Linear Regression With Unmatched Data: A Deconvolution Perspective

Mona Azadkia   M.AZADKIA@LSE.AC.UK
Department of Statistics
London School of Economics and Political Science
London, United Kingdom

Fadoua Balabdaoui   FADOUA.BALABDAOUI@STAT.MATH.ETHZ.CH
Department of Mathematics
ETH Zürich
Zürich, Switzerland

Editor: Marco Cuturi

Abstract

Consider the regression problem where the response $Y \in \mathbb{R}$ and the covariate $X \in \mathbb{R}^d$ for $d \geq 1$ are *unmatched*. Under this scenario, we do not have access to pairs of observations from the distribution of (X, Y), but instead, we have separate data sets $\{Y_i\}_{i=1}^{n_Y}$ and $\{X_j\}_{j=1}^{n_X}$, possibly collected from different sources. We study this problem assuming that the regression function is linear and the noise distribution is known, an assumption that we relax in the applications we consider. We introduce an estimator of the regression vector based on deconvolution and demonstrate its consistency and asymptotic normality under identifiability. Even when identifiability does not hold, we show in some cases that our estimator, the DLSE (Deconvolution Least Squared Estimator), is consistent in terms of an extended ℓ_2 norm. Using this observation, we devise a method for semi-supervised learning, i.e., when we have access to a small sample of matched pairs $\{(X_k, Y_k)\}_{k=1}^m$. Several applications with synthetic and real data sets are considered to illustrate the theory.

Keywords: denoising, convolution, shuffled, regression, semi-supervised, unmatched data, unlinked data

1. Introduction

Consider the standard regression setting

$$Y = m_0(X) + \epsilon \tag{1.1}$$

where $\epsilon \in \mathbb{R}$ is the noise variable, $X \in \mathbb{R}^d$ is the vector of features and $m_0 : \mathbb{R}^d \to \mathbb{R}$ is a measurable function. Given an i.i.d. sample of (Y_i, X_i), the problem of estimating m_0 has been vastly studied in the literature. The estimation problem is much harder when we do not have access to *matched/linked* data, i.e. the pairs (Y_i, X_i) from (1.1), but instead we have separate samples $\{Y_i\}_{i=1}^{n_Y}$ and $\{X_j\}_{j=1}^{n_X}$ such that Y_i and X_j's have the same distribution as Y and X. Under this setting, the exact equality in (1.1) is replaced by the following equality in distribution

$$Y \stackrel{d}{=} m_0(X) + \epsilon. \tag{1.2}$$

This type of data commonly arises in applications when the data has been collected through different sources. It can also occur when the link between the covariates and the response has been deleted because of privacy concerns. This problem is known as *unmatched* or *unlinked* regression in the literature and has been studied in different variations.

Unmatched regression can be seen as a generalization of a problem known as *permuted* regression or *unlabeled sensing*. In permuted regression, one again observes only the variables $\{Y_i\}_{i=1}^n$ and $\{X_i\}_{i=1}^n$ separately. The difference with unmatched regression is that in permuted regression, it is assumed that the link is lost because of the existence of some permutation π^* such that

$$Y_i = m_0(X_{\pi^*(i)}) + \epsilon_i. \tag{1.3}$$

Clearly, permuted regression is a special case of unmatched regression. However, note that the equality (1.3) is stronger than the equality in the distribution in (1.2) for unmatched regression since, in unmatched regression, such a permutation may not exist. Permuted regression has received considerable attention, especially when m_0 is a linear transformation, e.g., Unnikrishnan et al. (2018); Pananjady et al. (2018); Abid et al. (2017); Hsu et al. (2017); Slawski and Ben-David (2019); Tsakiris et al. (2020); Slawski et al. (2020); Tsakiris and Peng (2019); Zhang et al. (2021); Slawski et al. (2021); Rigollet and Weed (2019). In permuted regression, finding the corresponding permutation has also been the focus of many research works. Note that this is a much harder problem than just estimating m_0. If one estimates π^*, by some estimator $\hat{\pi}$ say, then estimating m_0 boils down to a simple estimation with matched data $(X_{\hat{\pi}(i)}, Y_i)$. For a detailed summary of the related work on permuted regression, we refer the reader to Slawski and Sen (2022).

The problem of unmatched regression is of special interest in areas such as microeconomics, where the variable of interest has not been observed jointly with some of the covariates. This problem is also known as data fusion, where multiple observational and experimental data sets exist and the links are unavailable. Popular approaches in this situation include methods based on *matching*, see e.g. Cohen and Richman (2002); Monge et al. (1996); Walter and Fritsch (1999). Matching-based methods rely on the access to an extra *contexual* variable such that they can use this variable to pair the response variable and the covariates.

In the absence of such contextual variables, there is little hope for pairing the variables. In a recent work Carpentier and Schlüter (2016), estimating m_0 in unmatched regression was studied for $d = 1$, assuming that the distribution of ϵ is known and m_0 is monotone without assuming any contextual variable. Note that when m_0 is monotone, this problem is also known as unmatched *isotonic* regression. In Carpentier and Schlüter (2016), authors have shown the close connection between unmatched isotonic regression and deconvolution. In Carpentier and Schlüter (2016) for estimating m_0, authors have resorted to kernel deconvolution as the main estimation technique. Under some smoothness assumptions, the authors have provided the rate of convergence for the proposed kernel estimator obtained with available deconvolution methods in the literature. In Balabdaoui et al. (2021), the authors made use of the tight relationship between the unmatched isotonic regression and deconvolution and provided an estimator for m_0 under the assumption that the noise distribution is known. Their method follows the idea of estimation of the mixing distribution for a normal mean, as done by Edelman (1988). The authors provide a rate of convergence

of a weighted ℓ_1-distance of their estimator of m_0 under several smoothness regimes of the distribution of the noise as well as some regularity conditions satisfied by m_0. For example, in the case where the noise distribution is ordinary smooth with a smoothness parameter equal to β, it follows from their Theorem 1 that the convergence rate of their monotone estimator to the truth when both are restricted on certain compacts cannot be slower than $n^{-1/(2(2\beta+1))}$. In Rigollet and Weed (2019), a minimum Wasserstein deconvolution estimator was suggested that achieves the rate of $O(\log(\log n)/\log n)$ and it was shown that for normally distributed errors, this rate is optimal. In Meis and Mammen (2021) the minimum Wasserstein deconvolution estimator achieves much better rate of $O(n^{-1/(2p)})$ for ℓ_p risks for discrete noise which is optimal. Later in Slawski and Sen (2022), the problem was considered for $d \geq 1$ where the authors have shown that a generalized notion of monotonicity, called cyclical monotonicity, of the regression function, is sufficient for estimation of the regression function m_0. Their method leverages ideas from the theory of optimal transport, specifically the Kiefer-Wolfowitz nonparametric maximum likelihood estimator. Slawski and Sen (2022) provide the rate of convergence of their estimator in terms of ℓ_2-distance under some smoothness assumption on m_0 and assuming that the noise is Gaussian.

Our Contribution: In this work, we consider the problem of estimating the regression vector in unmatched linear regression under the assumption that the noise distribution is known. Our work conceptually follows Balabdaoui et al. (2021), although there is a considerable difference between the linear model and the monotone one. Compared to Balabdaoui et al. (2021); Carpentier and Schlüter (2016), the main contribution of our work is that our theory is valid for any given dimension $d \geq 1$ (though not depending on the sample size of the observations). This framework, to the best of our knowledge, has been only considered in Slawski and Sen (2022). Our proposed deconvolution least squares estimator (DLSE) is not necessarily consistent; therefore, we cannot directly compare our results to those obtained in Slawski and Sen (2022). On the other hand, when we are in a setting in which our estimator is consistent, we do have the faster rate of convergence $O(n^{-1/2})$ compared to the rate of convergence $O(\log(\log n)/\log n)$ in Slawski and Sen (2022), without having to assume that noise is Gaussian. Such a fast rate may come as a big surprise given the non-standard situation of lacking any knowledge of the existing link, even partial, between the responses and covariates. The explanation is that, under the identifiability of the model, the estimation problem can be cast in the usual scope of the theory of M-estimators in parametric models. In the settings where the DLSE is not consistent, we show that a generalized notion of the estimator's norm is consistent, which can be used in semi-supervised scenarios.

Outline of the paper: In Section 2 we introduce our methodology for studying the problem. In Section 3, we provide the main results of the paper by first introducing the estimator and its properties and then applying it in a semi-supervised setting. Section 4 provides simulations on synthetic and real data. We provide the proofs of all lemmas and theorems in the supplementary material.

2. Methodology

Let X be a random vector in $\mathbb{R}^d$ for $d \geq 1$. Consider Y to be a random variable such that $Y \overset{d}{=} \beta_0^T X + \epsilon$ where ϵ is the noise random variable, and $\beta_0 \in \mathbb{R}^d$ is a deterministic

vector of coefficients. In this work we assume that we have access to two independent data sets $\{Y_i\}_{i=1}^{n_Y}$ and $\{X_i\}_{i=1}^{n_X}$ such that $Y_i, i = 1, \ldots, n_Y$ and $X_i, i = 1, \ldots, n_X$ are i.i.d random variables which are distributed as Y and X respectively.

For a random variable or vector Z, we denote by F^Z its cumulative distribution function, that is $F^Z(z) = \mathbb{P}(Z \leq z)$, $z \in \mathbb{R}$. In the case of a vector, the inequality should be considered component-wise. The equality in distribution $Y \stackrel{d}{=} \beta_0^T X + \epsilon$ yields

$$F^Y = F^{\beta_0^T X} \star F^\epsilon \tag{2.1}$$

where $\star$ denotes the convolution operator. Note that β_0 minimizes the function

$$\beta \mapsto \mathcal{D}_{p,\mathbb{Q}}(\beta) := \int \left| F^Y(y) - \left[F^{\beta_0^T X} \star F^\epsilon \right](y) \right|^p d\mathbb{Q}(y) \tag{2.2}$$

where $p \geq 1$ is an integer and $\mathbb{Q}$ is a positive measure on $\mathbb{R}$. In this work, we focus on the case where X has a continuous distribution. Also, we assume that F^ϵ is known. However, we do not assume that ϵ is necessarily Gaussian. Note that the assumption that the noise distribution is known was made in all the prior works on unmatched regression. In the applications below, we relax this assumption by letting the scale parameter of the distribution of the noise unspecified. Deriving the theory for such a relaxation is still needed as the estimation procedure needs to be extended to allow for the additional estimator of the scale parameter.

A natural choice of $\mathbb{Q}$ is F^Y, the probability measure induced by the distribution of Y on $\mathbb{R}$. Given a sample $\{Y_1, \cdots, Y_{n_Y}\}$ we can use the empirical estimate of F^Y

$$F_{n_Y}^Y(y) := n_Y^{-1} \sum_{i=1}^{n_Y} \mathbb{1}_{(-\infty, y]}(Y_i).$$

By using $F_{n_Y}^Y$ as a surrogate of $\mathbb{Q}$, we get the empirical version of (2.2)

$$\mathbb{D}_{n_Y, n_X, p}(\beta) := \int \left| F_{n_Y}^Y(y) - n_X^{-1} \sum_{i=1}^{n_X} F^\epsilon(y - \beta^T X_i) \right|^p dF_{n_Y}^Y(y). \tag{2.3}$$

Without loss of generality and for the sake of a simpler notation, we assume that $n_Y = n_X = n$. Also, we let $\mathbb{D}_{n,p}(\beta) := \mathbb{D}_{n_Y, n_X, p}(\beta)$. The main idea of this paper is to take the solution of the optimization problem $\min_{\beta \in \mathbb{R}^d} \mathbb{D}_{n,p}(\beta)$ as an estimate of β_0. In Section 3, we study this optimization problem in detail.

3. Main Result

Notation and Definition. Let C^k be the class of all functions on $\mathbb{R}$ such that their first k derivatives exist and are continuous. Let $\mathcal{B}_0$ be the set of all $\beta \in \mathbb{R}^d$ such that the unlinked linear regression model in (2.1) holds, i.e. $\mathcal{B}_0 = \{\beta^T X \stackrel{d}{=} \beta_0^T X\}$. Also, let us write $\mathcal{B}_{n,p} := \{\beta \in \mathbb{R}^d \text{ s.t. } \mathbb{D}_{n,p}(\beta) = \min_\beta \mathbb{D}_{n,p}(\beta)\}$. For $\beta \in \mathbb{R}^d$, the Euclidean norm of β is defined as $\|\beta\|_2 = \sqrt{\sum_{i=1}^d \beta_i^2}$. Let $S^{d-1} = \{x \in \mathbb{R}^d : \|x\|_2 = 1\}$ be the d-sphere. For Σ a positive-definite matrix, $\|x\|_{2,\Sigma} := \sqrt{x^T \Sigma x}$ for $x \in \mathbb{R}^d$ is a norm.

3.1 The Deconvolution Least Squares Estimator

Given samples $\{Y_i\}_{i=1}^n$ and $\{X_i\}_{i=1}^n$ let $\hat{\beta}_n := \operatorname{argmin}_{\beta \in \mathbb{R}^d} \mathbb{D}_{n,p}(\beta)$. In the following proposition, we show that when ϵ admits a continuous distribution, there exists at least a $\hat{\beta}_n \in \mathbb{R}^d$ such that it minimizes $\mathbb{D}_{n,p}(\beta)$.

Proposition 1 *Assume that F^ϵ is continuous on $\mathbb{R}$. For a given integer $p \geq 1$, and for $n \geq 8d$ with probability 1 we have $|\mathcal{B}_{n,p}| \geq 1$.*

Note that $\mathcal{B}_0$ may have more than one element. For example for $X = [X^1, \cdots, X^d]^T \in \mathbb{R}^d$ where X^i are exchangeable random variables it is easy to see that we have $\beta_0^\pi := [\beta_0^{\pi(1)}, \cdots, \beta_0^{\pi(d)}]^T \in \mathcal{B}_0$ where π is any permutation on $[d]$. The other interesting case is when $X = [X^1, \cdots, X^d]^T \in \mathbb{R}^d$ is distributed as $\mathcal{N}(0, \Sigma)$. In Lemma 2 we show that in this case $\mathcal{B}_0$ can be characterized in terms of $\|\cdot\|_{2,\Sigma}$.

Lemma 2 *Suppose the unlinked linear regression model in (2.1) holds with $X \sim \mathcal{N}(0, \Sigma)$ where Σ is non-singular. Then, there exists a constant $c > 0$ such that*

$$\mathcal{B}_0 = \left\{ \beta \in \mathbb{R}^d : \|\beta\|_{2,\Sigma} = c \right\}.$$

Since $\mathcal{B}_0$ is not necessarily a singleton, it is impossible to have a consistency result in the classical sense. On the other hand, with the characterization of $\mathcal{B}_0$ via the norm constraint, Theorem 3 shows that $\hat{\beta}_n$ is consistent in terms of such a norm constraint. From now on, we will take $p = 2$. This means that our deconvolution least squares estimator is any vector $\hat{\beta}_n$ which minimizes the criterion (2.3) for $p = 2$.

Theorem 3 *Suppose that F^ϵ is continuous on $\mathbb{R}$. If there exists a non-singular positive-definite $d \times d$ matrix Γ such that $\|\beta\|_{2,\Gamma} = c$ for all $\beta \in \mathcal{B}_0$ and some $c \in [0, \infty)$, then if $\|\hat{\beta}_n\|_2 = O_\mathbb{P}(1)$ we have*

$$\|\hat{\beta}_n\|_{2,\Gamma} \xrightarrow{\mathbb{P}} c.$$

Note that the result in Theorem 3 requires that $\|\hat{\beta}_n\|_2 = O_\mathbb{P}(1)$. The following proposition shows that when $X \sim \mathcal{N}(0, \Sigma)$ this requirement is met, and therefore we have the consistency of the DLSE $\hat{\beta}_n$ in terms of $\|\cdot\|_{2,\Sigma}$.

Proposition 4 *Suppose that $X \sim \mathcal{N}(0, \Sigma)$ for some non-singular Σ. Then, $\|\hat{\beta}_n\|_2 = O_\mathbb{P}(1)$ and*

$$\|\hat{\beta}_n\|_{2,\Sigma} \xrightarrow{\mathbb{P}} c, \quad and \quad \|\hat{\beta}_n\|_{2,\widehat{\Sigma}_n} \xrightarrow{\mathbb{P}} c,$$

where $c = \sqrt{\beta^T \Sigma \beta}$ for any $\beta \in \mathcal{B}_0$ for which the unlinked linear regression model in (2.1) holds, and $\widehat{\Sigma}_n$ is any consistent estimator of Σ.

In the case where $|\mathcal{B}_0| = 1$, meaning that there exists a unique $\beta_0 \in \mathbb{R}^d$ such that $Y \stackrel{d}{=} \beta_0^T X + \epsilon$ we show that $\hat{\beta}_n$ is a consistent estimator in the classical sense. In Theorem 5, we show that under this uniqueness assumption $\hat{\beta}_n \to \beta_0$ in probability.

Theorem 5 *Assume that F^ϵ admits a density f^ϵ and $|\mathcal{B}_0| = 1$. In addition assume that $\mathbb{P}(\alpha^T X = 0) = 0$ for any $\alpha \in \mathbb{R}^d$ such that $\alpha \neq 0$. Then for all $\hat{\beta}_n \in \mathcal{B}_{n,2}$ we have*

$$\hat{\beta}_n \xrightarrow{\mathbb{P}} \beta_0.$$

When $|\mathcal{B}_0| > 1$, then depending on the distribution of X, we might be able to narrow down $\mathcal{B}_{n,2}$ by more than just specifying the norm of the members. In Theorem 6 below, we show that for a certain family of probability distributions, $\mathcal{B}_0$ consists of all vectors that result from permuting the components of a member β_0.

Theorem 6 *Suppose that the components of the covariate X, $X^1, \cdots, X^d$ are i.i.d. with moment generating function that takes the form*

$$G_{X^1}(t) := E[e^{tX^1}] = \frac{1}{(\lambda - t)^\alpha} H(t), \quad for\ t < \lambda,$$

where $0 < \lambda$, $\alpha \in (0, \infty)$, and H is a continuous function such that

$$\lim_{t \nearrow \lambda} H(t) \in (0, \infty).$$

Then the set of all β such that $Y \overset{d}{=} \beta_0^T X + \epsilon$ is

$$\mathcal{B}_0 = \left\{ \beta \in \mathbb{R}^d : \exists\ permutation\ \pi\ such\ that\ \beta_{0,i} = \beta_{\pi(i)},\ i = 1, \ldots, d \right\}.$$

Note that this family includes exponential and Gamma distributions as well as any convolution thereof.

Finally, in Theorem 7, where there exists a unique regression vector for which the unmatched regression model holds we are able to show that under some smoothness assumptions the DLSE $\hat{\beta}_n$ is asymptotically normal with mean β_0 and covariance matrix that depends on the distribution of X, that of ϵ and the true regression vector β_0.

Theorem 7 *Suppose $F^\epsilon \in C^2$ and there exists $M > 0$ such that $f^\epsilon \leq M$ and $|(f^\epsilon)^{(1)}| \leq M$. Also assume that there exists an integer $m \geq 1$ and $a_1 < \cdots < a_m$ such that f^ϵ is monotone on $(-\infty, a_1], (a_1, a_2), \cdots, [a_m, \infty)$. Furthermore, we assume that $\mathbb{E}\|X\|_2^2 < \infty$, and the matrix*

$$U = \int \left(\int x f^\epsilon(y - \beta_0^T x) dF^X(x) \right) \left(\int x^T f^\epsilon(y - \beta_0^T x) dF^X(x) \right) dF^Y(y)$$

is positive-definite. Then when $|\mathcal{B}_0| = 1$ we have

$$\sqrt{n}(\hat{\beta}_n - \beta_0) \xrightarrow{d} U^{-1} \mathbb{V}$$

where

$$\mathbb{V} = - \int_{\mathbb{R}} \left(\mathbb{B}_1 \circ F^Y(y) - \int \mathbb{B}_2 \circ F^Z(z) f^\epsilon(y - z) dz \right)$$
$$\left(\int x f^\epsilon(y - \beta_0^T x) dF^X(x) \right) dF^Y(y),$$

with $Z = \beta_0^T X$, and $\mathbb{B}_1$ and $\mathbb{B}_2$ are two independent standard Brownian Bridges from $(0,0)$ to $(1,0)$. The random vector $U^{-1}\mathbb{V}$ is distributed as $\mathcal{N}(\mathbf{0}, U^{-1}\mathbb{E}[\mathbb{V}\mathbb{V}^T]U^{-1})$.

The conditions required for f^ϵ are satisfied by most of the well-known densities, including normal, Laplace, symmetric Gamma densities and any finite convolution thereof. The asymptotic covariance matrix has a complicated dependence on the parameters of the model. However, the exact knowledge of this asymptotic covariance matrix is not at all necessary to make useful inferences. In fact, the asymptotic result of Theorem 7 allows us to use re-sampling techniques to infer the true regression vector. Therefore, one can resort to bootstrapping in order to find an approximation of asymptotic confidence bounds for β_0.

Theorem 7 provides the asymptotic normality of the DLSE only under the identifiability of the unmatched linear regression model, and hence it cannot be used beyond the scope where $|\mathcal{B}_0| = 1$.

When we are in the situation of Theorem 6, and if the components of $\beta_0 \in \mathcal{B}_0$ are all distinct, then the model becomes identifiable if the components of β_0 are ordered from smallest to largest. In fact, the arguments in the proof of Theorem 8 can be used again to show the following result.

Theorem 8 *Suppose that F^ϵ satisfies the same conditions as in Theorem 6. Suppose also that the assumption of Theorem 6 holds and that the components of the vector $\beta_0 \in \mathcal{B}_0$ are all distinct. Denote $b_0 = [\beta_0^{(1)}, \cdots, \beta_0^{(d)}]^T$ the vector of ordered components of β_0 from smallest to largest, and $\hat{b}_n = [\hat{\beta}_n^{(1)}, \cdots, \hat{\beta}_n^{(d)}]^T$ the vector of ordered components of the DLSE, $\hat{\beta}_n$. Then,*

$$\sqrt{n}(\hat{b}_n - b_0) \xrightarrow{d} U^{-1}\mathbb{V},$$

where U and $\mathbb{V}$ are defined in Theorem 6.

To close this section, we would like to stress the fact that the fast rate of convergence, $1/\sqrt{n}$, is obtained under the very important condition of identifiability of β_0, or in other words its uniqueness. When this condition is satisfied, arguments from the theory of M-estimators can be used. In this case, the $1/\sqrt{n}$-rate can be obtained since the estimation problem is fully parametric. In case identifiability is not satisfied, Theorem 6 provides a sufficient condition on the distribution of the covariates which guarantees identifiability of the ordered values of the regression vector. Although the condition given in that theorem encompasses many distributions, a unified result with more general identifiability conditions is yet to be established.

3.2 Application to Semi-supervised Learning

As established above, the estimator $\hat{\beta}_n$ is not consistent in case of non-identifiability. Therefore, one cannot possibly think of $\hat{\beta}_n$ as something that will be close to some unique truth since the latter does not even exist. However, there are situations where some feature of the model is unique, and in this case, one would expect that the DLSE succeeds in remaining faithful to such features.

Consider the case where $X \sim \mathcal{N}(0, \Sigma)$ with a non-singular covariance matrix Σ. Lemma 2 describes the set $\mathcal{B}_0$ in terms of a norm with respect to Σ: There exists a constant $c > 0$ such that for any $\beta \in \mathcal{B}_0$ we have $\|\beta\|_{2,\Sigma} = c$. In addition, Proposition 4 guarantees that although we do not have consistency of $\hat{\beta}_n$, it holds that $\|\hat{\beta}_n\|_{2,\hat{\Sigma}_n} \xrightarrow{\mathbb{P}} c$. In this case if we have

access to a third sample of matched data, we can benefit from the result of Proposition 4 together with an estimate of β_0 from the matched sample.

Let $\{(\tilde{Y}_i, \tilde{X}_i)\}_{i=1}^{m}$ be a set of matched data such that $Y_i = \beta_0^T X_i + \epsilon_i$. Take $\tilde{\beta}_m$ to be an estimator of β_0 based on only $\{(\tilde{Y}_i, \tilde{X}_i)\}_{i=1}^{m}$. For example, one can take $\tilde{\beta}_m$ to be the ordinary least squared (OLS) estimator. Let $\widehat{\Sigma}_n$ be a consistent estimator of Σ, the covariance matrix of X using the sample $\{X_i\}_{i=1}^{n}$ (here, an estimate based on this sample is more accurate since n is much bigger than m). Now consider the following modified estimator

$$\beta_{n,m}^{\dagger} = \tilde{\beta}_m \frac{\|\hat{\beta}_n\|_{2,\widehat{\Sigma}_n}}{\|\tilde{\beta}_m\|_{2,\widehat{\Sigma}_n}}. \tag{3.1}$$

When $n \gg m$, we expect that $\|\hat{\beta}_n\|_{2,\widehat{\Sigma}_n}$ to be a better estimate of $\|\beta_0\|_{2,\Sigma}$ compared to $\|\tilde{\beta}_m\|_{2,\widehat{\Sigma}_n}$. Therefore, while the estimate based on the unlinked data fails to provide a meaningful estimate of the direction of β_0, we can still use $\|\hat{\beta}_n\|_{2,\widehat{\Sigma}_n}$ to modify the norm of $\tilde{\beta}_m$ and hence improve the performance of the latter.

Unfortunately, we do not yet have a rate of convergence for $\|\hat{\beta}_n\|_{2,\widehat{\Sigma}_n}$. In Section 4 we provide experimental results suggesting that for large enough n one can use $\|\hat{\beta}_n\|_{2,\widehat{\Sigma}_n}$ as an estimate of $\|\beta_0\|_{2,\Sigma}$. As we do not have the right arguments which show theoretically that the modified estimator $\beta_{n,m}^{\dagger}$ in (3.1) improves the performance $\tilde{\beta}_m$, we resort to another estimator of the norm $\|\beta_0\|_{2,\Sigma}$ and whose convergence rate can be established. The estimator is based on the following simple observation:

$$\mathrm{Var}(Y) = \beta_0^T \Sigma \beta_0 + \mathrm{Var}(\epsilon) = \|\beta_0\|_{2,\Sigma}^2 + \mathrm{Var}(\epsilon).$$

Therefore, under the assumption of known distribution of ϵ, it is possible to estimate $\|\beta_0\|_{2,\Sigma}$ by estimating $\mathrm{Var}(Y)$. The following remark shows that when $m = o(n)$, using the information from unmatched data improves the performance of the OLS estimator.

Remark 9 *Let $\tilde{\beta}_m$ be the ordinary least square estimate of β_0 using the matched data $\{(\tilde{Y}_i, \tilde{X}_i)\}_{i=1}^{m}$. Let $r_n^2 = \frac{1}{n-1}\sum_{i=1}^{n}(Y_i - \bar{Y}_n)^2 - \mathrm{Var}(\epsilon)$ where $\bar{Y}_n$ is the sample mean of $\{Y_i\}_{i=1}^{n}$. Define*

$$\tilde{\beta}_{n,m} = \tilde{\beta}_m \frac{r_n}{\|\tilde{\beta}_m\|_{2,\widehat{\Sigma}_n}},$$

where $\widehat{\Sigma}_n$ is the empirical covariance matrix of $\{X_i\}_{i=1}^{n}$. When $m = o(n)$

$$\mathbb{E}[\|\tilde{\beta}_{n,m} - \beta_0\|^2] < \mathbb{E}[\|\tilde{\beta}_m - \beta_0\|^2],$$

where expectations on the right and left sides are taken with respect to the distribution of the matched data and the product distribution of the matched and unmatched, respectively.

4. Applications to Synthetic and Real Data

In this section, we present the result of several simulations with the goal of illustrating the theoretical results derived above. In addition, we shall consider real applications using

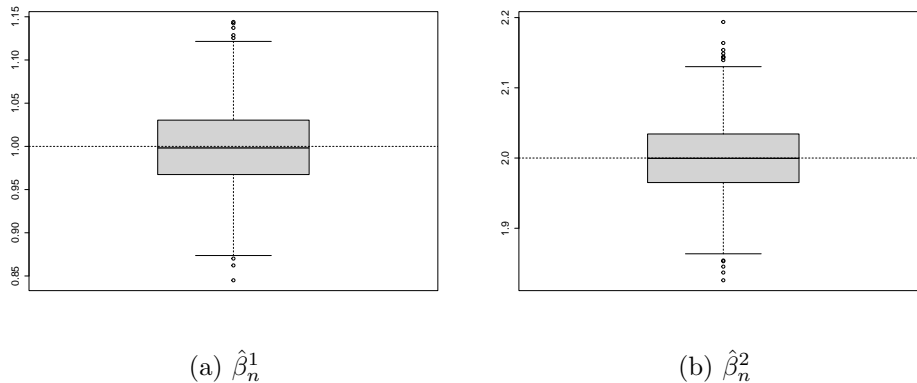(a) $\hat{\beta}_n^1$ (b) $\hat{\beta}_n^2$

Figure 1: Boxplots of $\hat{\beta}_n^1$ and $\hat{\beta}_n^2$ for 100 independent samples of size $n = 4000$. Root mean squared error of $\hat{\beta}_n^1$ is 0.047, and root mean squared error of $\hat{\beta}_n^2$ is 0.052.

two different real data sets: The (unmatched) inter-generational mobility data set already analyzed by several authors; e.g. Olivetti and Paserman (2015) and the (matched) Power Plant data set to which we apply ideas from semi-supervised learning. Computation of the DLSE, especially for large data sets (synthetic or not) and for several runs impose numerical challenges. One major issue is that the minimization problem is generally not convex and can be multi-modal especially in the case of non-identifiability. We use the default setting of the function "optim" from the package "stats" of the open software R. The default method of optimization is the method introduced by Nelder and Mead (1965).

4.1 Synthetic Data

Example 1 *For this example, we generated 1000 independent samples of Y and X where $Y \overset{d}{=} \beta_0^T X + \epsilon$ of size $n = 4000$ such that*

$$X = [X^1, X^2]^T \in \mathbb{R}^2, \qquad X^1 \sim \mathcal{N}(1, 1), \qquad X^2 \sim Exp(1),$$
$$\epsilon \sim \mathcal{N}(0, 1), \qquad \beta_0^T = [1, 2],$$

where X^1, X^2 and ϵ are independent. In this case, and as shown in the supplementary material, $\mathcal{B}_0$ contains only β_0. Therefore, we expect that $\hat{\beta}_n$ is a consistent estimator of β_0. Figure 1 shows the boxplot of $\hat{\beta}_n^1$ and $\hat{\beta}_n^2$ based on the 100 replications.

Theorem 7 shows that $\sqrt{n}(\hat{\beta}_n - \beta_0)$ has asymptotically a normal distribution. This type of result can be validated by plotting the quantiles $\sqrt{n}(\hat{\beta}_n - \beta_0)$ against those of a standard normal variable. Such a plot is commonly known under the name of qqplot. Figure 2 shows such qqplots for the components of $\sqrt{n}(\hat{\beta}_n - \beta_0)$. Their linear shape is much aligned with the asymptotic normality of our estimator as stated in Theorem 7.

Using unmatched data, at best, means that we only have access to the generating mechanism of X and Y without knowing the link between them. Lack of information about the link between the X's and Y's should come with a price paid on the performance of the estimate of β_0. To see how much our DLSE $\hat{\beta}_n$ suffers from this lack of knowledge, we compare the performance of $\hat{\beta}_n$ with the ordinary least square estimator $\tilde{\beta}_n$ using a matched sample
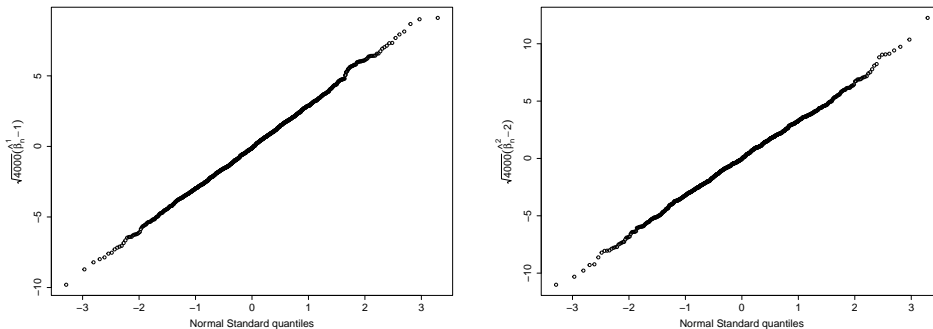
Figure 2: qqplot of $\sqrt{n}(\hat{\beta}_n^1 - 1)$ and $\sqrt{n}(\hat{\beta}_n^2 - 2)$ against $\mathcal{N}(0,1)$, with $n = 4000$. These plots are in line with the asymptotic result of Theorem 7.

of size $n = 4000$ over 100 independent replications. The comparison is done in terms of absolute prediction error. For this purpose we vary the noise strength level $\epsilon \sim \mathcal{N}(0, \sigma_\epsilon^2)$ for $\sigma_\epsilon \in \{0.6, 0.8, 1, 1.2\}$ to see how much it effects the performance. In Table 1, the result is summarized in terms of the ratio of the absolute error of the OLS estimator to the absolute error of DLSE for each value of σ_ϵ and for each component of β_0. Clearly, $\tilde{\beta}_n$ beats $\hat{\beta}_n$ as it uses more information, but given the fact that $\hat{\beta}_n$ is oblivious to the link between the response and covariate, its performance is rather quite satisfactory.

	$\sigma_\epsilon = 0.6$	$\sigma_\epsilon = 0.8$	$\sigma_\epsilon = 1$	$\sigma_\epsilon = 1.2$	$\sigma_\epsilon = 2$
$\text{mean}(|\tilde{\beta}_n^1 - \beta_0^1|)/\text{mean}(|\hat{\beta}_n^1 - \beta_0^1|)$	0.14	0.26	0.32	0.28	0.61
$\text{mean}(|\tilde{\beta}_n^2 - \beta_0^2|)/\text{mean}(|\hat{\beta}_n^2 - \beta_0^2|)$	0.18	0.16	0.18	0.25	0.55

Table 1: Ratio of mean absolute error of $\tilde{\beta}_n$ to $\hat{\beta}_n$ as a function of the standard deviation of the noise.

Example 2 *We consider the following setting*

$$Y \overset{d}{=} \beta_0^T X + \epsilon, \qquad X \sim \mathcal{N}(0, I_2), \qquad \epsilon \sim \mathcal{N}(0, 1), \qquad \beta_0^T = [1, 2],$$

where I_2 is the 2×2 identity matrix. Note that in this case, $\mathcal{B}_0$ contains more than one element. In fact by Lemma 2 we know that $\mathcal{B}_0 = \{\beta \in \mathbb{R}^2 : \|\beta\|_{2,\Sigma} = \sqrt{5}\}$. Proposition 4 guarantees that $\|\hat{\beta}_n\|_{2,\Sigma}$ converges to $\|\beta_0\| = \sqrt{5}$ in probability. The scatterplot of $\hat{\beta}_n$ shown in Figure 3 gives a clear illustration of this fact. The mean and standard deviation of $\|\hat{\beta}_n\|_2^2$ over the 1000 replications with sample size $n = 4000$ were found to be 5.01 and 0.15 respectively.
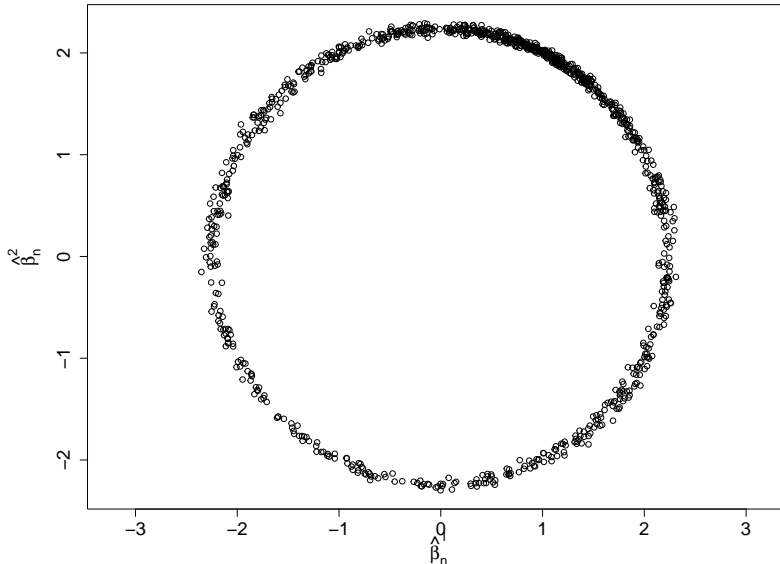
Figure 3: Plot of $\hat{\beta}_n$ for $X \sim \mathcal{N}(0, I_2)$ and $\beta_0^T = [1, 2]$. In this case $\mathcal{B}_0 = \{\beta \in \mathbb{R}^2 : \|\beta\|_2 = \sqrt{5}\}$.

Example 3 *In this example, we generated 1000 times independent samples of Y and X where $Y \stackrel{d}{=} \beta_0^T X + \epsilon$ with size $n = 8000$ such that*

$$X = [X^1, X^2, X^3]^T \in \mathbb{R}^3, \qquad X^i \sim Exp(1) \text{ for } i = 1, 2, 3 \;,$$
$$\epsilon \sim \mathcal{N}(0, 1), \qquad \beta_0^T = [1, 2, -1]$$

where X^1, X^2, X^3 and ϵ are independent. Theorem 6 implies that $|\mathcal{B}_0| = 3! = 6$, since all permutations of β_0 lead to the same distribution. Instead of $\hat{\beta}_n$ we take a look at $\hat{\beta}_{n,ordered} = [\hat{\beta}_n^{(1)}, \hat{\beta}_n^{(2)}, \hat{\beta}_n^{(3)}]^T$ where $\hat{\beta}_n^{(i)}$ is the i-th smallest element of $\hat{\beta}_n$. As mentioned earlier, when $|\mathcal{B}_0| > 1$, we do not have any asymptotic normality result but believe that $\hat{\beta}_{n,ordered}$ is asymptotically normal under some regularity assumptions. Figure 4 shows the qqplots of $\sqrt{8000}(\hat{\beta}_n^{(i)} - \beta_0^{(i)})$ against the quantiles of $\mathcal{N}(0, 1)$, and which is supported by Theorem 8. Figure 5 also shows the corresponding boxplots.

This example suggests that in cases where we do not have uniqueness, looking at the ordered version of $\hat{\beta}_n$ can provide additional information. Such a piece of information can be very valuable in the presence of a small matched data set or expert knowledge. In this case, one might even be able to recover the permutation that maps $\hat{\beta}_{n,ordered}$ to a consistent estimator of β_0.

Example 4 *For this example we generated 100 samples of size $n = 10^4$ from model (1.2) such that $Y \stackrel{d}{=} \beta_0^T X + \epsilon$ and*

$$X \sim \mathcal{N}(0, I_3), \qquad \epsilon \sim \mathcal{N}(0, 1), \qquad \beta_0^T = [1, -1, 2].$$

We also generate 100 samples of matched data $\{(\tilde{Y}_i, \tilde{X}_i)\}_{i=1}^m$ from model (1.1) for $m \in \{10, 20, \cdots, 100\}$. Note that $\mathcal{B}_0 = \{\beta \in \mathbb{R}^3 : \|\beta\|_2 = \sqrt{6}\}$. We use the unlinked data for
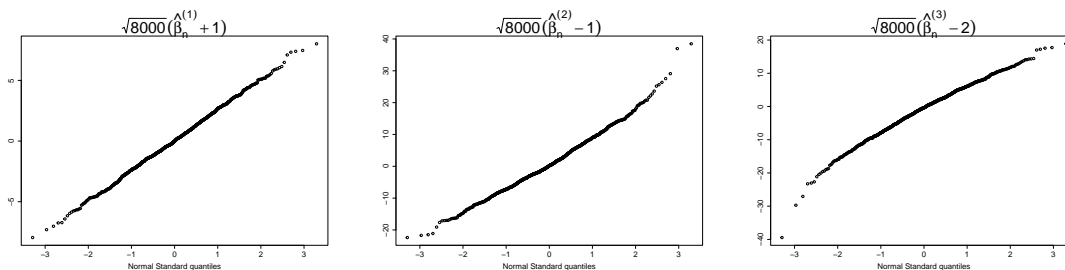
11

Figure 4: qqplot of $\sqrt{n}(\hat{\beta}_n^{(1)} + 1)$, $\sqrt{n}(\hat{\beta}_n^{(2)} - 1)$, and $\sqrt{n}(\hat{\beta}_n^{(3)} - 3)$ against $\mathcal{N}(0,1)$, with $n = 8000$. These plots are in line with the asymptotic result of Theorem 8.
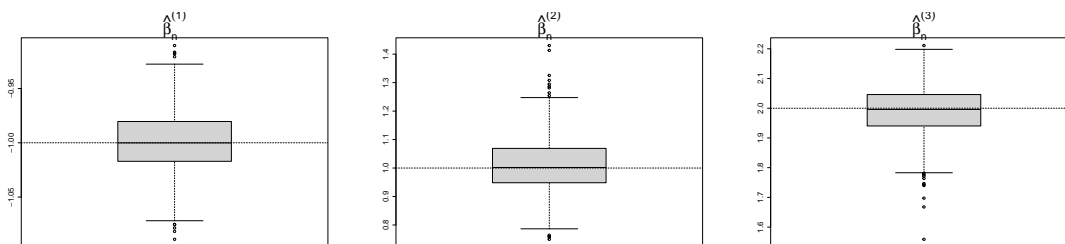

Figure 5: Boxplot of $\hat{\beta}_n^{(1)}$, $\hat{\beta}_n^{(2)}$, and $\hat{\beta}_n^{(3)}$ with $n = 8000$ for 1000 iterations. The standard deviations are 0.02, 0.09, and 0.07 respectively.

computing $\hat{\beta}_n$ and the linked data for computing the OLS estimator of β_0, $\tilde{\beta}_m$. In Table 2, the mean and standard deviation of these estimators over the 100 replications are shown. It can be seen that $\|\hat{\beta}_n\|_2^2$ is highly concentrated around $\|\beta_0\|_2^2$ compared to the OLS estimators with smaller sample sizes.

	$\|\tilde{\beta}_m\|_2^2$, square norm of OLS estimate with sample size m										$\|\hat{\beta}_n\|_2^2$
m	10	20	30	40	50	60	70	80	90	100	
mean	6.44	6.23	6.07	6.02	6.12	6.04	6.11	6.25	6.02	6.11	6.00
sd	2.03	1.23	0.87	0.80	0.66	0.65	0.58	0.56	0.52	0.55	0.16

Table 2: Mean and standard deviation of $\|\tilde{\beta}_m\|_2^2$ for values of $m \in \{10, 20, \cdots, 100\}$ for 100 independent iterations. The last column contains the mean and standard deviation of $\|\hat{\beta}_n\|_2^2$.

4.2 Intergenerational Mobility in the United States, 1850-1930

The degree to which the economic status is passed along generations is an important factor in quantifying the inequality in a society. Researchers have approached this problem by studying the relationship between the father/father-in-law's income and the son/son-in-law's income. It is intuitive to assume that there is an increasing relationship between the son's income and that of the father. If we assume that this relationship is linear, the magnitude of the coefficient can quantify the existing inequality: the stronger the relationship, the more inequality. For this purpose, we apply our method to data from 1850

to 1930 decennial censuses of the United States studied in Olivetti and Paserman (2015); D'Haultfoeuille et al. (2022) using the 1 percent IPUMS samples (Ruggles et al., 2010). We follow Olivetti and Paserman (2015) and focus only on white father/father-in-law and son/son-in-law relationships. In this available historical data on father-son income in the United States, the link between the father/father-in-law and son/son-in-law is not available. Other studies on this data have used information on the first name to reconstruct the link between the father/father-in-law and son/son-in-law, but we only use the unmatched data and do not look at these partially reconstructed links.

Since the exact value of the income in this data set is not available, we use the provided OCCSCORE. OCCSCORE assigns each occupation in all years a value representing the median total income (in hundreds of 1950 dollars) of all persons with that particular occupation in 1950. Therefore, using this score, we lose the within-occupation variation of the income.

We let Y be the son/son-in-law's OCCSCORE and X to be the father/father-in-law's OCCSCORE and assume that $Y \stackrel{d}{=} \beta_0 X + \epsilon$. The sample sizes in the data sets are quite large $(n_X, n_Y > 10^7)$ and therefore, for computational reasons, we select a subset of size 4000 of X and Y at random. We centered and normalized each sub-sample such that they have a mean of 0 and a variance of 1 but kept the same names (Y and X) for the transformed variables. In Figure 6, we show the sorted values of Y (Son's income) plotted against those of X (Father's income) for a sub-sample of size 10^6.

Since the true distribution of the noise variable ϵ is unknown in this problem, we consider two different families of centred distributions, Normal and Laplace. We consider different possible values for their scale parameters so that the standard deviation (sd) of the noise varies in the set $\{0.1, 0.2, \cdots, 1\}$. Recall that the standard deviation of a random variable distributed as $\mathcal{N}(0, \sigma^2)$ is σ and as Laplace(λ) is $\sqrt{2}\lambda$. Therefore, we choose the parameters σ and λ accordingly. Note that the data is discrete as there are many repeated values for each of X and Y. Varying σ and λ, the respective parameters of the distributions $\mathcal{N}(0, \sigma^2)$ and Laplace(λ), such that the standard deviation takes its value in the set $\{0.2, 0.3, \cdots, 1\}$ allows us to plot the values of $\mathbb{D}_{n,2}(\hat{\beta}_n)$ in Figure 7. Assuming that the noise is Gaussian, we observe that $\mathbb{D}_{n,2}(\hat{\beta}_n)$ takes its minimum over the grid of values of $\sigma \in \{0.1, 0.2, \cdots, 1\}$ at $\sigma = 0.2$ and the optimization procedure yields $\hat{\beta}_n = 1.27$. Similarly assuming that the noise follows Laplace(λ) with parameter λ such that $\sqrt{2}\lambda \in \{0.1, 0.2, \cdots, 1\}$ we find the minimum to be attained for $\lambda = 0.2/\sqrt{2}$ and $\hat{\beta}_n = 1.18$.

Relying on the observation that $\mathbb{D}_{n,2}(\hat{\beta}_n)$ is minimized for sd $= 0.2$ for both the Normal and Laplace distributions over 100 iterations we take independent sub-samples of size $n = 4000$ and calculate $\hat{\beta}_n$ using Normal and Laplace distribution for the noise with standard deviation equal 0.2. For both distributions, the estimate $\hat{\beta}_n$ shows a bi-modal behaviour with positive and negative modes with the same magnitude. This is to be expected in case the covariate has a zero expectation. However, as we know that the relationship between Y and X should be non-decreasing, we can consider the absolute value of $\hat{\beta}_n$ in all cases. Table 3 shows the values of mean and standard deviation for $|\hat{\beta}_n|$ using the Normal and Laplace distributions for the noise. Note that Theorem 7 guarantees asymptotic normality of $\hat{\beta}_n$ when $|\mathcal{B}_0| = 1$ under some regularity assumptions. Such assumptions do not seem to be fulfilled in this dataset because of its discreteness. However, one may use sub-sampling
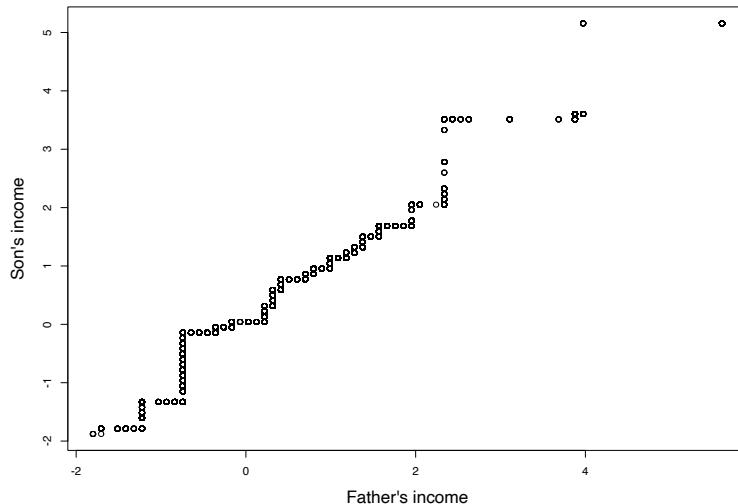
Figure 6: The sorted values of Y (Son's transformed income) are plotted against the sorted values of X (Father's transformed income) for a sub-sample of size 10^6.
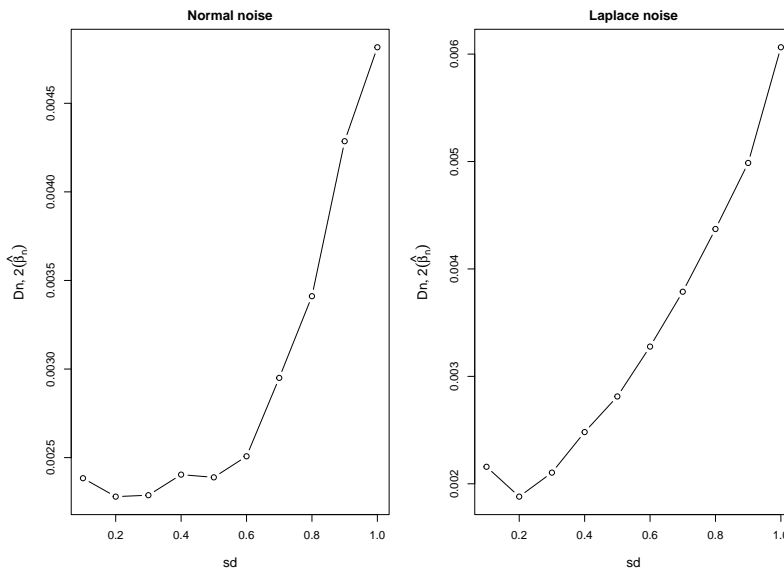

Figure 7: Values of $\mathbb{D}_{n,2}(\hat{\beta}_n)$ for different values of standard deviation (sd) when the noise distribution is assumed to be $\mathcal{N}(0, \sigma^2)$ (sd $= \sigma$), and Laplace(λ) (sd $= \sqrt{2}\lambda$).

ideas to create confidence intervals based on $\hat{\beta}_n$, hoping that some asymptotic normality holds.

14

$|\hat{\beta}_n|$	Normal (sd = 0.2)	Laplace (sd = 0.2)
mean	1.17	1.14
sd	0.06	0.09
95% Bootstrap confidence interval	$(1.07, 1.28)$	$(1.00, 1.26)$

Table 3: Mean and standard deviation for $|\hat{\beta}_n|$ using the Normal and Laplace as the noise distribution over 100 sub-samples of size $n = 4000$. The last row of the table shows the 95% Bootstrap confidence interval using the 100 sub-samples.

In this example, the covariate X is 1-dimensional, and the distribution of data is far from continuous. Still, we found this example interesting as it is the only real data problem with unmatched data that we could access to.

4.3 Power Plant Data Set

We consider the Power Plant data set from UCI Machine Learning Repository[1]. The data set contains 9568 matched data points collected from a combined cycle power plant. Features consist of hourly average ambient variables Temperature (T), Ambient Pressure (AP), Relative Humidity (RH) and Exhaust Vacuum (V) to predict the net hourly electrical energy output (EP) of the plant. Assuming that EP is a linear function of T, AP, RH, and V, we run ordinary least squares using all the data points. We get $R^2 = 0.93$, which supports the linearity of the relationship. We perform 100 independent simulations in the following manner: For each simulation, we select a sub-sample of matched data of size $m = 30$. We use this subsample of matched data for both $\tilde{\beta}_m$ and estimate the distribution of ϵ. Then, from the remainder of the data, we select a sub-sample of unmatched data of size $n = 4000$. This is done by selecting a sample of size $n = 4000$ from the remaining Y_i's and independently selecting a sample of size $n = 4000$ from the remaining X_i's. Since we do not have access to the population density in this case, we take the OLS estimate $\beta_{0,\mathrm{OLS}}$ using all the data points as the ground truth.

For the DLSE estimator $\hat{\beta}_n$, we need an estimate of the noise distribution, and for this, we use a Kernel density estimator based on the residuals of OLS $\tilde{\beta}_m$ obtained using the matched data. We use a Gaussian kernel and select the bandwidth according to Sheather and Jones (1991).[2]

Figure 8 depicts the scatterplots of pairs of components of the obtained DLSE $\hat{\beta}_n$ over 1000 sub-samples of size $n = 4000$ together with the projection of $\beta_{0,\mathrm{OLS}}$ onto the corresponding sub-spaces. The grey points in the scatterplots correspond to components of $\tilde{\beta}_m$. For example, the plot in the first row and the second column is the scatter plot of $\hat{\beta}_n^1$ vs $\hat{\beta}_n^2$ (and $\tilde{\beta}_m^1$ vs $\tilde{\beta}_m^2$) which shows that the first two components of $\hat{\beta}_n$ form 3 main clusters. Therefore, the scatterplots suggest that $\hat{\beta}_n$ is not converging to a unique value. On the other hand, they show that the vectors $\hat{\beta}_n$'s are concentrated around multiple modes similar to the phenomena described in Theorem 6. Note that in each scatterplot, one of the clusters can be represented by the projection of $\beta_{0,\mathrm{OLS}}$ and therefore can be recognized as the "true cluster". As in reality we do not have access to the whole matched data set, one can use $\tilde{\beta}_m$

---

1. https://archive.ics.uci.edu/
2. We use function *density* from R package "stats" with hyper-parameter "SJ" for the bandwidth and Gaussian kernel.

as a "guide" to pin down the right cluster. We intend to formalize this idea more concretely in the scope of future work.
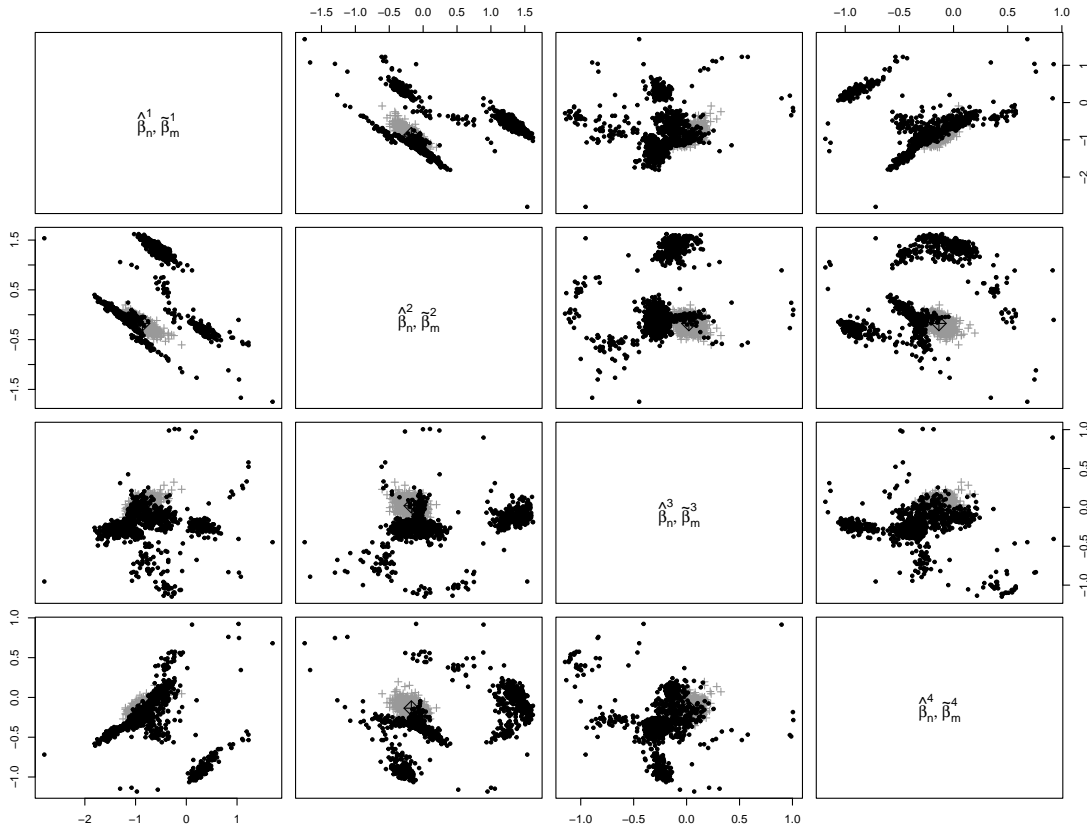

Figure 8: Scatterplots of pairs of components of $\hat{\beta}_n$ (filled black point) and $\tilde{\beta}_m$ (gray plus) together with projections of $\beta_{0,\mathrm{OLS}}$ (black diamond). Note that in each plot, one of the clusters of $\hat{\beta}_n$ is concentrated near the projection of $\beta_{0,\mathrm{OLS}}$. In a real-world scenario, we do not have access to $\beta_{0,\mathrm{OLS}}$ to choose the corresponding sub-clusters of $\hat{\beta}_n$. For this purpose one can use $\tilde{\beta}_m$ instead.

5. Discussion and Future Research Directions

In this paper, we have proposed an approach for making inference in a regression model in the case where the link between responses and covariates is not known. Here, our main goal is to estimate the regression vector. Note that the model we consider includes the permuted regression model studied by in Pananjady et al. (2018), where the principal focus is to recover the permutation under which the covariates are shuffled and which is responsible for the loss of the link. The main idea that we pursued in this work is to view the responses as random variables generated from the convolution of $\beta_0^T X$ and ϵ and find the vector which minimizes over all possible $\beta \in \mathbb{R}^d$ an ℓ_2-distance between the sample distribution of the

16

responses and a natural estimator of the distribution of the convolution of $\beta^T X$ and ϵ. As it becomes clear at the end of Section 3, the problem boils down to a deconvolution. It may come as very surprising that one is able to recover, under some identifiability conditions, the true regression vector or an ordered version thereof at the $n^{-1/2}$-rate. In fact, the convergence rates in deconvolution problems are known to be slow to very slow for smooth noise distributions. For example, if the noise is Gaussian and under minimal smoothness assumptions, then it follows from the seminal paper of Fan (1991) that $(\log n)^{-1/2}$ is the optimal rate of convergence for deconvoluting the distribution of X from the noise ϵ based on observations from the model $Y = X + \epsilon$. How do we reconcile this slow rate with the obtained rate $n^{-1/2}$ in our Theorem 7? The answer is that our problem is completely parametric, and the distribution to be deconvoluted is in a much smaller class than the one considered in Fan (1991).

The estimation method that is chosen and implemented in this work is partially motivated by convenience to some degree. In fact, the arguments from the theory of empirical process, although technical, seem to be less cumbersome as they are, for example, for the Maximum Likelihood estimator (MLE). Finding the MLE in this problem would have been a very appealing approach, and the estimator might be even more efficient. In this case, one would find a regression vector which maximizes the log-likelihood

$$\beta \mapsto \sum_{j=1}^n \log \left(n^{-1} \sum_{i=1}^n f^\epsilon(Y_j - \beta^T X_i) \right)$$

over $\mathbb{R}^d$. Optimization, in this case, presents harder numerical challenges because of the logarithm. Nevertheless, we believe that this estimator should be implemented. We leave this task to future research work, where we also plan to study the asymptotic properties of this estimator and compare its performance to the DLSE studied here. If the noise distribution is assumed to be known, we would like to stress the fact that it is possible to relax such an assumption. This can be done either by (1) estimating this distribution from a matched sample as done above for the Power Plant data set or (2) assuming that it belongs to some scaled parametric family and estimating the scale parameter, σ_0 say, together with the unknown regression vector. This approach was followed for the inter-generational mobility data set. We conjecture that under identifiability, the resulting estimators of β_0 and σ_0 are asymptotically Gaussian. A referee noted that it is possible to consider the case where the response Y is multivariate, of dimension p, in which case one needs to estimate a whole regression matrix. If the components of Y are independent, then the asymptotic theory developed here can be easily extended. However, the optimization problem, already difficult for $p = 1$, will impose additional numerical challenges. In case the components are not independent, we do not think that extending our estimator to this case would be straightforward. Computation of the empirical distribution of the observed responses becomes very tedious as well as the criterion, which now involves multivariate integration.

The great potential of the method described in this work was clearly seen in the semi-supervised learning situation, where some matched data is available. We view this scenario as the best and most realistic application for two reasons: (1) One is able to estimate the distribution of the noise from the matched proportion and not simply impose it. (2) The OLS found with the matched data can be used in combination with re-sampling from the

unmatched data to "guide" the deconvolution least square method in finding the closest DLSE to the OLS. The point in (2) is particularly useful in case the re-sampled data do not seem to agree on the DLSE.

Finally, we would like to point out that the linear regression model can be, of course, extended to other models, such as logistic regression for example. An extension is the high-dimensional setting, although we believe that many theoretical challenges will have to be tackled, both from theoretical and numerical perspectives.

Acknowledgments

Appendix A. Supplementary Material

We present proofs of all the theorems and lemmas in the following.

A.1 Proof of Proposition 1

Proof Our goal is to show that with probability one, there exists a vector $\beta \in \mathbb{R}^d$ such that it minimizes $\mathbb{D}_{n,p}$.

For $K > 0$ consider the compact set $\{\beta \in \mathbb{R}^d : \|\beta\| \leq K\}$. We show that our choice of K depends only on p. Let $u \in S^{d-1}$ be a fixed vector. Then we have

$$\lim_{\lambda \to \infty} \mathbb{D}_{n,p}(\lambda u) = \frac{1}{n^{p+1}} \sum_{j=1}^{n} \left| j - n_0 F^\epsilon(Y_{(j)}) - n_- \right|^p$$

with $n_- = |\{i : u^T X_i < 0\}|$ and $n_0 = |\{i : u^T X_i = 0\}|$. Note that n_- and n_0 depend on the vector u. Since X has a continuous distribution, $n_0 \leq d$ with probability one. If $n_0 \geq d+1$, then for some distinct $i_1, \cdots, i_{d+1}$, $X_{i_1}, \ldots, X_{i_{d+1}}$ are not linearly independent, which is of probability zero.

Suppose that $n_- \geq n/2$. Using $||a| - |b|| \leq |a - b|$, it follows that

$$\sum_{j=1}^{n} \left| j - n_0 F^\epsilon(Y_{(j)}) - n_- \right|^p \geq \sum_{j \leq n/4} \left| |j - n_-| - n_0 F^\epsilon(Y_{(j)}) \right|^p.$$

Since $j \leq n/4$ we have $|j - n_-| \geq n/4$. By taking $n > 8d$, and following from $n_0 F^\epsilon(Y_{(j)}) \leq n_0 \leq d$, with probability one we have

$$
\begin{aligned}
\sum_{j=1}^{n} \left| j - n_0 F^\epsilon(Y_{(j)}) - n_- \right|^p &\geq \sum_{j \leq n/4} \left(|j - n_-| - d \right)^p \\
&\geq \frac{n}{4} \left(\frac{n}{4} - d \right)^p \\
&\geq \frac{n}{4} \left(\frac{n}{8} \right)^p = \frac{n^{p+1}}{2^{3p+2}} > \frac{n^{p+1}}{8^{p+1}}.
\end{aligned}
$$

This implies that for $n \geq 8d$

$$\frac{1}{n^{p+1}} \sum_{j=1}^{n} \left| j - n_0 F^\epsilon(Y_{(j)}) - n_- \right|^p \geq \frac{1}{8^{p+1}}.$$

Now, for any small $\eta > 0$ and large enough λ'

$$\mathbb{D}_{n,p}(\lambda' u) \geq \lim_{\lambda \to \infty} \mathbb{D}_{n,p}(\lambda u) - \eta.$$

Let $\eta = 2^{-(3p+4)}$. Then, there exists $K_p > 0$ such that for $\lambda > K_p$

$$\mathbb{D}_{n,p}(\lambda u) \geq \frac{1}{16 \cdot 8^p} \tag{A.1}$$

19

with probability 1 provided that $n \geq 8d$. Note that for $n_- < n/2$, similar reasoning can be applied by summing over $j \geq n/4$. Now, consider the set

$$\mathcal{C}_p = \left\{ \beta \in \mathbb{R}^d : \|\beta\| \leq K_p \right\}. \tag{A.2}$$

where K_p is the same constant defined above. Since the lower bound shown in (A.1) does not depend on the vector u, we conclude that for all $n \geq 8d$

$$\inf_{\beta \in \mathcal{C}_p^c} \mathbb{D}_{n,p}(\beta) \geq \frac{1}{16 \cdot 8^p}. \tag{A.3}$$

On the other hand, and by using the equality $F^Y = F^{\beta_0^T X} \star F^\epsilon$,

$$\mathbb{D}_{n,p}(\beta_0) = \int \left| \left(F_n^Y(y) - F^Y(y) \right) - \left(n^{-1} \sum_{i=1}^n F^\epsilon(y - \beta_0^T X_i) - [F^{\beta_0^T X} \star F^\epsilon](y) \right) \right|^p dF_n(y)$$

$$\leq 2^{p-1} \int \left| F_n^Y(y) - F^Y(y) \right|^p dF_n(y)$$

$$+ 2^{p-1} \int \left| n^{-1} \sum_{i=1}^n F^\epsilon(y - \beta_0^T X_i) - [F^{\beta_0^T X} \star F^\epsilon](y) \right|^p dF_n(y).$$

Convexity of $x \mapsto |x|^p$ gives us

$$\mathbb{D}_{n,p}(\beta_0) \leq 2^{p-1} \|F_n^Y - F^Y\|_\infty + 2^{p-1} \left\| F^{\beta_0^T X} \star F^\epsilon - F_n^{\beta_0^T X} \star F^\epsilon \right\|_\infty$$

$$\leq 2^{p-1} \|F_n^Y - F^Y\|_\infty + 2^{p-1} \left\| F^{\beta_0^T X} - F_n^{\beta_0^T X} \right\|_\infty.$$

where $\|\cdot\|_\infty$ denotes the supremum norm, and $F_n^{\beta_0^T X}$ is the empirical distribution based on $\beta_0^T X_i, i = 1, \cdots, n$. By Glivenko-Cantelli we have

$$\mathbb{P}\left(\omega : \exists\, n_p(\omega)\ \forall\, n \geq n_p(\omega)\ \mathbb{D}_{n,p}(\beta_0) \leq \frac{1}{32 \cdot 8^p} \right) = 1. \tag{A.4}$$

Therefore by combining (A.3) and (A.4) we have

$$\mathbb{P}\Big(\omega : \exists\, n_p(\omega)\ \forall\, n \geq \max(n_p(\omega), 8d)\ \ \mathbb{D}_{n,p} \text{ admits a minimizer in } \mathcal{C}_p^c \Big)$$

$$= \mathbb{P}\Big(\omega : \exists\, n_p(\omega)\ \forall\, n \geq \max(n_p(\omega), 8d)\ \ \exists \beta \in \mathcal{C}_p^c : \mathbb{D}_{n,p}(\beta_0) \geq \mathbb{D}_{n,p}(\beta) \geq \frac{1}{16 \cdot 8^p} \Big)$$

$$= 0.$$

Since $\mathcal{C}_p$ defined in (A.2) is compact and $\mathbb{D}_{n,p}$ is Lebesgue-a.e continuous we have

$$\mathbb{P}\Big(\omega : \exists\, n_p(\omega)\ \forall\, n \geq \max(n_p(\omega), 8d)\ \ \mathbb{D}_{n,p} \text{ admits a minimizer in } \mathbb{R}^p \Big)$$

$$= \mathbb{P}\Big(\omega : \exists\, n_p(\omega)\ \forall\, n \geq \max(n_p(\omega), 8d)\ \ \mathbb{D}_{n,p} \text{ admits a minimizer in } \mathcal{C}_p \Big)$$

$$= 1.$$

■

A.2 A Useful Proposition and Its Proof

We will use the following proposition in the rest of the proofs.

Proposition 10 *For any $\hat{\beta}_n \in \mathcal{B}_{n,2}$ we have*

$$\int_{\mathbb{R}} \left(F_n^Y(y) - n^{-1} \sum_{i=1}^n F^\epsilon(y - \hat{\beta}_n^T X_i) \right)^2 dF_n^Y(y) = O_{\mathbb{P}}(n^{-1}).$$

Proof By definition of $\hat{\beta}_n$ we have

$$\int_{\mathbb{R}} \left(F_n^Y(y) - n^{-1} \sum_{i=1}^n F^\epsilon(y - \hat{\beta}_n^T X_i) \right)^2 dF_n^Y(y) \leq$$
$$\int_{\mathbb{R}} \left(F_n^Y(y) - n^{-1} \sum_{i=1}^n F^\epsilon(y - \beta_0^T X_i) \right)^2 dF_n^Y(y)$$

By adding and subtracting $F^Y(y)$ we have

$$\begin{aligned}
\mathbb{D}_{n,2}(\beta_0) &= \int_{\mathbb{R}} \left(F_n^Y(y) - n^{-1} \sum_{i=1}^n F^\epsilon(y - \beta_0^T X_i) \right)^2 dF_n^Y(y) \\
&\leq 2 \int_{\mathbb{R}} \left(F_n^Y(y) - F^Y(y) \right)^2 dF_n^Y(y) + \\
&\quad 2 \int \left(F^Y(y) - n^{-1} \sum_{i=1}^n F^\epsilon(y - \beta_0^T X_i) \right)^2 dF_n^Y(y).
\end{aligned}$$

By Dvoretzky–Kiefer–Wolfowitz inequality; Dvoretzky et al. (1956) we have

$$\int_{\mathbb{R}} \left(F_n^Y(y) - F^Y(y) \right)^2 dF_n^Y(y) \leq \| F_n^Y - F^Y \|_\infty^2 = O_{\mathbb{P}}(n^{-1}).$$

Using $F^Y(y) = \int F^\epsilon(y - \beta_0^T x) dF^X(x)$ it follows that

$$\begin{aligned}
&\int \left(F^Y(y) - n^{-1} \sum_{i=1}^n F^\epsilon(y - \beta_0^T X_i) \right)^2 dF_n^Y(y) \\
&= \int \left(\int F^\epsilon(y - \beta_0^T x) dF^X(x) - \int F^\epsilon(y - \beta_0^T x) dF_n^X(x) \right)^2 dF_n^Y(y) \\
&= \int \left(\int F^\epsilon(y - z) dF^Z(z) - \int F^\epsilon(y - z) dF_n^Z(z) \right)^2 dF_n^Y(y), \quad \text{with } Z = \beta_0^T X \\
&= \int \left(\int \left(F^Z(z) - F_n^Z(z) \right) f^\epsilon(y - z) dz \right)^2 dF_n^Y(y) \\
&\leq \| F_n^Z - F^Z \|_\infty^2 = O_{\mathbb{P}}(n^{-1}).
\end{aligned}$$

Therefore $\mathbb{D}_{n,2}(\beta_0) = O_{\mathbb{P}}(n^{-1})$. Since by definition $\mathbb{D}_{n,2}(\hat{\beta}_n) \leq \mathbb{D}_{n,2}(\beta_0)$ we conclude that

$$\mathbb{D}_{n,2}(\hat{\beta}_n) = \int_{\mathbb{R}} \left(F_n^Y(y) - n^{-1} \sum_{i=1}^n F^\epsilon(y - \hat{\beta}_n^T X_i) \right)^2 dF_n^Y(y) = O_{\mathbb{P}}(n^{-1}) \qquad (A.5)$$

which implies that

$$\int_{\mathbb{R}} \left(F^Y(y) - n^{-1} \sum_{i=1}^{n} F^\epsilon(y - \hat{\beta}_n^T X_i) \right)^2 dF_n^Y(y)$$

$$\leq 2 \int_{\mathbb{R}} \left(F^Y(y) - F_n^Y(y) \right)^2 dF_n^Y(y) +$$

$$2 \int_{\mathbb{R}} \left(F_n^Y(y) - n^{-1} \sum_{i=1}^{n} F^\epsilon(y - \hat{\beta}_n^T X_i) \right)^2 dF_n^Y(y)$$

$$= O_{\mathbb{P}}(n^{-1})$$

which completes the proof. ∎

A.3 Proof of Theorem 3

Proof For positive semi-definite matrix Γ and $z \in \mathbb{R}^2$, define $\|z\|_{2,\Gamma} := \sqrt{z^T \Gamma z}$. Note that $\|\hat{\beta}\|_2 = O_{\mathbb{P}}(1)$ is equivalent to $\|\hat{\beta}_n\|_{2,\Gamma} = O_{\mathbb{P}}(1)$. It follows from (van der Vaart, 1998, Theorem 5.14) that for any compact set $\mathcal{K} \subset \mathbb{R}^d$ and $\epsilon > 0$

$$\lim_{n \to \infty} \mathbb{P} \left(\inf_{\beta \in \mathcal{B}_0} \|\hat{\beta}_n - \beta\|_{2,\Gamma} \geq \epsilon, \hat{\beta}_n \in \mathcal{K} \right) = 0.$$

Assuming that $\|\hat{\beta}_n\|_{2,\Gamma} = O_{\mathbb{P}}(1)$, for any $\eta \in (0,1)$ we can find $K_\eta > 0$ such that

$$\mathbb{P}(\|\hat{\beta}_n\|_{2,\Gamma} \leq K_\eta) \geq 1 - \eta$$

for all n. Take $\mathcal{K}$ to be the closed Euclidean ball in $\mathbb{R}^d$ with centre 0 and radius K_η. Then,

$$\mathbb{P} \left(\inf_{\beta \in \mathcal{B}_0} \|\hat{\beta}_n - \beta\|_{2,\Gamma} \geq \epsilon \right) = \mathbb{P} \left(\inf_{\beta \in \mathcal{B}_0} \|\hat{\beta}_n - \beta\|_{2,\Gamma} \geq \epsilon, \hat{\beta}_n \in \mathcal{K} \right) +$$

$$\mathbb{P} \left(\inf_{\beta \in \mathcal{B}_0} \|\hat{\beta}_n - \beta\|_{2,\Gamma} \geq \epsilon, \hat{\beta}_n \notin \mathcal{K} \right)$$

$$\leq \mathbb{P} \left(\inf_{\beta \in \mathcal{B}_0} \|\hat{\beta}_n - \beta\|_{2,\Gamma} \geq \epsilon, \hat{\beta}_n \in \mathcal{K} \right) + \eta.$$

Hence

$$\lim_{n \to \infty} \mathbb{P} \left(\inf_{\beta \in \mathcal{B}_0} \|\hat{\beta}_n - \beta\|_{2,\Gamma} \geq \epsilon \right) \leq \eta.$$

Since $\eta > 0$ can be arbitrarily small, this implies

$$\lim_{n \to \infty} \mathbb{P} \left(\inf_{\beta \in \mathcal{B}_0} \|\hat{\beta}_n - \beta\|_{2,\Gamma} \geq \epsilon \right) = 0.$$

By the assumption of Theorem for any $\beta \in \mathcal{B}_0$ we have $\|\beta\|_{2,\Gamma} = c$. This, together with triangle inequality, implies

$$\|\|\hat{\beta}_n\|_{2,\Gamma} - \|\beta\|_{2,\Gamma}\| = \|\|\hat{\beta}_n\|_{2,\Gamma} - c\| \leq \|\hat{\beta}_n - \beta\|_{2,\Gamma},$$

hence

$$\|\hat{\beta}_n\|_{2,\Gamma} = \sqrt{\hat{\beta}_n^T \Gamma \hat{\beta}_n} \xrightarrow{\mathbb{P}} c$$

as $n \to \infty$. This completes the proof. ∎

A.4 Proof of Lemma 2

Proof For a random variable Z denote the characteristic function of Z by Φ_Z; i.e.,

$$\Phi_Z(t) = \mathbb{E}[\exp(itZ)], t \in \mathbb{R}.$$

Then, $\Phi_Y(t) = \Phi_{\beta^T X}(t)\Phi_\epsilon(t)$ for all $t \in \mathbb{R}$. Let $\mathcal{V} \subset \mathbb{R}$ be a small interval containing 0 such that for all $t \in \mathcal{V}$ we have $\Phi_\epsilon(t) \neq 0$. Then $\beta_1, \beta_2 \in \mathcal{B}_0$ and β_2 if and only if $\Phi_{\beta_1^T X}(t) = \Phi_{\beta_2^T X}(t)$ for all $t \in \mathcal{V}$. Since $\beta^T X \sim \mathcal{N}(0, \beta^T \Sigma \beta)$ we have

$$\exp\left(-\frac{t^2}{2}\beta_1^T \Sigma \beta_2\right) = \exp\left(-\frac{t^2}{2}\beta_2^T \Sigma \beta_2\right).$$

Therefore $\beta_1^T \Sigma \beta_1 = \beta_2^T \Sigma \beta_2$. Note that Σ is positive definite which implies $c = \beta^T \Sigma \beta > 0$ for $\beta \in \mathcal{B}_0$. ∎

A.5 Proof of Proposition 4

Proof Let us start with the case where $\Sigma = I_d$. Using (A.9) we have

$$\lim_{n \to \infty} \mathbb{P}\left(\forall \beta \in \mathbb{R}^d, \text{ and } j = 1,..,n, \frac{1}{n}\sum_{i=1}^n F^\epsilon(Y_j - \beta^T X_i) \geq \int F^\epsilon(Y_j - \beta^T x)dF^X(x) - \frac{1}{16}\right) = 1$$

where

$$\int F^\epsilon(Y_j - \beta^T x)dF^X(x) = \int F^\epsilon(Y_j - \|\beta\|_2 z)\varphi(z)dz, \quad \text{where } \varphi \text{ is the density of } \mathcal{N}(0,1).$$

This implies that

$$\lim_{n \to \infty} \mathbb{P}\left(\forall \beta \in \mathbb{R}^d, \text{ and } j = \lfloor n/5 \rfloor, \cdots, \lfloor n/4 \rfloor, \frac{1}{n}\sum_{i=1}^n F^\epsilon(Y_{(j)} - \beta^T X_i) \geq \int F^\epsilon(Y_{(j)} - \beta^T x)dF^X(x) - \frac{1}{16}\right) = 1.$$

For any fixed j and $\delta > 0$

$$\begin{aligned}
\int F^\epsilon(Y_{(j)} - \beta^T x)dF^X(x) &= \int_{z:\ z<0} + F^\epsilon(Y_{(j)} - \|\beta\|_2 z)\varphi(z)dz \\
&\quad \int_{z:\ z>0} F^\epsilon(Y_{(j)} - \|\beta\|_2 z)\varphi(z)dz \\
&\geq \int_{z:\ z<0} F^\epsilon(Y_{(j)} - \|\beta\|_2 z)\varphi(z)dz \\
&\geq \int_\delta^\infty F^\epsilon(Y_{(j)} + \|\beta\|_2 z)\varphi(z)dz.
\end{aligned}$$

Fix a small $\eta \in (0, 1)$. Since F^ϵ is a cumulative distribution function, there exists $M_\eta > 0$ such that $F^\epsilon(t) \geq 1 - \eta$ if $t > M_\eta$. This implies that for $\eta = 1/12$ and δ such $\int_\delta^\infty \varphi(z)dz \geq 3/8$

$$\|\beta\|_2 > \max((M_\eta - Y_{(j)})/\delta, 0) \implies \int_\delta^\infty F^\epsilon(Y_{(j)} + \|\beta\|_2 z)\varphi(z)dz \geq 3(1-\eta)/8 = \frac{11}{32}.$$

Let $M \geq \max\{M_\eta, [F^Y]^{(-1)}(1/4)+2\}$. Since $Y_{(n/4)}$ converges almost surely to $[F^Y]^{(-1)}(1/4)$, we have that

$$\lim_{n\to\infty} \mathbb{P}\Big(\frac{1}{n}\sum_{i=1}^n F^\epsilon(Y_{(j)} - \beta^T X_i) \geq \frac{11}{32} \text{ and } j = \lfloor n/5 \rfloor, \cdots, \lfloor n/4 \rfloor \text{ and } \|\beta\|_2 \geq \frac{M - [F^Y]^{(-1)}(1/4) + 1}{\delta}\Big) = 1.$$

Now, note that for $\beta \in \mathbb{R}^d$ such that $\int F^\epsilon(Y_{(j)} - \beta^T x)dF^X(x) \geq 11/32$ we have

$$
\begin{aligned}
\mathbb{D}_{n,2}(\beta) &\geq \frac{1}{n}\sum_{j=\lfloor n/5 \rfloor}^{\lfloor n/4 \rfloor}\left(F_n(Y_{(j)}) - n^{-1}\sum_{i=1}^n F^\epsilon(Y_j - \beta^T X_i)\right)^2 \\
&\geq \frac{1}{n}\left(\lfloor n/4 \rfloor - \lfloor n/5 \rfloor + 1\right)\left(11/32 - 1/4\right)^2 \\
&\geq \frac{1}{20}\frac{9}{32^2} := c.
\end{aligned}
$$

Therefore

$$\lim_{n\to\infty} \mathbb{P}\left(\inf_{\|\beta\|_2 \geq \frac{M-[F^Y]^{(-1)}(1/4)+1}{\delta}} \mathbb{D}_{n,2}(\beta) \geq c\right) = 1.$$

Now,

$$
\begin{aligned}
&\mathbb{P}\left(\|\hat{\beta}_n\|_2 \geq \frac{M - [F^Y]^{(-1)}(1/4) + 1}{\delta}\right) = \\
&\mathbb{P}\left(\|\hat{\beta}_n\|_2 \geq \frac{M - [F^Y]^{(-1)}(1/4) + 1}{\delta}, \mathbb{D}_{n,2}(\hat{\beta}_n) \geq c\right) + \\
&\mathbb{P}\left(\|\hat{\beta}_n\|_2 \geq \frac{M - [F^Y]^{(-1)}(1/4) + 1}{\delta}, \mathbb{D}_{n,2}(\hat{\beta}_n) < c\right) \\
&= \mathbb{P}_{n,1} + \mathbb{P}_{n,2}.
\end{aligned}
$$

Note that $\lim_{n\to\infty}\mathbb{P}_{n,2} = 0$ since the event $\{\|\hat{\beta}_n\|_2 \geq \frac{M-[F^Y]^{(-1)}(1/4)+1}{\delta}, \mathbb{D}_{n,2}(\hat{\beta}_n) < c\}$ is included in the complement of the event

$$\left\{\inf_{\|\beta\|_2 \geq \frac{M-[F^Y]^{(-1)}(1/4)+1}{\delta}} \mathbb{D}_{n,2}(\beta) \geq c\right\}.$$

On the other hand, (A.5) implies that $\mathbb{D}_{n,2}(\hat{\beta}_n) = O_{\mathbb{P}}(n^{-1})$. Since $\mathbb{P}_{n,1} \leq \mathbb{P}(\mathbb{D}_{n,2}(\hat{\beta}_n) \geq c)$

$$\lim_{n\to\infty}\mathbb{P}_{n,1} = 0.$$

Thus

$$\lim_{n \to \infty} \mathbb{P}\left(\|\hat{\beta}_n\|_2 \geq \frac{M - [F^Y]^{(-1)}(1/4) + 1}{\delta}\right) = 0.$$

this implies that $\hat{\beta}_n = O_{\mathbb{P}}(1)$. Now, consider the general case where $X \sim \mathcal{N}(\mu, \Sigma)$ with Σ positive definite. Then, the model in (2.1) can be written as

$$Y \stackrel{d}{=} (\Sigma^{1/2}\beta_0)^T \Sigma^{-1/2} X + \epsilon = \gamma_0^T \tilde{X} + \epsilon,$$

where $\tilde{X} \sim \mathcal{N}(0, I_d)$, independent of ϵ, and $\gamma_0 = \Sigma^{1/2}\beta_0$. Thus, as $\hat{\beta}_n = \Sigma^{1/2}\hat{\gamma}_n$ and $\|\hat{\gamma}_n\|_2 = O_{\mathbb{P}}(1)$, with $\hat{\gamma}_n$ the least squares estimator based on $Y_1, \ldots, Y_n$ and $\tilde{X}_1, \ldots, \tilde{X}_n$, it follows that $\|\hat{\beta}_n\|_2 = O_{\mathbb{P}}(1)$. By Lemma 2 and Proposition 3, we have

$$\sqrt{\hat{\beta}_n^T \Sigma \hat{\beta}_n} \stackrel{\mathbb{P}}{\to} c$$

where $c = \sqrt{\beta^T \Sigma \beta} > 0$ for all $\beta \in \mathcal{B}_0$ for this model. For $\widehat{\Sigma}_n$ a consistent estimator of Σ, we have

$$\hat{\beta}_n^T \widehat{\Sigma}_n \hat{\beta}_n^T = \hat{\beta}_n^T \Sigma \hat{\beta}_n^T + \hat{\beta}_n^T \left(\widehat{\Sigma}_n - \Sigma\right) \hat{\beta}_n^T$$

where the second term on the right is bounded by $\max_{1 \leq i,j \leq d} |\widehat{\Sigma}_{n,i,j} - \Sigma_{i,j}| \times \|\hat{\beta}_n\|^2 \stackrel{\mathbb{P}}{\to} 0$ since $\|\hat{\beta}_n\| = O_{\mathbb{P}}(1)$. This concludes the proof. ∎

A.6 Proof of Theorem 5

Proof Recall that

$$\mathcal{D}_{2,F^Y}(\beta) = \int \left(F^Y(y)dy - \int F^\epsilon(y - \beta^T x)dF^X(x)\right)^2 dF^Y(y).$$

For $\beta \in \mathbb{R}^d$ define

$$C_\beta(y) := \int F^\epsilon(y - \beta^T x)dF^X(x), \quad \text{and} \quad C_{n,\beta}(y) = \int F^\epsilon(y - \beta^T x)dF_n^X(x)$$

the convolution density and its empirical estimator. Then, we have that

$$
\begin{aligned}
|\mathbb{D}_{n,2}(\beta) - \mathcal{D}_{2,F^Y}(\beta)| &= \left|\int \left(F_n^Y(y) - C_{n,\beta}(y)\right)^2 dF_n^Y(y) - \int \left(F^Y(y) - C_\beta(y)\right)^2 dF^Y(y)\right| \\
&\leq \left|\int \left(F_n^Y(y) - C_{n,\beta}(y)\right)^2 d(F_n^Y(y) - F^Y(y))\right| \\
&\quad + \left|\int \left\{\left(F_n^Y(y) - C_{n,\beta}(y)\right)^2 - \left(F^Y(y) - C_\beta(y)\right)^2\right\} dF^Y(y)\right| \\
&= A_n + B_n.
\end{aligned}
$$

Let us consider A_n.

$$0 \leq A_n \leq \left| \int F_n^Y(y)^2 d(F_n^Y(y) - F^Y(y)) \right| + \left| \int C_{n,\beta}^2(y) d(F_n^Y(y) - F^Y(y)) \right|$$
$$+ 2 \left| \int F_n^Y(y) C_{n,\beta}(y) d(F_n^Y(y) - F^Y(y)) \right|.$$

Functions $y \mapsto F_n^Y(y)^2$, $y \mapsto C_{n,\beta}^2(y)$, and $y \mapsto F_n^Y(y) C_{n,\beta}(y)$ are all non-negative, monotone and bounded above by 1. Denote the class of real monotone functions f such that $\mathrm{Im}(f) \subseteq [0,1]$ by $\mathcal{M}$. We have

$$0 \leq A_n \leq 4 \sup_{f \in \mathcal{M}} \left| \int f d(F_n^Y(y) - F^Y(y)) \right|$$
$$= 4 n^{-1/2} \|\mathbb{G}_n\|_{\mathcal{M}}$$

where $\mathbb{G}_n^Y = \sqrt{n}(\mathbb{P}_n^Y - \mathbb{P}^Y)$. By Theorem 2.7.5 of van der Vaart and Wellner (1996) there exists a universal constant $K > 0$ such that for all $\eta > 0$

$$\log N_B(\eta, \mathcal{M}, L_2(\mathbb{P}^Y)) \leq \frac{K}{\eta}$$

where N_B denotes the bracketing covering number. Now, by Lemma 3.4.2 of van der Vaart and Wellner (1996) and using the fact that all functions in $\mathcal{M}$ are bounded above by 1, we have that

$$E[\|\mathbb{G}_n^Y\|_{\mathcal{M}}] \lesssim J(1) \left(1 + \frac{J(1)}{\sqrt{n}} \right),$$

where for small $\delta > 0$

$$J(\delta) = \int_0^\delta \sqrt{1 + \log N_B(\eta, \mathcal{M}, L_2(\mathbb{P}^Y))} dt$$
$$\leq \delta + \sqrt{K} \int_0^\delta \frac{1}{\sqrt{\eta}} dt$$
$$= \delta + 2\sqrt{K}\sqrt{\delta}.$$

It follows that $\mathbb{E}[\|\mathbb{G}_n^Y\|_{\mathcal{M}}] \lesssim 1$ for all $n \geq 1$. Using Markov's inequality, we conclude that

$$A_n = O_{\mathbb{P}}(n^{-1/2}). \tag{A.6}$$

Now, we focus on the term B_n. We have

$$\left(F_n^Y(y) - C_{n,\beta}(y) \right)^2 = \left(F_n^Y(y) - F^Y(y) + F^Y(y) - C_\beta(y) + C_\beta(y) - C_{n,\beta}(y) \right)^2$$
$$= \left(F_n^Y(y) - F^Y(y) \right)^2 + \left(F^Y(y) - C_\beta(y) \right)^2 + (C_\beta(y) - C_{n,\beta}(y))^2$$
$$+ 2 \left(F_n^Y(y) - F^Y(y) \right) \left(F^Y(y) - C_\beta(y) \right)$$
$$+ 2 \left(F_n^Y(y) - F^Y(y) \right) \left(C_\beta(y) - C_{n,\beta}(y) \right)$$
$$+ 2 \left(F^Y(y) - C_\beta(y) \right) \left(C_\beta(y) - C_{n,\beta}(y) \right).$$

Hence

$$0 \leq B_n \leq \int \left(F_n^Y(y) - F^Y(y) \right)^2 dF^Y(y) + \int \left(C_\beta(y) - C_{n,\beta}(y) \right)^2 dF^Y(y)$$

$$+ \ 2 \int \left| F_n^Y(y) - F^Y(y) \right| \times \left| F^Y(y) - C_\beta(y) \right| dF^Y(y)$$

$$+ \ 2 \int \left| F_n^Y(y) - F^Y(y) \right| \times \left| C_\beta(y) - C_{n,\beta}(y) \right| dF^Y(y)$$

$$+ \ 2 \int \left| F^Y(y) - C_\beta(y) \right| \times \left| C_\beta(y) - C_{n,\beta}(y) \right| dF^Y(y)$$

$$= I_n + II_n(\beta) + III_n(\beta) + IV_n(\beta) + V_n(\beta),$$

Note that

$$\mathbb{E}(I_n) = n^{-1} \int F^Y(y)(1 - F^Y(y))dF^Y(y) \leq n^{-1}$$

implying that $I_n = O_\mathbb{P}(n^{-1})$. Also, for all $\beta \in \mathbb{R}^d$

$$II_n(\beta) \ \leq \ \int \sup_{\gamma \in \mathbb{R}^d} |C_{n,\gamma}(y) - C_\gamma(y))| \, dF^Y(y),$$

where

$$|C_{n,\gamma}(y) - C_\gamma(y)| \ = \ |\int F^\epsilon(y - \gamma^T x)d(F_n^X(x) - F^X(x))|$$

$$= \ |\int F^\epsilon(y - z)d(F_n^Z(z) - F^Z(z))|$$

$$= \ |\int (F_n^Z(z) - F^Z(z))f^\epsilon(y - z)dz|.$$

using the change of variable $Z = \gamma^T X$, and integration by parts. Thus, it follows that

$$\sup_{y \in \mathbb{R}} |C_{n,\gamma}(y) - C_\gamma(y)| \ \leq \ \sup_{z \in \mathbb{R}} |F_n^Z(z) - F^Z(z)|$$

$$= \ \sup_z \left| \int \mathbb{1}\{\gamma^T x \leq z\}d(F_n^X(x) - F^X(x)) \right|. \qquad \text{(A.7)}$$

Next, we show that the supremum in (A.7) is $O_\mathbb{P}(n^{-1/2})$ independently of γ. Consider the class

$$\mathcal{F} = \{x \mapsto \gamma^T x - z, \gamma \in \mathbb{R}^d, z \in \mathbb{R}\}.$$

Note that $\mathcal{F}$ is a $d+1$ dimensional vector space. By Lemma 2.6.15 of van der Vaart and Wellner (1996), $\mathcal{F}$ is a VC-subgraph of index smaller than $d+1+2 = d+3$. Now, note that

$$\int \mathbb{1}\{\gamma^T x \leq z\}d(F_n^X(x) - F^X(x)) \ = \ \int \mathbb{1}\{\gamma^T x - z \leq 0\}d(F_n^X(x) - F^X(x))$$

$$= \ \int (\phi \circ f)(x)d(F_n^X(x) - F^X(x))$$

with $\phi(t) = \mathbb{1}\{t \le 0\}$ and $f(x) = \gamma^T x - z \in \mathcal{F}$. Since ϕ is monotone, by Lemma 2.6.18(viii) of van der Vaart and Wellner (1996), we conclude that the class $\phi \circ \mathcal{F}$ is a VC-subgraph. Since all elements in $\phi \circ \mathcal{F}$ are bounded by 1, the latter can be taken as its envelope. Thus, it follows from Theorem 2.6.7 in van der Vaart and Wellner (1996) that there exists $V \ge 2$ such that (taking $r = 2$) for all $\eta \in (0,1]$ and all probability measures $\mathbb{Q}$

$$N(\eta, \phi \circ \mathcal{F}, L_2(\mathbb{Q})) \le KV(162)^V \frac{1}{\eta^{2(V-1)}}$$

where $K > 0$ is some universal constant. Hence, there exists a constant $c \in \mathbb{R}$ such that

$$
\begin{aligned}
\sup_{\mathbb{Q}} \int_0^1 \sqrt{1 + \log N(\eta, \phi \circ \mathcal{F}, L_2(\mathbb{Q}))} d\eta &\le c + \sqrt{2(V-1)} \int_0^1 \sqrt{\log\left(\frac{1}{\eta}\right)} d\eta \\
&= c + \sqrt{2(V-1)} \int_1^\infty \frac{\sqrt{\log t}}{t^2} dt < \infty,
\end{aligned}
$$

using the fact that

$$\int_1^\infty \frac{\sqrt{\log(t)}}{t^2} dt \le \int_1^\infty \frac{1}{t^{3/2}} dx < \infty \tag{A.8}$$

using $\log(t) \le t, t \ge 1$. It follows from Theorem 2.14.1 in van der Vaart and Wellner (1996) that

$$\mathbb{E}[\|\mathbb{G}_n^X\|_{\phi \circ \mathcal{F}}] \lesssim 1,$$

where

$$\|\mathbb{G}_n^X\|_{\phi \circ \mathcal{F}} = \sup_{g \in \phi \circ \mathcal{F}} n^{1/2} \left| \int g(x) d(F_n^X(x) - F^X(x)) \right|.$$

Using Markov's inequality, it follows that

$$\sup_{g \in \phi \circ \mathcal{F}} \left| \int g d(F_n^X - F^X) \right| = O_{\mathbb{P}}(n^{-1/2}),$$

from which we conclude that

$$\sup_{y \in \mathbb{R}, \gamma \in \mathbb{R}^d} |C_{n,\gamma}(y) - C_\gamma(y)| = O_{\mathbb{P}}(n^{-1/2}). \tag{A.9}$$

Since $II_n(\beta) \le \sup_{y \in \mathbb{R}, \gamma \in \mathbb{R}^d} |C_{n,\gamma}(y) - C_\gamma(y)|$ for all $\beta \in \mathbb{R}^d$ it follows that

$$\sup_{\beta \in \mathbb{R}^d} II_n(\beta) = O_{\mathbb{P}}(n^{-1/2}).$$

By the Cauchy-Schwarz inequality, we have that

$$
\begin{aligned}
III_n(\beta) &\le 2 \left(\int |\mathbb{F}_n^Y(y) - F^Y(y)|^2 dF^Y(y) \right)^{1/2} \left(\int |F^Y(y) - C_\beta(y)|^2 dF^Y(y) \right)^{1/2} \\
&\le 2 \left(\int |\mathbb{F}_n^Y(y) - F^Y(y)|^2 dF^Y(y) \right)^{1/2},
\end{aligned}
$$

since $|F^Y(y) - C_\beta(y)| \leq 1$. Therefore

$$\sup_{\beta \in \mathbb{R}^d} III_n(\beta) = O_{\mathbb{P}}(n^{-1/2}).$$

Similarly, we have

$$\sup_{\beta \in \mathbb{R}^d} IV_n(\beta) = O_{\mathbb{P}}(n^{-1/2}).$$

Also,

$$
\begin{aligned}
\sup_{\beta \in \mathbb{R}^d} V_n(\beta) \quad &\leq \quad 2 \sup_{y \in \mathbb{R}, \beta \in \mathbb{R}^d} |C_{n,\beta}(y) - C_\beta(y)| \times \int \left| F^Y(y) - C_\beta(y) \right| dF^Y(y) \\
&\leq \quad 2 \sup_{y \in \mathbb{R}, \beta \in \mathbb{R}^d} |C_{n,\beta}(y) - C_\beta(y)| = O_{\mathbb{P}}(n^{-1/2}).
\end{aligned}
$$

It follows that

$$\sup_{\beta \in \mathbb{R}^d} |\mathbb{D}_{n,2}(\beta) - \mathcal{D}_{2,F^Y}(\beta)| \xrightarrow{\mathbb{P}} 0.$$

In the following, we show that

$$0 = \mathcal{D}_{2,F^Y}(\beta_0) < \inf_{\beta \notin \mathcal{O}} \mathcal{D}_{2,F^Y}(\beta) \qquad (A.10)$$

for open set $\mathcal{O} \subset \mathbb{R}^d$ such that $\beta_0 \in \mathcal{O}$. Note that $F^Y = C_{\beta_0}$, therefore

$$\mathcal{D}_{2,F^Y}(\beta) = \int (C_\beta(y) - C_{\beta_0}(y))^2 \, dF^Y(y).$$

Let $\mathcal{O} = B(\beta_0, r) \subset \mathbb{R}^d$ be the open Euclidean Ball of center β_0 and radius $r > 0$. Suppose that $\inf_{\beta \notin \mathcal{O}} \mathcal{D}_{2,F^Y} = 0$. This means that we can find a sequence $\{\beta_m\}_{m \geq 1} \notin B(\beta_0, r)$ such that

$$\lim_{m \to \infty} \int (C_{\beta_m}(y) - C_{\beta_0}(y))^2 \, dF^Y(y) = 0. \qquad (A.11)$$

Assume there exists a subsequence $\{\beta_{m'}\}$ such that $\lim_{m' \to \infty} \|\beta_{m'}\|_2 = \infty$. Let $u_{m'} = \beta_{m'}/\|\beta_{m'}\|_2$. Since the unit sphere S^{d-1} is compact, there exists a subsequence of $\{u_{m'}\}$, which without loss of generality we denote by $(u_{m'})_{m'}$, such that $\lim_{m' \to \infty} u_{m'} = u$ where $u \in S^{d-1}$. Note that for a fixed y

$$
\begin{aligned}
C_{\beta_{m'}}(y) \quad &= \quad \int F^\epsilon(y - \|\beta_{m'}\|_2 u_{m'}^T x) dF^X(x) \\
&= \quad \int F^\epsilon(y - \|\beta_{m'}\|_2 u_{m'}^T x) \mathbb{1}\{u^T x = 0\} dF^X(x) \\
&\quad + \int F^\epsilon(y - \|\beta_{m'}\|_2 u_{m'}^T x) \mathbb{1}\{u^T x < 0\} dF^X(x) \\
&\quad + \int F^\epsilon(y - \|\beta_{m'}\|_2 u_{m'}^T x) \mathbb{1}\{u^T x > 0\} dF^X(x) \\
&= \quad I_{m'}(y) + II_{m'}(y) + III_{m'}(y).
\end{aligned}
$$

Since X does not belong to any affine space with probability one, we have

$$\int \mathbb{1}\{u^T x = 0\} dF^X(x) = 0,$$

and therefore $I_{m'} = 0$. Also, for all x such that $u^T x < 0$, we have $\|\beta_{m'}\|_2 u_{m'}^T x \to -\infty$ and hence $F^\epsilon(y - \|\beta_{m'}\|_2 u_{m'}^T x) \to 1$ as $m' \to \infty$. Similarly, for all x such that $u^T x > 0$, $F^\epsilon(y - \|\beta_{m'}\|_2 u_{m'}^T x) \to 0$. By the Dominated Convergence Theorem, it follows that

$$\lim_{m' \to \infty} II_{m'}(y) = 1, \quad \text{and} \quad \lim_{m' \to \infty} III_{m'}(y) = 0.$$

Using the Dominated Convergence Theorem again, we conclude that

$$\lim_{m' \to \infty} \int \left(C_{\beta_{m'}}(y) - C_{\beta_0}(y) \right)^2 dF^Y(y) = \int \left(\mathbb{P}(u^T X < 0) - C_{\beta_0}(y) \right)^2 dF^Y(y),$$

which implies with (A.11) that $C_{\beta_0}(y) = F^Y(y) = \mathbb{P}(u^T X < 0)$ for all $y \in \mathbb{R}$, which is impossible. We conclude that it is impossible that the sequence $(\beta_{m'})_{m'}$ has an Euclidean norm that tends to ∞. This means that it is bounded by some constant $K > 0$. Then, there exists a subsequence, w.o.l.g. we denote it by $\{\beta_{m'}\}$, converging to some vector $\tilde{\beta}$. Using similar arguments as above, we have

$$\int \left(C_{\tilde{\beta}}(y) - C_{\beta_0}(y) \right)^2 dF^Y(y) = 0,$$

and hence $C_{\tilde{\beta}} = C_{\beta_0}$. Since β_0 is the unique vector such that $Y \overset{d}{=} \beta_0^T X + \epsilon$, we must have $\tilde{\beta} = \beta_0$. However, this is impossible because this would mean that $r = 0$. Since any open set $\mathcal{O}$ containing β_0 contains $\mathcal{B}(\beta_0, r)$ for some $r > 0$, we conclude that the separability condition in (A.10) must hold. As $\hat{\beta}_n$ minimizes $\mathbb{D}_{n,2}$, it follows that all the conditions of (i) in (van der Vaart and Wellner, 1996, Corollary 3.2.3) are fulfilled. Therefore

$$\hat{\beta}_n \overset{\mathbb{P}}{\to} \beta_0.$$

This concludes the proof. ∎

A.7 Proof of Theorem 6

Proof Note that $\beta \in \mathcal{B}_0$ if and only if

$$\mathbb{E}\left(e^{\beta_0^T X t} \right) = \mathbb{E}\left(e^{\beta^T X t} \right)$$

for all $t \in \mathbb{R}$ such that the expectations on the left and right are defined. To use a simpler notation, we write now G for G_{X^1}. By the i.i.d. assumption, we can write that

$$G(\beta_{0,1} t) \times \cdots \times G(\beta_{0,d} t) = G(\beta_1 t) \times \cdots \times G(\beta_d t)$$

for all t such that

$$\max_{1 \leq i \leq d} (\beta_{0,i} t) < \lambda, \quad \text{and} \quad \max_{1 \leq i \leq d} (\beta_i t) < \lambda. \tag{A.12}$$

By ordering the components of β_0 and β we will assume without generality that $\beta_{0,1} \leq \ldots \leq \beta_{0,d}$ and $\beta_1 \leq \ldots \leq \beta_d$. Suppose that $\beta_{0,d} \neq \beta_d$. By symmetry, we can assume without loss of generality that $\beta_d > \beta_{0,d}$. There are 3 cases to consider.

- $\beta_d > 0$. Then, let t such that $t \nearrow \lambda/\beta_d$. Then, t belongs to the permissible region in (A.12) and

$$\frac{1}{(\lambda - \beta_{0,1}t)^\alpha} \times \ldots \times \frac{1}{(\lambda - \beta_{0,d}t)^\alpha} = \frac{1}{(\lambda - \beta_1 t)^\alpha} \times \ldots \times \frac{1}{(\lambda - \beta_d t)^\alpha} K(t)$$

for some function K such that $K(t) \in (0, \infty)$ for such t. This implies that

$$\lim_{t \nearrow \lambda/\beta_d} \frac{1}{(\lambda - \beta_{0,1}t)^\alpha} \times \ldots \times \frac{1}{(\lambda - \beta_{0,d}t)^\alpha} = \infty$$

which is impossible. Hence, $\beta_{0,d} = \beta_d$.

- $\beta_{0,d} < 0$. Then, this implies that $\beta_{0,1} \leq \ldots \leq \beta_{0,d} < 0$ and $\beta_1 \leq \ldots \leq \beta_d < 0$. We show that we have $\beta_{0,1} = \beta_1$. Indeed, suppose that $\beta_1 > \beta_{0,1}$. Take now t such that $t\beta_1 \nearrow \lambda$. This means that $t \searrow \lambda/\beta_{0,1} < 0$. Then, $t\beta_{0,i} \leq t\beta_{0,1}$ for $i \geq 2$ implying that $\max_{1 \leq i \leq d}(\beta_{0,i}t) = \beta_{0,1}t$. Also, $\max_{1 \leq i \leq d}(\beta_{0,i}t) = \beta_1 t < \beta_{0,1}t$. Thus, t belongs to the permissible region in (A.12). Then, we have that

$$\lim_{t \searrow \lambda/\beta_{0,1}} \frac{1}{(\lambda - \beta_{0,1}t)^\alpha} \times \ldots \times \frac{1}{(\lambda - \beta_{0,d}t)^\alpha} = \infty$$

which implies that

$$\lim_{t \searrow \lambda/\beta_{0,1}} \frac{1}{(\lambda - \beta_1 t)^\alpha} \times \ldots \times \frac{1}{(\lambda - \beta_d t)^\alpha} = \infty$$

which is impossible. Using a similar argument, we can show that we cannot have $\beta_{0,1} < \beta_1$. Hence, $\beta_{0,1} = \beta_1$. Also, one can show successively that $\beta_{0,i} = \beta_i$ for $i = 2, \cdots, d-1$ and finally that $\beta_{0,d} = \beta_d$.

- $\beta_d = 0$. Suppose that for all $i = 1, \cdots, d$ $\beta_i = 0$. Then,

$$\frac{1}{(\lambda - \beta_{0,1}t)^\alpha} \times \ldots \times \frac{1}{(\lambda - \beta_{0,d}t)^\alpha} H(\beta_{0,1}t) \times \ldots \times H(\beta_d t) = \frac{1}{\lambda^{d\alpha}} H(0)^d$$

for all t such that $\max_{1 \leq i \leq d}(\beta_{0,i}t) < \lambda$. It is easy to see that this implies that $\beta_{0,1} = \cdots = \beta_{0,d}$, which can be shown by assuming that this does not hold. Now, suppose one of the coefficients $\beta_{0,i}, i < d \neq 0$. Consider then the integer r such that $\beta_{r+1} = \cdots = \beta_d = 0$ and $\beta_r \neq 0$. Note that we necessarily have $\beta_1 \leq \cdots \leq \beta_r < 0$. Recall also that we are in the case where $\beta_{0,1} \leq \cdots \leq \beta_{0,d} < 0$. Then, using the same argument as in the second case, we can successively show that

$$\beta_{0,1} = \beta_1, \cdots, \beta_{0,r} = \beta_r.$$

These equalities will enable us to show that $\beta_{0,r+1} = \cdots = \beta_{0,d} = 0$.

31

Thus, in all the cases considered, the assumption $\beta_{0,d} \neq \beta_d$ leads to a contradiction. Using a recursive reasoning, we conclude that we should have $\beta_{0,i} = \beta_i$ for all $i = 1, \cdots, d$. Therefore, in general, and without assuming that the elements of β_0 and β are ordered, $\mathcal{B}_0$ consists of all vectors β such that for some permutation π we have $\beta_i = \beta_{0,\pi(i)}$. Consequently $\|\beta\|_2 = \|\beta_0\|_2$ and hence $\mathcal{B}_0 \subset \{\beta : \|\beta\|_2 = \|\beta_0\|_2\}$. ∎

A.8 Auxiliary results for the proof of Theorem 7

In the sequel, we would need the following definition. For a given class of functions $\mathcal{F}$ with envelope F we define

$$J(\delta, \mathcal{F}) = \sup_{\mathbb{Q}} \int_0^\delta \sqrt{1 + \log N(\eta\|F\|_{\mathbb{Q},2}, \mathcal{F}, L_2(\mathbb{Q}))} d\eta \tag{A.13}$$

where $N(\nu, \mathcal{F}, \|\cdot\|)$ is the ν-covering number of $\mathcal{F}$ with respect to $\|\cdot\|$. The supremum in (A.13) is taken over probability measure $\mathbb{Q}$ such that $\|F\|_{\mathbb{Q},2}^2 = \int F^2 d\mathbb{Q} > 0$.

To make the notation used below more compact, we shall use the classical empirical process notation

$$\mathbb{G}_n^X := \sqrt{n}(\mathbb{P}_n^X - \mathbb{P}^X), \quad \text{and} \quad \mathbb{G}_n^Y := \sqrt{n}(\mathbb{P}_n^Y - \mathbb{P}^Y).$$

For random vector Z we define its norm in $\mathcal{F}$ as

$$\|\mathbb{G}_n^Z\|_{\mathcal{F}} := \sup_{f \in \mathcal{F}} |\mathbb{G}^Z f|.$$

Additionally, we use the fact that

$$\int_0^1 \sqrt{\log\left(\frac{1}{t}\right)} dt \quad = \quad \int_1^\infty \frac{\sqrt{\log(x)}}{x^2} dx$$

together with the inequality in (A.8) shown above

Proposition 11 *Consider the classes of functions $\mathcal{I}$, $\mathcal{M}_{\beta_0}$, and $\mathcal{G}$ defined as*

$$\mathcal{I} = \left\{ I_y(t) = \mathbb{1}\{t \leq y\}, y \in \mathbb{R} \right\},$$

$$\mathcal{M}_{\beta_0} := \left\{ f_y(x) = m_{\beta_0, y}, y \in \mathbb{R} \right\},$$

with $m_{\gamma, y}(x) = F^\epsilon(y - \gamma^T x)$, $\gamma \in \mathbb{R}^d$, and

$$\mathcal{G} := \left\{ g_{\beta, y}(x) = m_{\beta, y}(x) - m_{\beta_0, y}(x), \ \|\beta - \beta_0\|_2 \leq \delta, y \in \mathbb{R} \right\}.$$

Then, we have

$$\mathbb{E}\left[\|\mathbb{G}_n^Y\|_{\mathcal{I}}^2\right] \lesssim 1, \quad \text{and} \quad \mathbb{E}\left[\|\mathbb{G}_n^Y\|_{\mathcal{M}_{\beta_0}}^2\right] \lesssim 1, \quad \text{and} \quad \mathbb{E}\left[\|\mathbb{G}_n^X\|_{\mathcal{G}}^2\right] \lesssim \delta^2.$$

Proof (Proposition 11). We begin with the class $\mathcal{I}$. This class is a VC-class with index equal to 2; see Example 2.6.1 in van der Vaart and Wellner (1996). Since $F \equiv 1$ is an envelope for the class, it follows from Theorem 2.6.7 of van der Vaart and Wellner (1996) that for any probability measure $\mathbb{Q}$ and $\eta \in (0, 1)$

$$N(\eta, \mathcal{I}, L_2(\mathbb{Q})) \leq 2K(16e)^2 \frac{1}{\eta^2}$$

for some universal constant $K > 0$. With the definition given in (A.13) and inequality $\sqrt{a+b} \leq \sqrt{a} + \sqrt{b}$ for any $a \geq 0, b \geq 0$, it follows that

$$J(1, \mathcal{I}) \quad \leq \quad 1 + \left(\sqrt{\log(2K(16e)^2) \vee 0} + \sqrt{2} \int_0^1 \sqrt{\log\left(\frac{1}{\eta}\right)} \right) d\eta < \infty$$

by (A.8). By Theorem 2.4.1 in van der Vaart and Wellner (1996), we have that

$$\mathbb{E}\left[\|\mathbb{G}_n^Y\|_{\mathcal{I}}^2 \right] \lesssim 1.$$

For the class $\mathcal{M}_{\beta_0}$, we use a small adaptation of Lemma 2.6.16 in van der Vaart and Wellner (1996) with $\psi = F^\epsilon$ to claim that it is a VC-class with index 2. The main difference between our setting and the one in that lemma is that we have $\beta_0^T x$ instead of a univariate x. However, the main argument in the proof of this lemma remains the same. Since 1 is an envelope, we conclude using arguments similar to those above to show that

$$\mathbb{E}\left[\|\mathbb{G}_n^Y\|_{\mathcal{M}_{\beta_0}}^2 \right] \lesssim 1.$$

Now, we turn to the class $\mathcal{G}$, and we start by showing that it is contained in a VC-hull. Without loss of generality, suppose that $\beta_0^T x \leq \beta^T x$. Then,

$$\begin{aligned}
g(x) &= m_{\beta, y}(x) - m_{\beta_0, y}(x) \\
&= \int \mathbb{1}\{t \in [y - \beta^T x, y - \beta_0^T x]\} dF^\epsilon(t) \\
&= \int \mathbb{1}\{\beta^T x + t \geq y\} \cdot \mathbb{1}\{\beta_0^T x + t \leq y\} dF^\epsilon(t) \\
&= \int \phi_1(\beta^T x + t - y) \cdot \phi_2(\beta_0^T x + t - y) dF^\epsilon(t)
\end{aligned} \tag{A.14}$$

where

$$\phi_1(u) = \mathbb{1}\{u \geq 0\}, \quad \text{and} \quad \phi_2(u) = \mathbb{1}\{u \leq 0\}, \quad u \in \mathbb{R}.$$

Now, consider the class of functions

$$\mathcal{F} = \{x \mapsto \gamma^T x + t - y : \gamma \in \mathbb{R}^d, t, y \in \mathbb{R}\}.$$

Then, $\mathcal{F}$ is a finite-dimensional space with dimension $d + 1$. Lemma 2.6.15 of van der Vaart and Wellner (1996) implies that this class is a VC-subgraph of index smaller than $d + 3$. Let $\mathcal{H}$ be the class defined as

$$\mathcal{H} = \{x \mapsto \phi_1(\beta^T x + t - y) \cdot \phi_2(\beta_0^T x + t - y), \quad \beta \in \mathbb{R}^d, t, y \in \mathbb{R}\}.$$

33

Let us denote $\mathcal{H}_1 = \phi_1 \circ \mathcal{F}$ and $\mathcal{H}_2 = \phi_2 \circ \mathcal{F}$. The key observation is that

$$\mathcal{H} \subset \mathcal{H}_1 \cdot \mathcal{H}_2 = \mathcal{H}_1 \wedge \mathcal{H}_2$$

where $\wedge$ denotes the minimum. The equality above is true since ϕ_1 and ϕ_2 are indicator functions which implies that for any pair $(f_1, f_2) \in \mathcal{F}^2$ we have that $\phi_1(f_1)\phi_2(f_2) = \phi_1(f_1) \wedge \phi_2(f_1)$. Now, It follows by (i) of Lemma 2.6.18 in van der Vaart and Wellner (1996) that $\mathcal{H}_1 \wedge \mathcal{H}_2$ is VC. This implies that $\mathcal{H}$ is VC. Given the definition of $g \in \mathcal{G}$ in (A.14), it follows that the class $\mathcal{G}$ is a subset of the convex hull of $\mathcal{H}$. By the mean value Theorem, we can find some real number θ^* (depending on x, y, β_0 and β) such that

$$g(x) = (\beta - \beta_0)^T x f^\epsilon(\theta^*).$$

Thus, using the assumption that $f^\epsilon \leq M$ and the Cauchy-Schwarz inequality, it follows that

$$G(x) = M\delta \|x\|_2$$

is an envelope of $\mathcal{G}$. By Theorem 2.6.9 of van der Vaart and Wellner (1996), we can find an integer $V \geq 2$ and a universal constant $K > 0$ such that we have for any probability measure Q satisfying $\|G\|_{Q,2} \in (0, \infty)$, i.e., $\int \|x\|_2 d\mathbb{Q}(x) \in (0, \infty)$

$$\log N \left(\eta \|G\|_{\mathbb{Q},2}, \mathcal{G}, L_2(\mathbb{Q}) \right) \leq K \left(\frac{1}{\eta} \right)^{2V/(V+2)}.$$

Hence, with $a = V/(V+2)$

$$J(1, \mathcal{G}) \leq 1 + \sqrt{K} \int_0^1 \frac{1}{\eta^a} d\eta < \infty$$

since $a < 1$. Under the assumption that $\int \|x\|^2 dF^X(x) < \infty$, we are allowed to use Theorem 2.4.1 in van der Vaart and Wellner (1996) to conclude that

$$\mathbb{E}\left[\|\mathbb{G}_n^X\|_{\mathcal{G}}^2 \right] \lesssim \delta^2$$

which completes the proof. ∎

Proposition 12 *Let M be the class of monotone functions f such that $f \in [0,1]$. Consider the class*

$$\mathcal{C} = \mathcal{M} + \mathcal{M} - 2\mathcal{M} = \left\{ f : f = f_1 + f_2 - 2f_3, (f_1, f_2, f_3) \in \mathcal{M}^3 \right\}.$$

Then, there exists a universal constant $A > 0$ such that for all $\eta \in (0,1)$

$$\log N_B(\eta, \mathcal{C}, L_2(\mathbb{P})) \leq \frac{A}{\eta}$$

where $N_B(\eta, \mathcal{C}, L_2(\mathbb{P}))$ denotes the η-bracketing number for $\mathcal{C}$ with respect to $L_2(\mathbb{P})$.

Proof (Proposition 12). Fix η and let $N = N_B(\eta, \mathcal{M}, L_2(\mathbb{P}))$. For $f \in \mathcal{C}$, we can find $1 \leq i, j, k \leq N$ and brackets (L_i, U_i), (L_j, U_j) and (L_k, U_k) such that

$$L_i \leq f_1 \leq U_i, \ L_j \leq f_2 \leq U_j, \ L_k \leq f_3 \leq U_k$$

and $\int (U_i - L_i)^2 d\mathbb{P} \leq \eta^2$, $\int (U_j - L_j)^2 d\mathbb{P} \leq \eta^2$ and $\int (U_k - L_k)^2 d\mathbb{P} \leq \eta^2$. Then,

$$L_i + L_j - 2U_k \leq f \leq U_i + U_j - 2L_k$$

and

$$
\begin{aligned}
\int \left(U_i + U_j - 2L_k - L_i - L_j + 2U_k \right)^2 d\mathbb{P} &= \int \left(U_i - L_j + U_j - L_j - 2(U_k - L_k) \right)^2 d\mathbb{P} \\
&\leq 4 \int (U_i - L_i)^2 d\mathbb{P} + 4 \int (U_j - L_j)^2 d\mathbb{P} \\
&\quad + 8 \int (U_k - L_k)^2 d\mathbb{P} \\
&\leq 16\eta^2,
\end{aligned}
$$

using the fact that $(a+b+c)^2 \leq 2(a+b)^2 + 2c^2 \leq 4(a^2+b^2) + 2c^2$. Hence, $N_B(4\eta, \mathcal{C}, L_2(\mathbb{P})) \leq N^3$ implying that

$$\log N_B(4\eta, \mathcal{C}, L_2(\mathbb{P})) \leq 3\log N \leq \frac{3K}{\eta}, \ \text{for some universal constant } K > 0$$

where the last inequality follows from Theorem 2.7.5 of van der Vaart and Wellner (1996). It follows that

$$\log N_B(\eta, \mathcal{C}, L_2(\mathbb{P})) \leq \frac{A}{\eta}$$

with $A = 12K$. $\blacksquare$

Proposition 13 *Let f be a real function such that $0 \leq f \leq M$ for some constant $M > 0$. Suppose, there exist real numbers $a < b$ such that f is non-decreasing $(-\infty, a]$, non-increasing on (a, b) and non-increasing on $[b, \infty)$. Let $X \in \mathbb{R}^d$ be a random vector such that $\mathbb{E}\|X\|_2^2 < \infty$. Let F_n^X be the empirical distribution of $X_1, \cdots, X_n$ that are independent and identically distributed as X. Then*

$$\sup_{(\beta, z) \in \mathbb{R}^d \times \mathbb{R}} \left\| \int x f(\beta^T x + z) dF_n^X(x) \right\|_2 = O_{\mathbb{P}}(1).$$

Proof (Proposition 13). Consider the function class $\mathcal{K} = \{k_{\beta,z} : k_{\beta,z}(x) = \beta^T x + z, \beta \in \mathbb{R}^d, z \in \mathbb{R}\}$ and note that it is $d+1$ dimensional. By Lemma 2.6.15 of van der Vaart and Wellner (1996), $\mathcal{K}$ is a VC class with index $V \leq d+3$. Define the functions

$$
\begin{aligned}
f_{M,a}(t) &= f(t)\mathbb{1}\{t \leq a\} + M\mathbb{1}\{t > a\}, \\
f_{M,b}(t) &= f(t)\mathbb{1}\{t \geq b\} + M\mathbb{1}\{t < b\}, \\
f_{M,a,b}(t) &= f(t)\mathbb{1}\{a < t < b\} + M\mathbb{1}\{t \leq a\}.
\end{aligned}
$$

Fix h such that $\|h\|_2 \le K$. Then, we can write

$$
\begin{aligned}
\int x f(\beta^T x + z) dF_n^X(x) &= \int x f \circ k_{\beta,z}(x) dF_n^X(x) \\
&= \int x f \mathbb{1}\{k_{\beta,z}(x) \le a\} dF_n^X(x) + \int x f \mathbb{1}\{a < k_{\beta,z}(x) < b\} dF_n^X(x) \\
&\quad + \int x f \mathbb{1}\{k_{\beta,z}(x) \ge b\} dF_n^X(x) \\
&= \int x f_{M,a} \circ k_{\beta,z}(x) dF_n^X(x) - M \int x \mathbb{1}\{k_{\beta,z}(x) > a\} dF_n^X(x) \\
&\quad + \int x f_{M,a,b} \circ k_{\beta,z}(x) dF_n^X(x) - M \int x \mathbb{1}\{k_{\beta,z}(x) \le a\} dF_n^X(x) \\
&\quad + \int x f_{M,b} \circ k_{\beta,z}(x) dF_n^X(x) - M \int x \mathbb{1}\{k_{\beta,z}(x) < b\} dF_n^X(x) \\
&= \int x f_{M,a} \circ k_{\beta,z}(x) dF_n^X(x) + \int x f_{M,a,b} \circ k_{\beta,z}(x) dF_n^X(x) \\
&\quad + \int x f_{M,b} \circ k_{\beta,z}(x) dF_n^X(x) - M \int x \mathbb{1}\{k_{\beta,z}(x) < b\} dF_n^X(x) \\
&= T_{n,1}(\beta, z, h) + T_{n,2}(\beta, z, h) + T_{n,3}(\beta, z, h) + T_{n,4}(\beta, z, h).
\end{aligned}
$$

Functions $f_{M,a}$, $f_{M,b}$, $f_{M,a,b}$, and $\mathbb{1}\{\cdot < b\}$ are all bounded monotone functions. Following from Lemma 2.6.18 (viii) and (vi), Theorem 2.6.7, Theorem 2.14.1 of van der Vaart and Wellner (1996) for $j \in \{1, 2, 3, 4\}$ we have

$$
\mathbb{E}\left[\left(\sup_{\beta, z} |T_{n,j}(\beta, z)| \right)^2 \right] \lesssim \mathbb{E}[\|X\|_2^2] < \infty.
$$

Application of Chebychev's inequality yields the claim. ∎

Proposition 14 *There exists a constant $C > 0$ depending only on β^0 such that for $\delta > 0$ small enough we have that*

$$
\mathbb{E} \sup_{\|\beta - \beta_0\|_2 < \delta} \sqrt{n} |(\mathbb{D}_{n,2}(\beta) - \mathcal{D}_{2,F^Y}(\beta)) - (\mathbb{D}_{n,2}(\beta_0) - \mathcal{D}_{2,F^Y}(\beta_0))| \le C \left(\delta + \frac{\delta}{\sqrt{n}} + \frac{1}{\sqrt{n}} \right) \equiv \phi_n(\delta).
$$

Proof (Proposition 14). We begin with rewriting $\mathbb{D}_{n,2}(\beta)$ and $\mathcal{D}_{2,F^Y}(\beta)$ for a given $\beta \in \mathbb{R}^d$. For $y \in \mathbb{R}$ and $\beta \in \mathbb{R}^d$ define the function

$$
m_{\beta,y}(x) = F^\epsilon(y - \beta^T x), \ x \in \mathbb{R}^d.
$$

Then,

$$
\mathbb{D}_{n,2}(\beta) = \int \left(F_n^Y(y) - \mathbb{E}_n^X m_{\beta,y} \right)^2 dF_n^Y(y).
$$

Since $F^Y(y) = \int F^\epsilon(y - \beta_0^T x) dF^X(x) = \mathbb{E}^X m_{\beta_0, y}$ for all $y \in \mathbb{R}$, we have

$$\mathcal{D}_{2, F^Y}(\beta) = \int \left(F^Y(y) - \mathbb{E}^X m_{\beta, y} \right)^2 dF^Y(y) = \int \left(\mathbb{E}^X (m_{\beta, y} - m_{\beta_0, y}) \right)^2 dF^Y(y).$$

Therefore,

$$
\begin{aligned}
\mathbb{D}_{n,2}(\beta) &= \int \left(F_n^Y(y) - \mathbb{E}_n^X m_{\beta_0, y} - \mathbb{E}_n^X(m_{\beta, y} - m_{\beta_0, y}) \right)^2 dF_n^Y(y) \\
&= \int \left(F_n^Y(y) - \mathbb{E}_n^X m_{\beta_0, y} \right)^2 dF_n^Y(y) \\
&\quad - 2 \int \left(F_n^Y(y) - \mathbb{E}_n^X m_{\beta_0, y} \right) \mathbb{E}_n^X(m_{\beta, y} - m_{\beta_0, y}) dF_n^Y(y) \\
&\quad + \int \left(\mathbb{E}_n^X(m_{\beta, y} - m_{\beta_0, y}) \right)^2 dF_n^Y(y) \\
&= \mathbb{D}_{n,2}(\beta_0) - 2 \int \left(F_n^Y(y) - \mathbb{E}_n^X m_{\beta_0, y} \right) \mathbb{E}_n^X(m_{\beta, y} - m_{\beta_0, y}) dF_n^Y(y) \\
&\quad + \int \left(\mathbb{E}_n^X(m_{\beta, y} - m_{\beta_0, y}) \right)^2 dF_n^Y(y) \\
&= \mathbb{D}_{n,2}(\beta_0) \\
&\quad - 2 \int \left(F_n^Y(y) - \mathbb{E}_n^X m_{\beta_0, y} \right) \left(\mathbb{E}_n^X(m_{\beta, y} - m_{\beta_0, y}) - \mathbb{E}^X(m_{\beta, y} - m_{\beta_0, y}) \right) dF_n^Y(y) \\
&\quad - 2 \int \left(F_n^Y(y) - \mathbb{E}_n^X m_{\beta_0, y} \right) \mathbb{E}^X(m_{\beta, y} - m_{\beta_0, y}) dF_n^Y(y) \\
&\quad + \int \left(\mathbb{E}_n^X(m_{\beta, y} - m_{\beta_0, y}) \right)^2 dF_n^Y(y)
\end{aligned}
$$

with

$$
\begin{aligned}
&\int \left(\mathbb{E}_n^X(m_{\beta, y} - m_{\beta_0, y}) \right)^2 dF_n^Y(y) \\
&= \int \left[\left(\mathbb{E}_n^X(m_{\beta, y} - m_{\beta_0, y}) - \mathbb{E}^X(m_{\beta, y} - m_{\beta_0, y}) \right) \right]^2 d\mathbb{P}_n^Y(y) \\
&\quad + 2 \int \left(\mathbb{E}_n^X(m_{\beta, y} - m_{\beta_0, y}) - \mathbb{E}^X(m_{\beta, y} - m_{\beta_0, y}) \right) \mathbb{E}^X |m_{\beta, y} - m_{\beta_0, y}| dF_n^Y(y) \\
&\quad + \int \left(\mathbb{E}^X(m_{\beta, y} - m_{\beta_0, y}) \right)^2 dF_n^Y(y) \\
&= \int \left[\left(\mathbb{E}_n^X(m_{\beta, y} - m_{\beta_0, y}) - \mathbb{E}^X(m_{\beta, y} - m_{\beta_0, y}) \right) \right]^2 dF_n^Y(y) \\
&\quad + 2 \int \left(\mathbb{E}_n^X(m_{\beta, y} - m_{\beta_0, y}) - \mathbb{E}^X(m_{\beta, y} - m_{\beta_0, y}) \right) \mathbb{E}^X(m_{\beta, y} - m_{\beta_0, y}) dF_n^Y(y) \\
&\quad + \mathcal{D}_{2, F^Y}(\beta) + \int \left(\mathbb{E}^X(m_{\beta, y} - m_{\beta_0, y}) \right)^2 d(F_n^Y(y) - F^Y(y))
\end{aligned}
$$

37

These together with $\mathcal{D}_{2,F^Y}(\beta_0) = 0$ gives

$$\sqrt{n}\left|(\mathbb{D}_{n,2}(\beta) - \mathcal{D}_{2,F^Y}(\beta)) - (\mathbb{D}_{n,2}(\beta_0) - \mathcal{D}_{2,F^Y}(\beta_0))\right|$$

$$\leq 2\sqrt{n}\int \left|F_n^Y(y) - \mathbb{E}_n^X m_{\beta_0,y}\right|\left|(\mathbb{E}_n^X(m_{\beta,y} - m_{\beta_0,y}) - \mathbb{E}^X(m_{\beta,y} - m_{\beta_0,y}))\right|dF_n^Y(y)$$

$$+ 2\sqrt{n}\int \left|F_n^Y(y) - \mathbb{E}_n^X m_{\beta_0,y}\right|\mathbb{E}^X(m_{\beta,y} - m_{\beta_0,y})dF_n^Y(y)$$

$$+ \sqrt{n}\int \left[\left(\mathbb{E}_n^X(m_{\beta,y} - m_{\beta_0,y}) - \mathbb{E}^X(m_{\beta,y} - m_{\beta_0,y})\right)\right]^2 dF_n^Y(y)$$

$$+ 2\sqrt{n}\int \left|(\mathbb{E}_n^X(m_{\beta,y} - m_{\beta_0,y}) - \mathbb{E}^X(m_{\beta,y} - m_{\beta_0,y}))\right|\mathbb{E}^X(m_{\beta,y} - m_{\beta_0,y})dF_n^Y(y)$$

$$+ \sqrt{n}\left|\int \left(\mathbb{E}^X(m_{\beta,y} - m_{\beta_0,y})\right)^2 d(F_n^Y(y) - F^Y(y))\right|$$

$$= 2T_{n,1}(\beta) + 2T_{n,2}(\beta) + T_{n,3}(\beta) + T_{n,4}(\beta) + 2T_{n,5}(\beta).$$

Since $F^Y(y) = \mathbb{E}^X m_{\beta_0,y}$, the triangle inequality gives us

$$\sup_{\beta:\|\beta-\beta_0\|_2\leq\delta} T_{n,1}(\beta)$$

$$\leq \sqrt{n}\sup_{y\in\mathbb{R}}\left|(F_n^Y - F^Y)(y)\right| \times$$

$$\sup_{y\in\mathbb{R},\beta:\|\beta-\beta_0\|_2\leq\delta}\left|(\mathbb{E}_n^X(m_{\beta,y} - m_{\beta_0,y}) - \mathbb{E}^X(m_{\beta,y} - m_{\beta_0,y}))\right|$$

$$+ \sqrt{n}\sup_{y\in\mathbb{R}}\left|(\mathbb{E}_n^X m_{\beta_0,y} - \mathbb{E}^X m_{\beta_0,y})\right| \times$$

$$\sup_{y\in\mathbb{R},\beta:\|\beta-\beta_0\|_2\leq\delta}\left|(\mathbb{E}_n^X(m_{\beta,y} - m_{\beta_0,y}) - \mathbb{E}^X(m_{\beta,y} - m_{\beta_0,y}))\right|$$

$$= \frac{1}{\sqrt{n}}\left(\|\mathbb{G}^Y\|_{\mathcal{I}} + \|\mathbb{G}_n^X\|_{\mathcal{M}_{\beta_0}}\right)\|\mathbb{G}_n^X\|_{\mathcal{G}}.$$

Then by Cauchy-Schwarz inequality

$$\mathbb{E}\left[\sup_{\beta:\|\beta-\beta_0\|_2\leq\delta} T_{n,1}(\beta)\right] \leq \frac{1}{\sqrt{n}}\mathbb{E}\left[\left(\|\mathbb{G}^Y\|_{\mathcal{I}} + \|\mathbb{G}_n^X\|_{\mathcal{M}_{\beta_0}}\right)\|\mathbb{G}_n^X\|_{\mathcal{G}}\right]$$

$$\leq \frac{1}{\sqrt{n}}\mathbb{E}\left[\left(\|\mathbb{G}^Y\|_{\mathcal{I}} + \|\mathbb{G}_n^X\|_{\mathcal{M}_{\beta_0}}\right)^2\right]^{1/2}\mathbb{E}\left[\|\mathbb{G}_n^X\|_{\mathcal{G}}^2\right]^{1/2}$$

$$\lesssim \frac{\delta}{\sqrt{n}}.$$

We also have that

$$\sup_{\beta:\|\beta-\beta_0\|_2\leq\delta} T_{n,2}(\beta) \leq \left(\|\mathbb{G}_n^Y\|_{\mathcal{I}} + \|\mathbb{G}_n^X\|_{\mathcal{M}_{\beta_0}}\right)\sup_{y\in\mathbb{R},\beta:\|\beta-\beta_0\|_2\leq\delta}\mathbb{E}^X|m_{\beta,y} - m_{\beta_0,y}|$$

$$\leq M\delta\left(\|\mathbb{G}_n^Y\|_{\mathcal{I}} + \|\mathbb{G}_n^X\|_{\mathcal{M}_{\beta_0}}\right)\mathbb{E}\|X\|_2$$

$$\lesssim \delta$$

since $\mathbb{E}\|X\|_2^2 < \infty$ implies $\mathbb{E}\|X\|_2 < \infty$. This implies that

$$
\begin{aligned}
\mathbb{E}\left[\sup_{\beta:\|\beta-\beta_0\|_2\leq\delta} T_{n,2}(\beta)\right] &\lesssim \left(\mathbb{E}\|\mathbb{G}_n^Y\|_{\mathcal{I}} + \mathbb{E}\|\mathbb{G}_n^X\|_{\mathcal{M}_{\beta_0}}\right)\delta \\
&\leq \left(\sqrt{\mathbb{E}\|\mathbb{G}_n^Y\|_{\mathcal{I}}^2} + \sqrt{\mathbb{E}\|\mathbb{G}_n^X\|_{\mathcal{M}_{\beta_0}}^2}\right)\delta \\
&\lesssim \delta.
\end{aligned}
$$

Now,

$$
\begin{aligned}
\sup_{\beta:\|\beta-\beta_0\|_2\leq\delta} T_{n,3}(\beta) &\leq \frac{1}{\sqrt{n}}\sup_{y\in\mathbb{R}}\sup_{\beta:\|\beta-\beta_0\|_2\leq\delta}\left|\sqrt{n}(\mathbb{E}_n^X(m_{\beta,y}-m_{\beta_0,y}) - \mathbb{E}^X(m_{\beta,y}-m_{\beta_0,y}))\right|^2 \\
&\leq \frac{1}{\sqrt{n}}\|\mathbb{G}_n^X\|_{\mathcal{G}}^2
\end{aligned}
$$

and hence

$$
\mathbb{E}\left[\sup_{\beta:\|\beta-\beta_0\|_2\leq\delta} T_{n,3}(\beta)\right] \leq \frac{1}{\sqrt{n}}\mathbb{E}\|\mathbb{G}_n^X\|_{\mathcal{G}}^2 \lesssim \frac{\delta^2}{\sqrt{n}}.
$$

We have that

$$
\begin{aligned}
\sup_{\beta:\|\beta-\beta_0\|_2\leq\delta} T_{n,4}(\beta) &\leq \|\mathbb{G}_n^X\|_{\mathcal{G}}\sup_{y\in\mathbb{R},\beta:\|\beta-\beta_0\|_2\leq\delta}\mathbb{E}^X|m_{\beta,y}-m_{\beta_0,y}| \\
&\lesssim \delta\|\mathbb{G}_n^X\|_{\mathcal{G}},
\end{aligned}
$$

which implies

$$
\mathbb{E}\left[\sup_{\beta:\|\beta-\beta_0\|_2\leq\delta} T_{n,4}(\beta)\right] \lesssim \delta\sqrt{\mathbb{E}\left[\|\mathbb{G}_n^X\|_{\mathcal{G}}^2\right]} \lesssim \delta^2.
$$

Finally, to bound $T_{n,5}$, note that

$$
T_{n,5} \leq \|\mathbb{G}_n^Y\|_{\mathcal{K}}
$$

where

$$
\mathcal{K} = \left\{y \mapsto k_\beta(y) = (\mathbb{E}^X(m_{\beta,y}-m_{\beta_0,y}))^2 : \|\beta-\beta_0\|_2 \leq \delta\right\}
$$

More explicitly,

$$
\begin{aligned}
k_\beta(y) &= \left(\int (m_{\beta,y}(x) - m_{\beta_0,y}(x))\,dF^X(x)\right)^2 \\
&= \left(\int m_{\beta,y}(x)dF^X(x)\right)^2 + \left(\int m_{\beta_0,y}(x)dF^X(x)\right)^2 \\
&\quad -2\left(\int m_{\beta_0,y}(x)dF^X(x)\right)\left(\int m_{\beta,y}(x)dF^X(x)\right) \\
&= k_{1,\beta}(y) + k_{2,\beta}(y) - 2k_{3,\beta}(y).
\end{aligned}
$$

39

Functions $k_{i,\beta}$ are non-decreasing and $k_{i,\beta} \in [0,1]$ for $i = 1, 2, 3$. Thus, $\mathcal{K} \subset \mathcal{C}$ defined in Proposition 12. For $\nu > 0$ define

$$\tilde{J}(\nu) = \int_0^\nu \sqrt{1 + \log N_B(\eta, \mathcal{K}, L_2(F^Y))} d\eta.$$

Since $\mathbb{E}k_\beta^2 \lesssim \delta^4$ and $\|k_\beta\|_\infty \lesssim \delta^2$, it follows from Lemma 3.4.2 of van der Vaart and Wellner (1996) that

$$\mathbb{E}\|\mathbb{G}_n^Y\|_{\mathcal{K}} \leq \tilde{J}(\delta^2)\left(1 + \frac{\tilde{J}(\delta^2)}{\delta^4\sqrt{n}}\right).$$

Now, by Proposition 12 and for small enough $\nu > 0$ we have

$$\begin{aligned}
\tilde{J}(\nu) &\leq \int_0^\nu \sqrt{1 + \log N_B(\eta, \mathcal{C}, L_2(F^Y)} d\eta \\
&\leq \nu + \sqrt{A}\int_0^\nu \eta^{-1/2} d\eta = \mu + 2\sqrt{A}\sqrt{\nu} \lesssim \sqrt{\nu}.
\end{aligned}$$

Thus,

$$\mathbb{E}\|\mathbb{G}_n^Y\|_{\mathcal{K}} \lesssim \delta\left(1 + \frac{\delta}{\delta^4\sqrt{n}}\delta^2\right) = \delta + \frac{1}{\sqrt{n}}.$$

Taking $\delta \in (0,1)$, there exists a constant $D > 0$ such that

$$\begin{aligned}
\mathbb{E}\sup_{\|\beta-\beta_0\|_2 < \delta} \sqrt{n}\left|(\mathbb{D}_{n,2}(\beta) - \mathcal{D}_{2,F^Y}(\beta)) - (\mathbb{D}_{n,2}(\beta_0) - \mathcal{D}_{2,F^Y}(\beta_0))\right| &\leq D\left(3\delta + 2\frac{\delta}{\sqrt{n}} + \frac{1}{\sqrt{n}}\right) \\
&\lesssim \delta + \frac{\delta}{\sqrt{n}} + \frac{1}{\sqrt{n}},
\end{aligned}$$

which completes the proof. ∎

Proposition 15 *Assume that F^ϵ admits a density $f^\epsilon \in C^2$ such that f^ϵ, and $|(f^\epsilon)^{(1)}|$ are bounded by some constant $M > 0$, and $\mathbb{E}\|X\|_2^2 < \infty$. Also, assume that matrix*

$$U = \int\left(\int xf^\epsilon(y - \beta_0^T x)dF^X(x)\right)\left(\int x^T f^\epsilon(y - \beta_0^T x)dF^X(x)\right)dF^Y(y)$$

is positive definite. Then, there exists a small neighborhood of β_0 and a constant $C > 0$ such that for all β in this neighborhood

$$\mathcal{D}_{2,F^Y}(\beta) \geq C\|\beta - \beta_0\|_2^2.$$

Proof (Proposition 15). We need to show that

$$\psi(\beta) \equiv \mathcal{D}_{2,F^Y}(\beta), \ \beta \in \mathbb{R}^d$$

is twice continuously differentiable in a small neighbourhood of β_0 such that

$$\frac{\partial \psi(\beta)}{\partial \beta}\Big|_{\beta_0} = \mathbf{0}, \tag{A.15}$$

and the Hessian matrix $\frac{\partial^2 \psi(\beta)}{\partial \beta \partial \beta^T}\big|_{\beta_0}$ is positive definite. We can write

$$\psi(\beta) = \int (\Delta(\beta, y))^2 dF^Y(y),$$

where

$$\Delta(\beta, y) = \int F^\epsilon(y - \beta^T x) dF^X(x) - \int F^\epsilon(y - \beta_0^T x) dF^X(x), \ (\beta, y) \in \mathbb{R}^d \times \mathbb{R}.$$

By assumption, the distribution function F^ϵ is continuously differentiable on $\mathbb{R}$, and hence

$$\frac{\partial}{\partial \beta} F^\epsilon(y - \beta^T x) = -x f^\epsilon(y - \beta^T x).$$

Furthermore, and by the assumption of the proposition, we have $\|x f^\epsilon(y - \beta^T x)\|_2 \leq M\|x\|_2$ and $\mathbb{E}\|X\|_2 < \infty$. Hence, for any fixed y, $\beta \mapsto \|\frac{\partial}{\partial \beta} F^\epsilon(y - \beta^T x)\| \leq M\|x\|_2$ which is integrable. Thus, by the Dominated Convergence Theorem, it follows that for any fixed y

$$\frac{\partial \Delta(\beta, y)}{\partial \beta} = -\int x f^\epsilon(y - \beta^T x) dF^X(x),$$

and hence

$$\frac{\partial (\Delta(\beta, y))^2}{\partial \beta} = -2\Delta(\beta, y) \int x f^\epsilon(y - \beta^T x) dF^X(x).$$

Since $|F^\epsilon(y - \beta^T x) - F^\epsilon(y - \beta_0^T x)| \leq 1$, it follows that $|\Delta(\beta, y)| \leq 1$ for all β, y. Thus,

$$\left\| \frac{\partial (\Delta(\beta, y))^2}{\partial \beta} \right\|_2 \leq 2 \left\| \Delta(\beta, y) \int x f^\epsilon(y - \beta^T x) dF^X(x) \right\|_2 \leq M\mathbb{E}\|X\|_2.$$

From the Dominated Convergence Theorem, it follows that for any $\beta \in \mathbb{R}^d$

$$\frac{\partial \psi(\beta)}{\partial \beta} = -2 \int \Delta(\beta, y) \left(\int x f^\epsilon(y - \beta^T x) dF^X(x) \right) dF^Y(y).$$

In particular, (A.15) holds. Now, we show that the multivariate

$$f(\beta, y) = -2\Delta(\beta, y) \left(\int x f^\epsilon(y - \beta^T x) dF^X(x) \right)$$

is continuously differentiable on $\mathbb{R}$ with a positive definite Hessian at β_0. Since we have already shown that $\beta \mapsto \Delta(\beta, y)$ is differentiable, we need to show differentiability of

$$\beta \mapsto \int x f^\epsilon(y - \beta^T x) dF^X(x). \tag{A.16}$$

41

Using the assumption that f^ϵ is continuously differentiable such that $|(f^\epsilon)^{(1)}| \leq M$, we can use arguments that are very similar to those above to show that the function in (A.17) is indeed differentiable with gradient

$$\beta \mapsto -\int xx^T (f^\epsilon)^{(1)}(y - \beta^T x)dF^X(x). \tag{A.17}$$

Thus, for a fixed $y \in \mathbb{R}$

$$\begin{aligned}
\frac{\partial f(\beta, y)}{\beta^T} &= 2\Delta(\beta, y)\left(\int xx^T (f^\epsilon)^{(1)}(y - \beta^T x)dF^X(x)\right) \\
&\quad + 2\left(\int xf^\epsilon(y - \beta^T x)dF^X(x)\right)\left(\int x^T f^\epsilon(y - \beta^T x)dF^X(x)\right).
\end{aligned}$$

For $1 \leq i, j \leq d$, we have that

$$\begin{aligned}
&\left|\Delta(\beta, y)\left(\int [xx^T]_{ij}(f^\epsilon)^{(1)}(y - \beta^T x)dF^X(x)\right)\right. \\
&\quad \left. + 2\left(\int x_i f^\epsilon(y - \beta^T x)dF^X(x)\right)\left(\int x_j f^\epsilon(y - \beta^T x)dF^X(x)\right)\right| \\
&\leq M\int |[xx^T]_{ij}|dF^X(x) + 2M\left(\int |x_i|dF^X(x)\right)\left(\int |x_j|dF^X(x)\right) \\
&= M\int |x_i x_j|dF^X(x) + 2M\left(\int |x_i|dF^X(x)\right)\left(\int |x_j|dF^X(x)\right) \\
&\leq M\mathbb{E}\|X\|_2^2 + 2M\left(\mathbb{E}\|X\|_2\right)^2 \tag{A.18}
\end{aligned}$$

using the Cauchy-Schwarz inequality and the fact that $|x_i| \leq \|x\|_2$. The term on the right of (A.18) is a constant and hence integrable with respect to F^Y. Using the Dominated convergence Theorem again, it follows that ψ is twice continuously differentiable and

$$\begin{aligned}
\frac{\partial^2 \psi(\beta)}{\partial\beta\partial\beta^T} &= 2\int \Delta(\beta, y)\left(\int xx^T (f^\epsilon)^{(1)}(y - \beta^T x)dF^X(x)\right)dF^Y(y) \\
&\quad + 2\int\left(\int xf^\epsilon(y - \beta^T x)dF^X(x)\right)\left(\int x^T f^\epsilon(y - \beta^T x)dF^X(x)\right)dF^Y(y).
\end{aligned}$$

Therefore we have

$$\frac{\partial\psi(\beta)}{\partial\beta\partial\beta^T}\Big|_{\beta_0} = 2\int\left(\int xf^\epsilon(y - \beta_0^T x)dF^X(x)\right)\left(\int x^T f^\epsilon(y - \beta_0^T x)dF^X(x)\right)dF^Y(y) = 2U,$$

which is positive definite by assumption. Using Taylor expansion of ψ up to the second order, and using $\psi(\beta_0) = 0$ and (A.15) we can write

$$\begin{aligned}
\psi(\beta) &= \psi(\beta_0) + (\beta - \beta_0)^T \frac{\partial\psi(\beta)}{\partial\beta}\Big|_{\beta_0} + (\beta - \beta_0)^T U(\beta - \beta_0) + o(\|\beta - \beta_0\|_2^2) \\
&= (\beta - \beta_0)^T U(\beta - \beta_0) + o(\|\beta - \beta_0\|_2^2).
\end{aligned}$$

If λ_1 is the smallest eigenvalue of U, then $\lambda_1 > 0$ and also $(\beta - \beta_0)^T U (\beta - \beta_0) \geq \lambda_1 \|\beta - \beta_0\|_2^2$. Also, there exist a small enough neighborhood of β_0 such that when β is in this neighborhood we have $o(\|\beta - \beta_0\|_2^2) > -\lambda_1 \|\beta - \beta_0\|_2^2/2$. This implies that for β in this neighbourhood

$$\psi(\beta) = \mathcal{D}_{2,F^Y}(\beta) \geq \frac{\lambda_1}{2}\|\beta - \beta_0\|_2^2$$

which completes the proof. ∎

Theorem 16 *Let $K > 0$, $h \in B(\mathbf{0}, K) = \{h \in \mathbb{R}^d : \|h\|_2 \leq K\}$, and $\mathbb{B}_1, \mathbb{B}_2$ two independent standard Brownian Bridges from $(0,0)$ to $(1,0)$. Assume that F^ϵ admits a density f^ϵ which is continuously differentiable and that f^ϵ, and $|(f^\epsilon)^{(1)}|$ are bounded by $M > 0$. There exist an integer $m \geq 1$ and $a_1 < a_2 < \ldots < a_m$ such that f^ϵ is monotone on $(-\infty, a_1]$, (a_i, a_{i+1}), $i = 1, \ldots, m-1$ and non-increasing on $[a_m, \infty)$ and $\mathbb{E}\|X\|_2^2 < \infty$. Let*

$$\mathbb{Q}_n(h) = n\mathbb{D}_{n,2}(\beta_0 + hn^{-1/2}).$$

Then,

$$\mathbb{Q}_n \Rightarrow \mathbb{Q}$$

in $\ell^\infty(B(\mathbf{0}, K))$, where

$$\mathbb{Q}(h) =$$
$$\int \left\{ \mathbb{B}_1 \circ F^Y(y) - \int F^\epsilon(y - \beta_0^T x) d\mathbb{B}_2 \circ F^X(x) + h^T \int x f^\epsilon(y - \beta_0^T x) dF^X(x) \right\}^2 dF^Y(y)$$
$$= \int \left\{ \mathbb{B}_1 \circ F^Y(y) - \int \mathbb{B}_2 \circ F^Z(z) f^\epsilon(y - z) dz + h^T \int x f^\epsilon(y - \beta_0^T x) dF^X(x) \right\}^2 dF^Y(y)$$

with $Z = \beta_0^T X$. Furthermore, if the $d \times d$ matrix

$$U = \int \left(\int x f^\epsilon(y - \beta_0^T x) dF^X(x) \right) \left(\int x^T f^\epsilon(y - \beta_0^T x) dF^X(x) \right) dF^Y(y)$$

is positive definite then $h^0 = U^{-1}V$ is the unique minimizer of the process $\mathbb{Q}$ where

$$V = -\int \left(\mathbb{B}_1 \circ F^Y(y) - \int F^\epsilon(y - \beta_0^T x) d\mathbb{B}_2 \circ F^X(x) \right) \int \left(x f^\epsilon(y - \beta_0^T x) dF^X(x) \right) dF^Y(y)$$
$$= -\int \left(\mathbb{B}_1 \circ F^Y(y) - \int \mathbb{B}_2 \circ F^Z(z) f^\epsilon(y - z) dz \right) \int \left(x f^\epsilon(y - \beta_0^T x) dF^X(x) \right) dF^Y(y)$$

Proof (Therem 16). Let $h \in B(\mathbf{0}, K)$ for a fixed $K > 0$. Using Taylor expansion, we have

$$F^\epsilon(y - (\beta_0 + hn^{-1/2})^T x) = F^\epsilon(y - \beta_0^T x - n^{-1/2}h^T x)$$
$$= F^\epsilon(y - \beta_0^T x) - n^{-1/2}h^T x f^\epsilon(y - \beta_0^T x) + \frac{1}{2}n^{-1}(h^T x)^2 (f^\epsilon)^{(1)}(\theta_n^*),$$

for some $\theta_n^* = \theta^*(y, x, n)$ between $y - \beta_0^T x$ and $y - \beta_0^T x - n^{-1/2} h^T x$. Furthermore, we shall use the well-known fact that if $Z_1, \ldots, Z_n$ are i.i.d. $\sim F$ then there exists a Brownian Bridge $\mathbb{B}$ defined on the same probability space as $(Z_1, \ldots, Z_n)$ such that

$$\|\mathbb{B}_n - \mathbb{B}\|_\infty \to 0$$

almost surely, where $\mathbb{B}_n$ is defined such that for all for all $z \in \mathbb{R}$ we have

$$F_n(z) = n^{-1} \sum_{i=1}^n \mathbb{1}\{Z_i \le z\} = F(z) + n^{-1/2} \mathbb{B}_n \circ F(z).$$

Now,

$$\mathbb{Q}_n(h) = \int \mathbb{K}_n^2(y, h) dF_n^Y(y)$$

where

$$
\begin{aligned}
\mathbb{K}_n^2(y, h) &= n \Big\{ F_n^Y(y) - \int F^\epsilon(y - \beta_0^T x) dF_n^X(x) + n^{-1/2} \int h^T x f^\epsilon(y - \beta_0^T x) dF_n^X(x) \\
&\quad - \frac{1}{2} n^{-1} \int (h^T x)^2 (f^\epsilon)^{(1)}(\theta_n^*) dF_n^X(x) \Big\}^2 \\
&= \Big\{ \sqrt{n}(F_n^Y(y) - F^Y(y)) - \int F^\epsilon(y - \beta_0^T x) d\sqrt{n}(F_n^X(x) - F^X(x)) \\
&\quad + \int h^T x f^\epsilon(y - \beta_0^T x) dF_n^X(x) - \frac{1}{2} n^{-1/2} \int (h^T x)^2 (f^\epsilon)^{(1)}(\theta_n^*) dF_n^X(x) \Big\}^2.
\end{aligned}
$$

Using again $F^Y(y) = \int F^\epsilon(y - \beta_0^T x) dF^X(x)$

$$
\begin{aligned}
\mathbb{K}_n^2(y, h) &= \Big\{ \mathbb{B}_{n,1}(F^Y(y)) - \int \mathbb{B}_{n,2}(F^Z(z)) f^\epsilon(y - z) dz + \int h^T x f^\epsilon(y - \beta_0^T x) dF_n^X(x) \\
&\quad - \frac{1}{2} n^{-1/2} \int (h^T x)^2 (f^\epsilon)^{(1)}(\theta_n^*) dF_n^X(x) \Big\}^2,
\end{aligned}
$$

where $Z = \beta_0^T X$, and $\mathbb{B}_{n,1} \circ F^Y(y) = \sqrt{n}(F_n^Y(y) - F^Y(y))$, and $\mathbb{B}_{n,2} \circ F^X(y) = \sqrt{n}(F_n^X(y) - F^X(y))$.

Now let $Z = \beta_0^T X$ and $z = \beta_0^T x$. Then,

$$
\begin{aligned}
K_n(y) &= n \Big\{ F_n^Y(y) + \int f^\epsilon(y - z) F_n^Z(z) + n^{-1/2} \int h^T x f^\epsilon(y - \beta_0^T x) dF_n^X(x) \\
&\quad - \frac{1}{2} n^{-1} \int (h^T x)^2 (f^\epsilon)^{(1)}(\theta_n^*) dF_n^X(x) \Big\}^2.
\end{aligned}
$$

Define now the function

$$\mathbb{K}^2(y, h) = \left(\mathbb{B}_1 \circ F^Y(y) - \int \mathbb{B}_2 \circ F^Z(z) f^\epsilon(y - z) dz + \int h^T x f^\epsilon(y - \beta_0^T x) dF^X(x) \right)^2$$

for $y \in \mathbb{R}$, where $\mathbb{B}_1$ and $\mathbb{B}_2$ are two independent standard Brownian Bridges from $(0,0)$ to $(1,0)$ defined on the same probability space as $(X_1, \ldots, X_n)$ and $(Y_1, \ldots, Y_n)$ such that $\|\mathbb{B}_{n,1} - \mathbb{B}_1\|_\infty \to 0$ and $\|\mathbb{B}_{n,2} - \mathbb{B}_2\|_\infty \to 0$ almost surely. We show now that

$$\sup_{y \in \mathbb{R}, \|h\|_2 \le K} |\mathbb{K}_n^2(y,h) - \mathbb{K}^2(y,h)| = o_\mathbb{P}(1).$$

Let $E_1 = \mathbb{B}_{n,1} \circ F^Y - \mathbb{B}_1 \circ F^Y$ and $E_2 = \mathbb{B}_{n,2} \circ F^Z - \mathbb{B}_2 \circ F^Z$. Then,

$$
\begin{aligned}
\mathbb{K}_n^2(y,h) &= \left\{ \mathbb{B}_1 \circ F^Y(y) - \int \mathbb{B}_2 \circ F^Z(z) f^\epsilon(y-z)dz + \int h^T x f^\epsilon(y - \beta_0^T x)dF^X(x) \right. \\
&\quad + E_1(y) - \int E_2(z)f^\epsilon(y-z)dz + n^{-1/2} \int h^T x f^\epsilon(y - \beta_0^T x)d\mathbb{G}_n^X(x) \\
&\quad \left. - \frac{1}{2}n^{-1/2} \int (h^T x)^2 (f^\epsilon)^{(1)}(\theta_n^*)dF_n^X(x) \right\}^2 \\
&= \mathbb{K}^2(y,h) + \mathbb{K}(y,h)R_n(y,h) + R_n^2(y,h)
\end{aligned}
$$

where

$$
\begin{aligned}
R_n(y,h) &= E_1(y) - \int E_2(z)f^\epsilon(y-z)dz + n^{-1/2}\int h^T x f^\epsilon(y - \beta_0^T x)d\mathbb{G}_n^X(x) \\
&\quad - \frac{1}{2}n^{-1/2}\int (h^T x)^2 (f^\epsilon)^{(1)}(\theta_n^*)dF_n^X(x).
\end{aligned}
$$

Since $|\mathbb{K}(y,h)| \le 2 + KM \int \|\|x\|_2 dF^X(x) \equiv C$ we have

$$|\mathbb{K}_n^2(y,h) - \mathbb{K}^2(y,h)| \le C|R_n(y,h)| + R_n^2(y,h)$$

where

$$
\begin{aligned}
|R_n(y,h)| &\le \|\mathbb{B}_{n,1} - \mathbb{B}_1\|_\infty + \|\mathbb{B}_{n,2} - \mathbb{B}_2\|_\infty \\
&\quad + n^{-1/2}\left| \int h^T x f^\epsilon(y - \beta_0^T x)d\mathbb{G}_n^X(x) \right| + \frac{M}{2}n^{-1/2}\int (h^T x)^2 dF_n^X(x).
\end{aligned}
$$

We show now that $\sup_{\|h\|_2 \le K, y \in \mathbb{R}} \left| h^T \int x f^\epsilon(y - \beta_0^T x)d\mathbb{G}_n^X(x) \right| = O_\mathbb{P}(1)$. As it is easy to adapt the proof of Proposition 13 for any integer $m \ge 1$, we have that

$$\sup_{y \in \mathbb{R}} \left\| \int x f^\epsilon(y - \beta_0^T x)d\mathbb{G}_n^X(x) \right\|_2 = O_\mathbb{P}(1).$$

Now, for any $h \in \mathbb{R}^d : \|h\|_2 \le K$ we have that

$$
\begin{aligned}
\left| \int h^T x f^\epsilon(y - \beta_0^T x)d\mathbb{G}_n^X(x) \right| &= \left| h^T \int x f^\epsilon(y - \beta_0^T x)d\mathbb{G}_n^X(x) \right| \\
&\le \|h\|_2 \left\| \int x f^\epsilon(y - \beta_0^T x)d\mathbb{G}_n^X(x) \right\|_2, \\
&\le K \sup_y \left\| \int x f^\epsilon(y - \beta_0^T x)d\mathbb{G}_n^X(x) \right\|_2 \\
&= O_\mathbb{P}(1),
\end{aligned}
$$

where the last inequality follows from Cauchy-Schwarz. It follows that

$$n^{-1/2} \sup_{h \in B(\mathbf{0}, K), y \in \mathbb{R}} \left| \int h^T x f^\epsilon (y - \beta_0^T x) d\mathbb{G}_n^X(x) \right| = O_{\mathbb{P}}(n^{-1/2}).$$

To handle the last term, we apply the SLLN to show that

$$\int (h^T x)^2 dF_n^X(x) \to \int (h^T x)^2 dF^X(x) < \infty,$$

since $\int (h^T x)^2 dF^X(x) \le K^2 \int \|x\|_2^2 dF^X(x) < \infty$ by assumption. It follows that

$$
\begin{aligned}
\left| \mathbb{Q}_n(h) - \int \mathbb{K}^2(y, h) dF_n^Y(y) \right| &= \left| \int \mathbb{K}_n^2(y, h) dF_n^Y(y) - \int \mathbb{K}^2(y, h) dF_n^Y(y) \right| \\
&\le \sup_{y \in \mathbb{R}, \|h\|_2 \le K} |\mathbb{K}_n^2(y, h) - \mathbb{K}^2(y, h)| \to 0
\end{aligned}
$$

in probability. Now, we show that

$$\int \mathbb{K}^2(y, h) dF_n^Y(y) \Rightarrow \int \mathbb{K}^2(y, h) dF^Y(y)$$

in $\ell^\infty(B(\mathbf{0}, K))$. Here, we will apply Theorem 1.5.4 of van der Vaart and Wellner (1996). First, we show that the process

$$\left(\int \mathbb{K}^2(y, h) dF_n^Y(y) \right)_{n, h \in B(\mathbf{0}, K)} \equiv (Z_{n, h})_{n, h \in B(\mathbf{0}, K)}$$

is asymptotically tight. We have that for all $y \in \mathbb{R}, h \in B(\mathbf{0}, K)$

$$K^2(h, y) \le \left(\|\mathbb{B}_1\|_\infty + \|\mathbb{B}_2\|_\infty + KM \int \|x\|_2 dF^X(x) \right)^2.$$

Using a classical result about Brownian Bridge, it follows that for any $\delta > 0$ we can find $D_\delta > 0$ such that

$$\mathbb{P}(\|\mathbb{B}_1\|_\infty + \|\mathbb{B}_2\|_\infty \le D_\delta) \ge 1 - \delta.$$

This implies that with $D_\delta' = (D_\delta + KM \int \|x\|_2 dF^x(x))^2$

$$\mathbb{P}\left((Z_{n, h})_{n, h \in B(\mathbf{0}, K)} \in [0, D_\delta'] \right) \ge 1 - \delta$$

Next, denote $Z_h = \int \mathbb{K}^2(y, h) dF^Y(y)$. We will show that for any finite integer $k \ge 1$, and $h_1, \dots, h_k \in B(\mathbf{0}, K)$,

$$(Z_{n, h_1}, \dots, Z_{n, h_k}) \xrightarrow{d} (Z_{h_1}, \dots, Z_{h_k})$$

as $n \to \infty$. We will show the stronger result that the convergence above occurs almost surely. Let $\alpha_1, \ldots, \alpha_k \in \mathbb{R}$. The distribution function F^Y is continuous on $\mathbb{R}$ and hence

$$y \mapsto \sum_{j=1}^{k} \alpha_j \mathbb{K}^2(y, h_j)$$

$$= \sum_{j=1}^{k} \alpha_j \left(\mathbb{B}_1 \circ F^Y(y) + \int \mathbb{B}_2 \circ F^Z(z) f^\epsilon(y - z) dz + \int h_j^T x f^\epsilon(y - \beta_0^T x) dF^X(x) \right)^2$$

(A.19)

is continuous. In fact, it is known that the Brownian Bridge is continuous on $[0, 1]$. Also, the function

$$y \mapsto \int \mathbb{B}_2 \circ F^Z(z) f^\epsilon(y - z) dz = \int \mathbb{B}_2 \circ F^Z(y - z) f^\epsilon(z) dz$$

is continuous at every point $y \in \mathbb{R}$ by the dominated convergence Theorem: The function $y \mapsto \mathbb{B}_2 \circ F^Z(y - z)$ is continuous and

$$\left| \mathbb{B}_2 \circ F^Z(y - z) f^\epsilon(z) \right| \leq \|\mathbb{B}_2\|_\infty f^\epsilon(z)$$

is integrable. By the same theorem, we show that the function $y \mapsto \int h_j^T x f^\epsilon(y - \beta_0^T x) dF^X(x)$ is continuous: $y \mapsto f^\epsilon(y - \beta_0^T x)$ is continuous and $\left| h_j^T x f^\epsilon(y - \beta_0^T x) \right| \leq \|h\|_2 \|x\|_2 M$ is integrable with respect to F^X. Also, the function in (A.19) is bounded by

$$\sum_{j=1}^{k} |\alpha_j| \left(\|\mathbb{B}_1\|_\infty + \|\mathbb{B}_2\|_\infty + \|h_j\|_2 M \int \|x\|_2 dF^X(x) \right)^2.$$

Since F_n^Y converges weakly to F^Y almost surely and the continuity of the function, it follows that

$$\sum_{j=1}^{k} \alpha_j Z_{n, h_j} \overset{a.s.}{\to} \sum_{j=1}^{k} \alpha_j Z_{h_j}.$$

Since $\alpha_1, \ldots, \alpha_k$ are arbitrary it follows that

$$(Z_{n, h_1}, \ldots, Z_{n, h_k}) \overset{a.s.}{\to} (Z_{h_1}, \ldots, Z_{h_k})$$

as $n \to \infty$. This completes the proof that

$$\mathbb{Q}_n \Rightarrow \mathbb{Q}$$

in $\ell^\infty(B(\mathbf{0}, K))$. Lastly, let us put

$$\mathbb{W}(y) = \mathbb{B}_1 \circ F^Y(y) - \int \mathbb{B}_2 \circ F^Z(z) f^\epsilon(y - z) dz, \quad \text{and} \quad S(y) = \int x f^\epsilon(y - \beta_0^T x) dF^X(x).$$

47

Then, we can write

$$
\begin{aligned}
\mathbb{Q}(h) &= \int \left(\mathbb{W}(y) + h^T S(y) \right)^2 dF^Y(y) \\
&= \int \left(\mathbb{W}(y) + h^T S(y) \right) \left(\mathbb{W}(y) + S(y)^T h \right) dF^Y(y) \\
&= \int \mathbb{W}(y)^2 dF^Y(y) + 2 \int \mathbb{W}(y) S(y)^T h dF^Y(y) + h^T \int S(y) S(y)^T dF^Y(y) h
\end{aligned}
$$

which is a quadratic form. It admits a unique minimum if the matrix

$$
\int S(y) S(y)^T dF^Y(y) = \int \left(x f^\epsilon(y - \beta_0^T x) dF^X(x) \right) \left(x f^\epsilon(y - \beta_0^T x) dF^X(x) \right)^T dF^Y(y) = U
$$

is positive definite, in which case the unique minimizer h^0 solves the equation

$$
U h^0 = - \int \mathbb{W}(y) S(y) dF^Y(y) = V,
$$

or equivalently $h^0 = U^{-1} V$. ∎

A.9 Proof of Theorem 7

Proof From Proposition 14 we know that there exists a constant $C > 0$ such that

$$
\phi_n(\delta) \leq C \left(\delta + \frac{\delta}{\sqrt{n}} + \frac{1}{\sqrt{n}} \right)
$$

for all $\delta \in (0, 1)$. Now, $\delta \mapsto \phi_n(\delta)/\delta$ is decreasing. Also, if we take $r_n = \sqrt{n}$ we get

$$
r_n^2 \phi_n(1/r_n) = C \, n \left(\frac{1}{\sqrt{n}} + \frac{1}{n} + \frac{1}{\sqrt{n}} \right) = C \left(2\sqrt{n} + \frac{1}{\sqrt{n}} \right) \leq 4C\sqrt{n}
$$

for n large enough. Combining this with Proposition 15 it follows from Theorem 3.2.5 of van der Vaart and Wellner (1996) that

$$
\sqrt{n}(\hat{\beta}_n - \beta_0) = O_{\mathbb{P}}(1).
$$

∎

A.10 Proof of Theorem 8

Proof The arguments are the same as those used in the proof of Theorem 7 except that we replace X by $\tilde{X}$ as defined in the statement of the theorem. ∎

A.11 Proof of Proposition 9

Proof Let $\beta_m^0 = \tilde{\beta}_m \frac{\|\beta_0\|_{2,\widehat{\Sigma}}}{\|\tilde{\beta}_m\|_{2,\widehat{\Sigma}}}$. Note that with probability one, we have

$$\|\tilde{\beta}_m - \beta_0\| \geq \|\beta_m^0 - \beta_0\|.$$

We can rewrite $\tilde{\beta}_{n,m}$ as

$$\tilde{\beta}_{n,m} = \beta_m^0 \frac{r_n}{\|\beta_0\|_{2,\widehat{\Sigma}}}.$$

We have

$$
\begin{aligned}
\frac{\mathbb{E}[\|\tilde{\beta}_{n,m} - \beta_0\|_{2,\widehat{\Sigma}}^2]}{\mathbb{E}[\|\tilde{\beta}_m - \beta_0\|_{2,\widehat{\Sigma}}^2]} &= \frac{\mathbb{E}[\|\beta_m^0 \frac{r_n}{\|\beta_0\|_{2,\widehat{\Sigma}}} - \beta_0\|_{2,\widehat{\Sigma}}^2]}{\mathbb{E}[\|\tilde{\beta}_m - \beta_0\|_{2,\widehat{\Sigma}}^2]} \\
&= \frac{\mathbb{E}[\|\beta_m^0 - \beta_0 + \beta_m^0(\frac{r_n}{\|\beta_0\|_{2,\widehat{\Sigma}}} - 1)\|_{2,\widehat{\Sigma}}^2]}{\mathbb{E}[\|\tilde{\beta}_m - \beta_0\|_{2,\widehat{\Sigma}}^2]} \\
&= \frac{\mathbb{E}[\|\beta_m^0 - \beta_0\|_{2,\widehat{\Sigma}}^2] + 2T_2 + T_3}{\mathbb{E}[\|\tilde{\beta}_m - \beta_0\|_{2,\widehat{\Sigma}}^2]},
\end{aligned}
$$

where

$$T_2 = \mathbb{E}[(\frac{r_n - \|\beta_0\|_{2,\widehat{\Sigma}}}{\|\beta_0\|_{2,\widehat{\Sigma}}})(\beta_m^0 - \beta_0)^T \widehat{\Sigma} \beta_m^0]$$

and

$$T_3 = \mathbb{E}[\frac{(r_n - \|\beta_0\|_{2,\widehat{\Sigma}})^2}{\|\beta_0\|_{2,\widehat{\Sigma}}^2}].$$

To complete the proof, it is enough to show that there exist constants $C_1, C_2,$ and C_3 such that

$$\frac{C_1}{m} \leq \mathbb{E}[\|\tilde{\beta}_m - \beta_0\|_{2,\widehat{\Sigma}}^2] \leq \frac{C_2}{m}, \tag{A.20}$$

and

$$T_3 \leq \frac{C_3}{n}. \tag{A.21}$$

With (A.20), and (A.21) we have

$$\frac{2T_2}{\mathbb{E}[\|\hat{\beta}_m - \beta_0\|_{2,\widehat{\Sigma}}^2]} \lesssim \sqrt{\frac{m}{n}},$$

and

$$\frac{T_3}{\mathbb{E}[\|\hat{\beta}_m - \beta_0\|_{2,\widehat{\Sigma}}^2]} \lesssim \frac{m}{n}$$

49

and therefore, we can conclude that there exists $D_1, D_2 > 0$ such that

$$\frac{\mathbb{E}[\|\tilde{\beta}_{n,m} - \beta_0\|_{2,\widehat{\Sigma}}^2]}{\mathbb{E}[\|\tilde{\beta}_m - \beta_0\|_{2,\widehat{\Sigma}}^2]} \leq \frac{\mathbb{E}[\|\beta_m^0 - \beta_0\|_{2,\widehat{\Sigma}}^2]}{\mathbb{E}[\|\tilde{\beta}_m - \beta_0\|_{2,\widehat{\Sigma}}^2]} + D_1\sqrt{\frac{m}{n}} + D_2\frac{m}{n}$$

$$< 1 + D_1\sqrt{\frac{m}{n}} + D_2\frac{m}{n}.$$

Therefore for $m = o(n)$ we have

$$\frac{\mathbb{E}[\|\tilde{\beta}_{n,m} - \beta_0\|_{2,\widehat{\Sigma}}^2]}{\mathbb{E}[\|\tilde{\beta}_m - \beta_0\|_{2,\widehat{\Sigma}}^2]} < 1.$$

To complete the proof, we need to prove (A.20), and (A.21). Let's start by proving (A.20). First note that

$$\tilde{\beta}_m = \beta_0 + [\tilde{X}^T\tilde{X}]^{-1}\tilde{X}^T\tilde{\epsilon},$$

where $\tilde{X}$ is a $m \times d$ matrix such that its rows are $\tilde{X}_i$ from the matched data. Also $\tilde{\epsilon} = [\tilde{\epsilon}_1, \cdots, \tilde{\epsilon}_m]$ with $\tilde{\epsilon}_i = \tilde{Y}_i - \beta_0^T\tilde{X}_i$. Let $M = [\tilde{X}^T\tilde{X}]^{-1}\tilde{X}^T$ then it follows that

$$\|\tilde{\beta}_m - \beta_0\|_{2,\widehat{\Sigma}}^2 = \tilde{\epsilon}M^T\widehat{\Sigma}M\tilde{\epsilon}$$

$$= tr(\tilde{\epsilon}M^T\widehat{\Sigma}M\tilde{\epsilon})$$

$$= tr(\widehat{\Sigma}M\tilde{\epsilon}\tilde{\epsilon}M^T),$$

where in the above equality, we are using the fact about trace that $tr(AB) = tr(BA)$. Now let $A = M\tilde{\epsilon}\tilde{\epsilon}^T M^T$, then we have

$$tr(\widehat{\Sigma}A) = \sum_{i=1}^{d}(\widehat{\Sigma}A)_{ii}$$

$$= \sum_{i=1}^{d}\sum_{j=1}^{d}\widehat{\Sigma}_{ij}A_{ji}.$$

Note that A is based on the matched sample $\{(\tilde{X}_i, \tilde{Y}_i)\}_{i=1}^{m}$ and $\widehat{\Sigma}$ is computed using the unmatched sample $\{Y_i\}_{i=1}^{n}$ and $\{X_i\}_{i=1}^{n}$. Therefore, A and $\widehat{\Sigma}$ are independent of each other. This gives us

$$\mathbb{E}[\|\tilde{\beta}_m - \beta_0\|_{2,\widehat{\Sigma}}^2] = \sum_{i=1}^{d}\sum_{j=1}^{d}\mathbb{E}[\widehat{\Sigma}_{ij}]\mathbb{E}[A_{ji}]$$

$$= \sum_{i=1}^{d}\sum_{j=1}^{d}\Sigma_{ij}\mathbb{E}[A_{ji}]$$

$$= tr(\Sigma\mathbb{E}[A])$$

According to Kleinman and Athans (1968) we have

$$\lambda_d(\Sigma) tr(\mathbb{E}[A]) \le tr(\Sigma \mathbb{E}[A]) \le \lambda_1(\Sigma) tr(\mathbb{E}[A]),$$

where $\lambda_d(\Sigma)$ and $\lambda_1(\Sigma)$ are respectively the smallest and largest eigenvalues of Σ. To bound $tr(\mathbb{E}[A]) = \mathbb{E}[tr(A)]$, let $K = (\tilde{\epsilon}\tilde{\epsilon}^T)$ and $P = M^T M$ and note that

$$\begin{aligned}
\mathbb{E}[tr(A)] &= \mathbb{E}[tr(M\tilde{\epsilon}\tilde{\epsilon}^T M^T)] \\
&= \mathbb{E}[tr(\tilde{\epsilon}\tilde{\epsilon}^T M^T M)] \\
&= \mathbb{E}[tr(KP)] \\
&= \sum_{i=1}^m \sum_{j=1}^m \mathbb{E}[K_{ij} P_{ji}] \\
&= Var(\epsilon)\mathbb{E}[tr(P)] \\
&= Var(\epsilon)\mathbb{E}[tr([\tilde{X}^T \tilde{X}]^{-1})] \\
&= \frac{Var(\epsilon)}{m}\mathbb{E}[tr([\frac{\tilde{X}^T \tilde{X}}{m}]^{-1})].
\end{aligned}$$

Therefore

$$\frac{Var(\epsilon)}{m}\mathbb{E}[\frac{1}{\lambda_1(\frac{\tilde{X}^T\tilde{X}}{m})}] \le \mathbb{E}[tr(A)] \le d\frac{Var(\epsilon)}{m}\mathbb{E}[\frac{1}{\lambda_d(\frac{\tilde{X}^T\tilde{X}}{m})}].$$

By Weyl's theorem, we have

$$\max_{j=1,\cdots,d}|\lambda_j(\Sigma) - \lambda_j(\frac{1}{m}\tilde{X}^T\tilde{X})| \le \|\Sigma - \frac{1}{m}\tilde{X}^T\tilde{X}\|_2.$$

On the other hand, with high probability, we have

$$\|\Sigma - \frac{1}{m}\tilde{X}^T\tilde{X}\|_2 \le \lambda_1(\Sigma)(2\sqrt{\frac{d}{m}} + 2\delta + (\sqrt{\frac{d}{m}} + \delta)^2)$$

for all $\delta > 0$. Putting these together gives us a high probability

$$\frac{Var(\epsilon)}{m}(\lambda_1(\Sigma) - C_1\sqrt{\frac{d}{m}})^{-1} \le \mathbb{E}[tr(A)] \le d\frac{Var(\epsilon)}{m}(\lambda_d(\Sigma) + C_2\sqrt{\frac{d}{m}})^{-1},$$

where C_1, C_2 are constants which depends on $\lambda_1(\Sigma)$ and δ. This gives us

$$\mathbb{E}[\|\tilde{\beta}_m - \beta_0\|_{2,\widehat{\Sigma}}^2] \in [\frac{Var(\epsilon)\lambda_d(\Sigma)}{m(\lambda_1(\Sigma) - C_1\sqrt{\frac{d}{m}})}, \frac{dVar(\epsilon)\lambda_1(\Sigma)}{m(\lambda_d(\Sigma) + C_2\sqrt{\frac{d}{m}})}],$$

and this conclude (A.20).

To show (A.21) for simplicity let $\hat{r}_0 = \|\beta_0\|_{2,\widehat{\Sigma}}$ and note that

$$\begin{aligned}
\mathbb{E}[(r_n - \hat{r}_0)^2] &= \mathbb{E}[(r_n - r_0 + r_0 - \hat{r}_0)^2] \\
&\le 2\mathbb{E}[(r_n - r_0)^2] + 2\mathbb{E}[(r_0 - \hat{r}_0)^2],
\end{aligned}$$

51

where $r_0^2 = \|\beta_0\|_{2,\Sigma}^2 = \text{Var}(Y) - \text{Var}(\epsilon)$. Assuming that $m_2 = \mathbb{E}[Y^2]$, $m_4 = \mathbb{E}[Y^4]$ exist we have (Exercise 5.8 in Casella and Berger (1990))

$$\mathbb{E}[(r_n^2 - r_0^2)^2] \leq \frac{1}{n}(m_4 - \frac{n-3}{n-1}m_2^2).$$

Hence

$$
\begin{aligned}
\mathbb{E}[(r_n - r_0)^2] &\leq \frac{1}{r_0^2}\mathbb{E}[(r_n^2 - r_0^2)^2] \\
&\leq \frac{1}{nr_0^2}(m_4 - \frac{n-3}{n-1}m_2^2).
\end{aligned}
$$

Since $\|\cdot\|_{2,\Gamma}$ is a continuous function of Γ and since $\|\widehat{\Sigma} - \Sigma\| = O(n^{-1})$ we have

$$
\begin{aligned}
\mathbb{E}[(r_0 - \hat{r}_0)^2] &= \mathbb{E}[(\|\beta_0\|_{2,\Sigma} - \|\beta_0\|_{2,\widehat{\Sigma}})^2] \\
&\lesssim \frac{1}{n}.
\end{aligned}
$$

Furthermore for large enough n we have $\hat{r}_0/r_0 \geq 1/2$, therefore

$$T_3 = \mathbb{E}[\frac{(r_n - \hat{r}_0)^2}{\hat{r}_0^2}] = \mathbb{E}[\frac{(r_n - \hat{r}_0)^2}{r_0^2}\frac{r_0^2}{\hat{r}_0^2}] \lesssim \frac{1}{n},$$

and this completes the proof of (A.21). ∎

A.12 Uniqueness in Example 1

We show that for $Y \stackrel{d}{=} \beta_0^T X + \epsilon$ such that

$$X = [X^1, X^2]^T \in \mathbb{R}^2, X^1 \sim \mathcal{N}(0,1), X^2 \sim \text{Exp}(1), \epsilon \in \mathcal{N}(0,1), \beta_0^T = [1,2]$$

we have $\mathcal{B}_0 = \{\beta_0\}$. We prove this by contradiction.

Proof Assume $|\mathcal{B}_0| > 1$. Take $\beta \neq \tilde{\beta} \in \mathcal{B}_0$. Let $\beta = [\beta_1, \beta_2]^T$ and $\tilde{\beta} = [\tilde{\beta}_1, \tilde{\beta}_2]^T$. Note that in this case both β and $\tilde{\beta}$ lead to the same distribution for Y and hence using the moment generating function of Y we have

$$\exp(\mu\beta_1 t + \frac{\beta_1^2 t^2}{2})\frac{1}{1-\beta_2 t} = \exp(\mu\tilde{\beta}_1 t + \frac{\tilde{\beta}_1^2 t^2}{2})\frac{1}{1-\tilde{\beta}_2 t} \tag{A.22}$$

for all t such that $\beta_2 t < 1$ and $\tilde{\beta}_2 t < 1$. First suppose $\beta_2 = 0$. By rewriting (A.22), we have

$$\exp(\mu(\beta_1 - \tilde{\beta}_1)t + \frac{(\beta_1^2 - \tilde{\beta}_1^2)t^2}{2}) = \frac{1}{1-\tilde{\beta}_2 t}$$

for all t such that $\tilde{\beta}_2 t < 1$. This equality holds if and only if $\tilde{\beta}_2 = 0$ and $\beta_1 = \tilde{\beta}_1$. By symmetry, we reach the same conclusion assuming $\tilde{\beta}_2 = 0$. Now assume $\beta_2 \neq 0$ and $\tilde{\beta}_2 \neq 0$. Without loss of generality assume $\beta_2 < \tilde{\beta}_2$. Then we rewrite (A.22) as

$$\exp(\mu(\beta_1 - \tilde{\beta}_1)t + \frac{(\beta_1^2 - \tilde{\beta}_1^2)t^2}{2})\frac{1-\tilde{\beta}_2 t}{1-\beta_2 t} = 1 \tag{A.23}$$

for all t such that $\beta_2 t < 1$ and $\tilde{\beta}_2 t < 1$. We study this within the following cases:

Case 1: $\beta_2 < \tilde{\beta}_2 < 0$. In this case we have $|\tilde{\beta}_2| < |\beta_2|$ and hence the equality in (A.23) holds for $t > -1/|\beta_2|$. Now note that

$$\lim_{t \searrow -1/|\beta_2|} \exp(\mu(\beta_1 - \tilde{\beta}_1)t + \frac{(\beta_1^2 - \tilde{\beta}_1^2)t^2}{2})\frac{1 - \tilde{\beta}_2 t}{1 - \beta_2 t} = \infty,$$

which is contradiction with (A.23).

Case 2: $\beta_2 < 0 < \tilde{\beta}_2$. In this case (A.23) must hold for $-1/|\beta_2| < t < 1/\tilde{\beta}_2$. Taking the limit $t \searrow -1/|\beta_2|$ again we reach contradiction.

Case 3: $0 < \beta_2 < \tilde{\beta}_2$. In this case (A.23) must hold for $t < 1/\tilde{\beta}_2$. Taking the $t \nearrow 1/\beta_2$ implies

$$\lim_{t \nearrow 1/|\tilde{\beta}_2|} \exp(\mu(\beta_1 - \tilde{\beta}_1)t + \frac{(\beta_1^2 - \tilde{\beta}_1^2)t^2}{2})\frac{1 - \tilde{\beta}_2 t}{1 - \beta_2 t} = 0,$$

which is again a contradiction. Therefore we must have $\beta_2 = \tilde{\beta}_2$ which implies

$$\exp(\mu(\beta_1 - \tilde{\beta}_1)t + \frac{(\beta_1^2 - \tilde{\beta}_1^2)t^2}{2}) = 1$$

for all $\beta_2 t < 1$ which means

$$\mu(\beta_1 - \tilde{\beta}_1)t + \frac{(\beta_1^2 - \tilde{\beta}_1^2)t^2}{2} = 0.$$

Since $\mu \neq 0$ this implies $\beta_1 = \tilde{\beta}_1$. ∎

References

Abubakar Abid, Ada Poon, and James Zou. Linear regression with shuffled labels. *arXiv preprint arXiv:1705.01342*, 2017.

Fadoua Balabdaoui, Charles R Doss, and Cécile Durot. Unlinked monotone regression. *The Journal of Machine Learning Research*, 22(1):7766–7825, 2021.

Alexandra Carpentier and Teresa Schlüter. Learning relationships between data obtained independently. In *Proceedings of the 19th International Conference on Artificial Intelligence and Statistics*, pages 658–666, 2016.

George Casella and Roger L. Berger. *Statistical inference.* The Wadsworth & Brooks/Cole Statistics/Probability Series. Wadsworth & Brooks/Cole Advanced Books & Software, Pacific Grove, CA, 1990.

William W Cohen and Jacob Richman. Learning to match and cluster large high-dimensional data sets for data integration. In *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 475–480, 2002.

Xavier D'Haultfoeuille, Christophe Gaillac, and Arnaud Maurel. Partially linear models under data combination. Technical report, National Bureau of Economic Research, 2022.

A. Dvoretzky, J. Kiefer, and J. Wolfowitz. Asymptotic minimax character of the sample distribution function and of the classical multinomial estimator. *Ann. Math. Statist.*, 27: 642–669, 1956.

David Edelman. Estimation of the mixing distribution for a normal mean with applications to the compound decision problem. *Ann. Statist.*, 16(4):1609–1622, 1988.

Jianqing Fan. On the optimal rates of convergence for nonparametric deconvolution problems. *Ann. Statist.*, 19(3):1257–1272, 1991.

Daniel J Hsu, Kevin Shi, and Xiaorui Sun. Linear regression without correspondence. *Advances in Neural Information Processing Systems*, 30, 2017.

David L. Kleinman and Michael Athans. The design of suboptimal linear time-varying systems. *IEEE Trans. Automatic Control*, AC-13:150–159, 1968.

Jan Meis and Enno Mammen. Uncoupled isotonic regression with discrete errors. In *Advances in Contemporary Statistics and Econometrics*, pages 123–135. Springer, 2021.

Alvaro E Monge, Charles Elkan, et al. The field matching problem: algorithms and applications. In *Kdd*, volume 2, pages 267–270, 1996.

John A Nelder and RA Mead. Simplex method for function minimization, the computer journal, 7. 1965.

Claudia Olivetti and M Daniele Paserman. In the name of the son (and the daughter): Intergenerational mobility in the united states, 1850-1940. *American Economic Review*, 105(8):2695–2724, 2015.

Ashwin Pananjady, Martin J Wainwright, and Thomas A Courtade. Linear regression with shuffled data: statistical and computational limits of permutation recovery. *IEEE Trans. Inform. Theory*, 64(5):3286–3300, 2018.

Philippe Rigollet and Jonathan Weed. Uncoupled isotonic regression via minimum Wasserstein deconvolution. *Inf. Inference*, 8(4):691–717, 2019.

Steven Ruggles, J Trent Alexander, Katie Genadek, Ronald Goeken, Matthew B Schroeder, Matthew Sobek, et al. Integrated public use microdata series: Version 5.0 [machinereadable database]. *Minneapolis: University of Minnesota*, 42, 2010.

S. J. Sheather and M. C. Jones. A reliable data-based bandwidth selection method for kernel density estimation. *J. Roy. Statist. Soc. Ser. B*, 53(3):683–690, 1991.

Martin Slawski and Emanuel Ben-David. Linear regression with sparsely permuted data. *Electron. J. Stat.*, 13(1):1–36, 2019.

Martin Slawski and Bodhisattva Sen. Permuted and unlinked monotone regression in $\mathbb{R}^d$: an approach based on mixture modeling and optimal transport. *arXiv preprint arXiv:2201.03528*, 2022.

Martin Slawski, Emanuel Ben-David, and Ping Li. A two-stage approach to multivariate linear regression with sparsely mismatched data. *J. Mach. Learn. Res.*, 21(204):1–42, 2020.

Martin Slawski, Guoqing Diao, and Emanuel Ben-David. A pseudo-likelihood approach to linear regression with partially shuffled data. *Journal of Computational and Graphical Statistics*, 30(4):991–1003, 2021.

Manolis C. Tsakiris and Liangzu Peng. Homomorphic sensing. *Proceedings of the 36th International Conference on Machine Learning*, pages 6335–6344, 2019.

Manolis C. Tsakiris, Liangzu Peng, Aldo Conca, Laurent Kneip, Yuanming Shi, and Hayoung Choi. An algebraic-geometric approach for linear regression without correspondences. *IEEE Transactions on Information Theory*, 66(8):5130–5144, 2020.

Jayakrishnan Unnikrishnan, Saeid Haghighatshoar, and Martin Vetterli. Unlabeled sensing with random linear measurements. *IEEE Trans. Inform. Theory*, 64(5):3237–3253, 2018.

A. W. van der Vaart. *Asymptotic statistics*, volume 3 of *Cambridge Series in Statistical and Probabilistic Mathematics*. Cambridge University Press, Cambridge, 1998.

Aad W. van der Vaart and Jon A. Wellner. *Weak convergence and empirical processes.* Springer Series in Statistics. Springer-Verlag, 1996.

Volker Walter and Dieter Fritsch. Matching spatial data sets: a statistical approach. *International Journal of geographical information science*, 13(5):445–473, 1999.

Hang Zhang, Martin Slawski, and Ping Li. The benefits of diversity: Permutation recovery in unlabeled sensing from multiple measurement vectors. *IEEE Transactions on Information Theory*, 2021.