilon} \sim \mathcal{N}_p(\mathbf{0}, \Gamma)$ and $\boldsymbol{\epsilon} \perp\!\!\!\perp S$, and noting the definitions of γ_0, Γ and η_0. (B.19) therefore establishes the first bound in result (v) of theorem 3.4. The subsequent bounds in result (v) now follow from (B.19) along with straightforward uses of the first inequality (both the upper and lower bounds) given in lemma B.4 noting that $\boldsymbol{\beta}_0'\boldsymbol{\gamma}_0 > 0$ by assumption, as well as using the fact that $\sigma_s \boldsymbol{\beta}_0'\boldsymbol{\gamma}_0 \equiv \widetilde{\rho}_0 \eta_0$. All the claims in result (v) are now established, and the proof of theorem 3.4 is complete. ∎

References

Alarcón-Segovia, D. (2005). Shared Autoimmunity: A Concept for Which the Time Has Come. *Autoimmunity* **38** 201–203.

Andrews, D. W. (1995). Nonparametric Kernel Estimation for Semiparametric Models. *Econometric Theory* **11** 560–586.

Belkin, M., Niyogi, P. and Sindhwani, V. (2006). Manifold Regularization: A Geometric Framework for Learning from Labeled and Unlabeled Examples. *The Journal of Machine Learning Research* **7** 2399–2434.

Bielinski, S. J., Chai, H. S., Pathak, J. et al. (2011). Mayo Genome Consortia: A Genotype-Phenotype Resource for Genome-Wide Association Studies with an Application to the Analysis of Circulating Bilirubin Levels. *Mayo Clinic Proceedings* **86** 606–614.

Buldygin, V. V. and Moskvichova, K. K. (2013). The Sub-Gaussian Norm of a Binary Random Variable. *Theory of Probability and Mathematical Statistics* **86** 33–49.

Buonaccorsi, J. P. (2010). *Measurement Error: Models, Methods, and Applications*. CRC Press.

Burkardt, J. (2014). The Truncated Normal Distribution. Tech. rep., Department of Scientific Computing, Florida State University, USA.

Carroll, R. J. (1998). Measurement Error in Epidemiologic Studies. *Encyclopedia of Biostatistics* .

Carroll, R. J., Ruppert, D., Stefanski, L. A. and Crainiceanu, C. M. (2006). *Measurement Error in Nonlinear Models: A Modern Perspective*. CRC Press.

Castelli, V. and Cover, T. M. (1995). The Exponential Value of Labeled Samples. *Pattern Recognition Letters* **16** 105–111.

Castelli, V. and Cover, T. M. (1996). The Relative Value of Labeled and Unlabeled Samples in Pattern Recognition with an Unknown Mixing Parameter. *IEEE Transactions on Information Theory* **42** 2102–2117.

Chakrabortty, A. and Cai, T. (2015). Efficient and Adaptive Linear Regression in Semi-Supervised Settings. *Preprint* .

Chapelle, O., Schölkopf, B. and Zien, A. (2006). *Semi-Supervised Learning*. MIT Press, Cambridge, MA, USA.

Chen, Y., Caramanis, C. and Mannor, S. (2013). Robust Sparse Regression under Adversarial Corruption. In *Proceedings of the 30th International Conference on Machine Learning (ICML-13)*.

Chen, Y., Wang, J. Z. and Krovetz, R. (2005). CLUE: Cluster-Based Retrieval of Images by Unsupervised Learning. *IEEE Transactions on Image Processing* **14** 1187–1201.

Chiani, M., Dardari, D. and Simon, M. K. (2003). New Exponential Bounds and Approximations for the Computation of Error Probability in Fading Channels. *IEEE Transactions on Wireless Communications* **2** 840–845.

Chung, S. A. and Criswell, L. A. (2007). PTPN22: Its Role in SLE and Autoimmunity. *Autoimmunity* **40** 582–590.

Cios, K. J., Swiniarski, R. W., Pedrycz, W. and Kurgan, L. A. (2007). Unsupervised Learning: Clustering. In *Data Mining*. Springer.

Cook, R. D. (1998). Principal Hessian Directions Revisited (with Discussion). *Journal of the American Statistical Association* **93** 84–100.

Cook, R. D. (2009). *Regression Graphics: Ideas for Studying Regressions through Graphics*, vol. 482. John Wiley & Sons.

Cook, R. D. and Lee, H. (1999). Dimension Reduction in Binary Response Regression. *Journal of the American Statistical Association* **94** 1187–1200.

Cook, R. D. and Weisberg, S. (1991). Discussion of "Sliced Inverse Regression" by K.-C. Li. *Journal of the American Statistical Association* **86** 328–332.

Cotsapas, C., Voight, B. F., Rossin, E. et al. (2011). Pervasive Sharing of Genetic Effects in Autoimmune Disease. *PLoS Genetics* **7** e1002254.

Cozman, F. G. and Cohen, I. (2001). Unlabeled Data Can Degrade Classification Performance of Generative Classifiers. Tech. Rep. HPL-2001-234, HP Laboratories, Palo Alto, CA, USA.

Cozman, F. G., Cohen, I. and Cirelo, M. C. (2003). Semi-Supervised Learning of Mixture Models. In *Proceedings of the Twentieth ICML*.

Culp, M. (2013). On the Semi-Supervised Joint Trained Elastic Net. *Journal of Computational and Graphical Statistics* **22** 300–318.

Denny, J. C., Ritchie, M. D., Basford, M. A. et al. (2010). PheWAS: Demonstrating the Feasibility of a Phenome-Wide Scan to Discover Gene–Disease Associations. *Bioinformatics* **26** 1205–1210.

Duan, N. and Li, K.-C. (1991). Slicing Regression: A Link-Free Regression Method. *The Annals of Statistics* **19** 505–530.

Düembgen, L. (2010). Bounding Standard Gaussian Tail Probabilities. *arXiv Preprint (arXiv:1012.2063)* .

Feng, J., Xu, H., Mannor, S. and Yan, S. (2014). Robust Logistic Regression and Classification. In *Advances in Neural Information Processing Systems*.

Gllavata, J., Ewerth, R. and Freisleben, B. (2004). Text Detection in Images Based on Unsupervised Classification of High-Frequency Wavelet Coefficients. In *Proceedings of the 17th International Conference on Pattern Recognition, 2004. ICPR 2004.*, vol. 1. IEEE.

Hall, P. and Li, K.-C. (1993). On Almost Linearity of Low Dimensional Projections from High Dimensional Data. *The Annals of Statistics* 867–889.

Hansen, B. E. (2008). Uniform Convergence Rates for Kernel Estimation with Dependent Data. *Econometric Theory* **24** 726–748.

Hastie, T. J., Tibshirani, R. J. and Friedman, J. H. (2008). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Second Edition, Springer Series in Statistics, Springer, Berlin.

Henegar, C., Clément, K. and Zucker, J.-D. (2006). Unsupervised Multiple-Instance Learning for Functional Profiling of Genomic Data. In *Machine Learning: ECML 2006*. Springer, 186–197.

Hofmann, T. (2001). Unsupervised Learning by Probabilistic Latent Semantic Analysis. *Machine Learning* **42** 177–196.

Horrace, W. C. (2004). On Ranking and Selection from Independent Truncated Normal Distributions. *Journal of Econometrics* **126** 335–354.

Horrace, W. C. (2005). Some Results on the Multivariate Truncated Normal Distribution. *Journal of Multivariate Analysis* **94** 209–221.

Hsing, T. and Carroll, R. J. (1992). An Asymptotic Theory for Sliced Inverse Regression. *The Annals of Statistics* **20** 1040–1061.

Jacques, L., Laska, J. N., Boufounos, P. T. and Baraniuk, R. G. (2013). Robust 1-bit Compressive Sensing via Binary Stable Embeddings of Sparse Vectors. *IEEE Transactions on Information Theory* **59** 2082–2102.

Johnson, N. L., Kotz, S. and Balakrishnan, N. (1994). *Continuous Univariate Distributions: Volume 1*. John Wiley & Sons, New York, USA.

Kawakita, M. and Kanamori, T. (2013). Semi-Supervised Learning with Density-Ratio Estimation. *Machine Learning* **91** 189–209.

Ko, Y. and Seo, J. (2000). Automatic Text Categorization by Unsupervised Learning. In *Proceedings of the 18th Conference on Computational Linguistics-Volume 1*. Association for Computational Linguistics.

Kohane, I. S. (2011). Using Electronic Health Records to Drive Discovery in Disease Genomics. *Nature Reviews Genetics* **12** 417–428.

Kohane, I. S., Churchill, S. E. and Murphy, S. N. (2012). A Translational Engine at the National Scale: Informatics for Integrating Biology and the Bedside. *Journal of the American Medical Informatics Association* **19** 181–185.

Lafferty, J. D. and Wasserman, L. (2007). Statistical Analysis of Semi-Supervised Regression. *Advances in Neural Information Processing Systems* **20** 801–808.

Laska, J. N., Davenport, M. and Baraniuk, R. G. (2009). Exact Signal Recovery from Sparsely Corrupted Measurements through the Pursuit of Justice. In *Conference Record of the Forty-Third Asilomar Conference on Signals, Systems and Computers, 2009*. IEEE.

Li, K.-C. (1991). Sliced Inverse Regression for Dimension Reduction. *Journal of the American Statistical Association* **86** 316–327.

Li, K.-C. (1992). On Principal Hessian Directions for Data Visualization and Dimension Reduction: Another Application of Stein's Lemma. *Journal of the American Statistical Association* **87** 1025–1039.

Li, K.-C. and Duan, N. (1989). Regression Analysis under Link Violation. *The Annals of Statistics* **17** 1009–1052.

Li, X. (2013). Compressed Sensing and Matrix Completion with Constant Proportion of Corruptions. *Constructive Approximation* **37** 73–99.

Liao, K. P., Cai, T., Gainer, V. et al. (2010). Electronic Medical Records for Discovery Research in Rheumatoid Arthritis. *Arthritis Care and Research* **62** 1120–1127.

Liao, K. P., Kurreeman, F., Li, G. et al. (2013). Associations of Autoantibodies, Autoimmune Risk Alleles, and Clinical Diagnoses from the Electronic Medical Records in Rheumatoid Arthritis Cases and Non–Rheumatoid Arthritis Controls. *Arthritis & Rheumatism* **65** 571–581.

Masry, E. (1996). Multivariate Local Polynomial Regression for Time Series: Uniform Strong Consistency and Rates. *Journal of Time Series Analysis* **17** 571–600.

Merkl, D. and Rauber, A. (2000). Document classification with Unsupervised Artificial Neural Networks. In *Soft Computing in Information Retrieval*. Springer, 102–121.

Murphy, S., Churchill, S., Bry, L. et al. (2009). Instrumenting the Health Care Enterprise for Discovery Research in the Genomic Era. *Genome Research* **19** 1675–1681.

Natarajan, N., Dhillon, I. S., Ravikumar, P. K. and Tewari, A. (2013). Learning with Noisy Labels. In *Advances in Neural Information Processing Systems*.

Negahban, S. N., Ravikumar, P., Wainwright, M. J. and Yu, B. (2012). A Unified Framework for High-Dimensional Analysis of M-Estimators with Decomposable Regularizers. *Statistical Science* **27** 538–557.

Newey, W. K. (1994). Kernel Estimator of Partial Means and a Generalized Variance Estimator. *Econometric Theory* **10** 1–21.

Newey, W. K., Hsieh, F. and Robins, J. (1998). Undersmoothing and Bias Corrected Functional Estimation. Tech. Rep. 98-17, Dept. of Economics, MIT, USA.

Newey, W. K. and McFadden, D. (1994). Large Sample Estimation and Hypothesis Testing. *Handbook of Econometrics* **4** 2111–2245.

Nigam, K., McCallum, A. K., Thrun, S. and Mitchell, T. (2000). Text Classification from Labeled and Unlabeled Documents Using EM. *Machine Learning* **39** 103–134.

Nigam, K. P. (2001). *Using Unlabeled Data to Improve Text Classification.* Ph.D. thesis, Carnegie Mellon University, USA. CMU-CS-01-126.

Plan, Y. and Vershynin, R. (2013a). One-Bit Compressed Sensing by Linear Programming. *Communications on Pure and Applied Mathematics* **66** 1275–1297.

Plan, Y. and Vershynin, R. (2013b). Robust 1-Bit Compressed Sensing and Sparse Logistic Regression: A Convex Programming Approach. *IEEE Transactions on Information Theory* **59** 482–494.

Seeger, M. (2002). Learning with Labeled and Unlabeled Data. Tech. Rep. EPFL-REPORT-161327, University of Edinburgh, UK.

Shi, T. and Horvath, S. (2006). Unsupervised Learning with Random Forest Predictors. *Journal of Computational and Graphical Statistics* **15**.

Tallis, G. M. (1961). The Moment Generating Function of the Truncated Multi-normal Distribution. *Journal of the Royal Statistical Society. Series B (Methodological)* **23** 223–229.

Van de Geer, S. and Lederer, J. (2013). The Bernstein–Orlicz Norm and Deviation Inequalities. *Probability Theory and Related Fields* **157** 225–250.

Van der Vaart, A. W. (2000). *Asymptotic Statistics*, vol. 3. Cambridge University Press.

Van Der Vaart, A. W. and Wellner, J. A. (1996). *Weak Convergence and Empirical Processes*. Springer, New York.

Vershynin, R. (2010). Introduction to the Non-Asymptotic Analysis of Random Matrices. *arXiv preprint arXiv:1011.3027* .

Wei, S. and Kosorok, M. R. (2013). Latent Supervised Learning. *Journal of the American Statistical Association* **108** 957–970.

Wilke, R. A., Xu, H., Denny, J. C., Roden, D. M. et al. (2011). The Emerging Role of Electronic Medical Records in Pharmacogenomics. *Clinical Pharmacology & Therapeutics* **89** 379–386.

Yu, S., Chakrabortty, A., Liao, K. P., Cai, T. et al. (2015). Surrogate-Assisted Feature Extraction for High-Throughput Phenotyping. *Preprint* .

Zhang, T. and Oles, F. J. (2000). The Value of Unlabeled Data for Classification Problems. In *Proceedings of the Seventeenth ICML*.

Zhu, L.-X. and Ng, K. W. (1995). Asymptotics of Sliced Inverse Regression. *Statistica Sinica* **5** 727–736.

Zhu, X. (2005). *Semi-Supervised Learning through Graphs.* Ph.D. thesis, Carnegie Mellon University, USA. CMU-LTI-05-192.

Zhu, X. (2008). Semi-Supervised Learning Literature Survey. Tech. Rep. 1530, Computer Sciences, University of Wisconsin-Madison, USA.