Journal of Machine Learning Research 27 (2026) 1-39 Submitted 3/25; Revised 5/26; Published 6/26

Deconvolution in unlinked linear models

Fadoua Balabdaoui FADOUA.BALABDAOUI@STAT.MATH.ETHZ.CH
Seminar for Statistics, Department of Mathematics, ETH Zurich

Antonio Di Noia ANTONIO.DINOIA@STAT.MATH.ETHZ.CH
Seminar for Statistics, Department of Mathematics, ETH Zurich
Faculty of Economics, Euler Institute, Università della Svizzera italiana

Cécile Durot CECILE.DUROT@PARISNANTERRE.FR
MODAL'X, UMR CNRS 9023, Université Paris Nanterre

Editor: Sivan Sabato

Abstract

Unlinked regression, in which covariates and responses are observed separately without known correspondence, has recently gained increasing attention. Deconvolution, on the other hand, is a fundamental and challenging problem in nonparametric statistics with the aim of estimating the distribution of a latent random variable Z based on observations contaminated by some additive noise. The complexity of this task is heavily influenced by the smoothness of the noise distribution and often leads to slow estimation rates. In this paper, we combine the recent unlinked linear regression problem with the classical deconvolution framework. Specifically, we study nonparametric deconvolution under the assumption that Z is a linear function of an observable multidimensional covariate. This structural constraint allows us to introduce a nonparametric estimator of the distribution of Z which achieves the parametric rate of convergence in the Wasserstein distance of order 1, where the smoothness of the noise does not affect the rate. Furthermore, we introduce nonparametric estimators for the unconditional density of Z and the conditional density of Z given an observed response. This allows us to study the problem of estimating the value of the latent linear predictor, whose link to the observed response is not accessible. Through several simulations, we illustrate the fast convergence rate of our deconvolution estimator and the performance of the proposed conditional estimators of the latent predictor in different simulation scenarios.

Keywords: deconvolution, inverse problem, unlinked linear regression, rate of convergence, Wasserstein distance, empirical process theory

1. Introduction

1.1 Background and contributions

The unlinked or unmatched regression has recently regained attention in the statistics and machine learning literature. In this problem, we are given independent copies of covariates $X_1, \dots, X_n$ and responses $Y_1, \dots, Y_m$ with the same distribution as random variables X and Y respectively. Here, the sample sizes $n \geq 1$ and $m \geq 1$ are not necessarily equal. Assuming that $Y \stackrel{d}{=} f(X) + \epsilon$, the goal is to reconstruct the regression function f under the condition that the noise ϵ has a known distribution. This problem arises in various fields, including sociology, economics, data privacy, computer vision, biology, and multi-tracking scenarios;

see for example Durot and Mukherjee (2024) and Slawski and Sen (2024) for a detailed account of the applications of unlinked regression.

Theoretical developments in unlinked regression often leverage techniques from the deconvolution literature. For instance, the case where f is monotone has been studied under various assumptions on noise distribution using deconvolution techniques (Carpentier and Schlüter, 2016; Balabdaoui et al., 2021; Rigollet and Weed, 2019; Meis and Mammen, 2020). In the unlinked linear regression setting, Azadkia and Balabdaoui (2024) proposed a deconvolution least squares estimator for parameter estimation, highlighting that identifiability issues may arise. Under some technical conditions on the noise and covariate distributions, consistency and asymptotic normality were established. The linear case has been previously addressed only in a closely related sub-problem, namely the *permuted regression* or *shuffled regression*. In the latter setting, there exists a link between covariates and responses, however, this link is not known due to some unknown shuffling of the covariates (or equivalently the responses). The main focus in the existing literature on permuted regression is either exact or approximate recovery of the unknown permutation; see Slawski et al. (2020), Hsu et al. (2017), Pananjady et al. (2017), Zhang et al. (2021), Slawski et al. (2021), Slawski and Ben-David (2019), Tsakiris et al. (2020) and Unnikrishnan et al. (2018). The gap between unlinked and permuted regression becomes particularly pronounced when the noise vanishes at a rate smaller than a threshold of order close to $n^{-1/2}$, as formalized in Durot and Mukherjee (2024), where a phase transition in minimax rates under the Wasserstein distance was established: the minimax rates of estimation are the same in the two models when the noise level is larger than the threshold (in particular, under fixed noise level) and becomes smaller in the permuted regression than in unlinked regression when the noise level is of smaller order than $n^{-1/2}$.

Deconvolution, on the other hand, is an old fundamental inverse problem in nonparametric statistics that arises in a wide range of applied fields, including econometrics, biometrics, medical statistics, image analysis, and signal processing. The main objective of deconvolution is to estimate an unknown density when only observations contaminated with additive noise are available. Note that it is often the case that the noise distribution is assumed to be known. Formally, given a latent random variable Z, we seek to estimate its distribution using observed data from $Y = Z + \epsilon$, where ϵ represents the additive noise, assumed to be independent of Z. The estimation task requires inverting the convolution induced by ϵ to recover the unknown distribution of Z.

Despite its seemingly straightforward formulation, the deconvolution problem is notably challenging. Its complexity, often measured in terms of the convergence rates in a chosen metric (e.g., L^p or Wasserstein distance), is driven by the smoothness properties of the noise distribution. Carroll and Hall (1988) established optimal pointwise estimation rates in the Gaussian case showing that they are very slow, specifically, a power of the logarithm of the sample size. The seminal papers Fan (1991b, 1992, 1993) generalized this finding by showing that in the case where the characteristic function of the noise ϵ decays exponentially (as for the Gaussian noise), known as the *supersmooth* case, the pointwise and global optimal rates in the L^p-norm are also a power of the logarithm of the sample size. In Fan (1991b, 1993), it is shown that for the *ordinary smooth* case, i.e. where the characteristic function decays polynomially, the pointwise and global optimal rates in the L^p-norm are faster, specifically, a power of the sample size.

The statistical literature on deconvolution is very extensive. Here, we shall provide only a few references thereof, and refer the reader to the book Meister (2009) for a nice overview. A variety of methodologies have been proposed in deconvolution problems, including kernel-based estimators which are based on the empirical characteristic function and Fourier inversion (Liu and Taylor, 1989; Stefanski and Carroll, 1990; Fan, 1991a,b; Es and Uh, 2005). Extensions to settings where the noise distribution is unknown, partially or fully, have been explored e.g. in Butucea and Matias (2005); Meister (2006); Delaigle and Hall (2014). Alternative estimation techniques include series estimators (Pensky and Vidakovic, 1999; Lounici and Nickl, 2011; Carrasco and Florens, 2011; Carroll and Hall, 2004; Hall and Qiu, 2005), nonparametric maximum likelihood methods (Groeneboom and Jongbloed, 2003; Guan, 2021), and penalization approaches (Comte et al., 2006, 2007).

Our contribution in this paper bridges nonparametric deconvolution with the unlinked linear regression. Specifically, we focus on estimating the distribution of $Z \stackrel{d}{=} \beta_0^\top X$ (for some non-unique β_0) from realizations of $Z + \epsilon$ and of the covariate X whose distribution is unknown. This framework extends the classical deconvolution problem through incorporating the structure of the unlinked linear model. We develop an estimator based on the deconvolution least squares approach from Azadkia and Balabdaoui (2024) and the empirical measure of $X_1, \ldots, X_n$. Using empirical process techniques, we show that this additional structure allows us to obtain the parametric rate of convergence in the Wasserstein distance for estimating the distribution of the latent variable Z, where the smoothness of the noise does not affect the rate. We emphasize that the Wasserstein distance is a natural risk measure to consider for estimating a distribution, especially in cases where no regularity assumption is made. See e.g. Caillerie et al. (2013); Dedecker and Michel (2013); Dedecker et al. (2015). In contrast, risk measures based for example on some L^p distance between probability density functions can only be used for distributions that are assumed to have a probability density. Such assumptions are not required in this paper for the distribution of Z to achieve the fast convergence rate in the sense of the Wasserstein risk. It is of crucial importance to emphasize that although the model incorporates a finite-dimensional parameter, it does not fall within the framework of semi-parametric models. In fact, our goal is not the estimate the finite-dimensional regression vector β_0 but rather an infinite-dimensional object: the distribution of $\beta_0^\top X$.

We believe that our results are valuable for at least two reasons: (i) introducing a structured nonparametric deconvolution problem where the parametric rate of convergence is attainable, which is uncommon in classical deconvolution problems, and (ii) being able to achieve such rate of convergence without requiring restrictive assumptions on the noise distribution. Another important contribution we make in this work is to propose nonparametric estimators for the unconditional density of Z and for the conditional density of $Z|Y = y$ for any observed response value y. Such estimators can be employed to make inference about the latent linear predictor linked to this specific realisation of the response variable.

1.2 The setting

Let X be a random vector in $\mathbb{R}^d$ for $d \geq 1$. Consider Y to be a random variable defined on the same probability space as X such that

$$Y = \beta_0^\top X + \epsilon \tag{1}$$

where ϵ is the noise random variable with known distribution and independent of X, and $\beta_0 \in \mathbb{R}^d$ is a deterministic vector of unknown coefficients. In this work, we assume that we only have access to $\mathcal{X}_n = \{X_1, \ldots, X_n\}$, a set of independent and identically distributed (i.i.d.) copies of X and $\mathcal{Y}_m = \{Y_1, ..., Y_m\}$, a set of i.i.d. copies of Y. Here, the data $\mathcal{X}_n$ and $\mathcal{Y}_m$ can potentially be collected from different sources (i.e., they can be independent), or originally measured for the same individuals but the link between the responses and covariates is missing, or anything in between (i.e., there is a non-trivial intersection). Note that it is possible that $n \neq m$. Still, for ease of exposition, we restrict ourselves to the case of samples of the same size n.

Let $\mathcal{B}_0$ be the set of all vectors $\beta \in \mathbb{R}^d$ such that the unlinked linear model in (1) holds. In other words,

$$\mathcal{B}_0 = \{\beta \in \mathbb{R}^d,\ \beta^\top X \overset{d}{=} \beta_0^\top X\}.$$

We will not assume that β_0 is identifiable, which means that we might be in the case where $|\mathcal{B}_0| > 1$. Under the assumption that (X, ϵ) is independent of $\{\mathcal{X}_n, \mathcal{Y}_n\}$, our goal is to predict the distribution μ_0 of the conditional expectation $\mathbb{E}(Y|X) = \beta_0^\top X,\ \beta_0 \in \mathcal{B}_0$, of a future response Y taken from (1) given observations $\{\mathcal{X}_n, \mathcal{Y}_n\}$ and using our knowledge of the noise distribution μ_ϵ. We emphasize that the two samples $\mathcal{X}_n$ and $\mathcal{Y}_n$ are not linked. We use the word *future* to stress the fact that Y is not observed yet. To do that, we consider $\widehat{Z}_n := \widehat{\beta}_n^\top X^*$ where $\widehat{\beta}_n$ is a deconvolution least squares estimator as defined in Azadkia and Balabdaoui (2024), and X^* is built as follows: Define J_n to be a random variable which is independent of $\{\mathcal{X}_n, \mathcal{Y}_n\}$ and uniformly distributed on $\{1, \ldots, n\}$ and let $X^* = X_{J_n}$. Note that X^* is drawn from the empirical distribution of the observations in sample $\mathcal{X}_n$. This also means that $\widehat{Z}_n$ is a random draw from the empirical distribution of $\widehat{\beta}_n^\top X_1, \ldots, \widehat{\beta}_n^\top X_n$. We consider the distribution of $\widehat{Z}_n$ as an estimator of the distribution μ_0.

1.3 Notation

All the random variables involved are defined on the same probability space $(\Omega, \mathcal{A}, \mathbb{P})$ that is equipped with a probability measure denoted by $\mathbb{P}$. We denote the corresponding expectation by $\mathbb{E}$. To alleviate notation, we denote by P and E the conditional probability and expectation given $\{\mathcal{X}_n, \mathcal{Y}_n\}$. Moreover, we will allow ourselves to write $\mathbb{P}(A)$ or $\mathbb{E}(X)$ without justifying that the set A belongs to the sigma-algebra $\mathcal{A}$ or that X is a measurable function: In case $A \subset \Omega$ is not in $\mathcal{A}$ or the function X is not measurable, then $\mathbb{P}$ and $\mathbb{E}$ must be interpreted as outer probability and outer integral as defined in van der Vaart and Wellner (2023).

For a given $\beta \in \mathbb{R}^d$ let

$$d(\beta, \mathcal{B}_0) := \inf_{\beta_0 \in \mathcal{B}_0} \|\beta - \beta_0\|$$

denote the Euclidean distance of β to $\mathcal{B}_0$. Moreover, for a given $\beta \in \mathbb{R}^d$, we will use the following notation:

- μ^*_β denotes the conditional distribution of the random variable $\beta^\top X^*$ given $\{\mathcal{X}_n, \mathcal{Y}_n\}$ (recall it is equal to the empirical distribution of $\beta^\top X_1, \dots, \beta^\top X_n$),
- μ_β denotes the distribution of $\beta^\top X$.

Hence, the distribution of $\mathbb{E}(Y|X) = \beta_0^\top X$ from (1) can be re-written as μ_0 if $\beta_0 \in \mathcal{B}_0$ to emphasize that the distributions μ_{β_0} are all the same in this case. In contrast, note that $\mu^*_{\beta_0}$ might depend on β_0 even if $\beta_0 \in \mathcal{B}_0$.

Let $\widehat{\beta}_n$ denote a deconvolution least squares estimator (DLSE) as defined in Azadkia and Balabdaoui (2024), that is

$$\widehat{\beta}_n \in \arg\min_{\beta \in \mathbb{R}^d} \mathbb{D}_n(\beta).$$

where

$$\mathbb{D}_n(\beta) := \int \Big(F_n^Y(y) - \frac{1}{n}\sum_{i=1}^n F^\epsilon(y - \beta^\top X_i)\Big)^2 dF_n^Y(y),$$

for $\beta \in \mathbb{R}^d$. Here, F_n^Y is the empirical distribution of the sample $\mathcal{Y}_n$ given by

$$F_n^Y(y) := \frac{1}{n}\sum_{i=1}^n \mathbb{1}_{(-\infty, y]}(Y_i)$$

for all $y \in \mathbb{R}$, and F_ϵ is the (known) distribution function of ϵ. Using the same notation as above, $\mu^*_{\widehat{\beta}_n}$ is reserved for the conditional distribution of $\widehat{\beta}_n^\top X^*$ given $\{\mathcal{X}_n, \mathcal{Y}_n\}$, while $\mu_{\widehat{\beta}_n}$ denotes the conditional distribution of $\widehat{\beta}_n^\top X$ given the same samples. Note that randomness in $\mu^*_{\widehat{\beta}_n}$ and $\mu_{\widehat{\beta}_n}$ is carried by the uniformly distributed random variable $J_n \in \{1, \cdots, n\}$ and X respectively.

For arbitrary probability measures μ and ν on $\mathbb{R}$ with finite first moment, we denote by $W_1(\mu, \nu)$ the Wasserstein distance of order 1 between the two measures, that is

$$W_1(\mu, \nu) = \inf_{Z_1 \sim \mu,\ Z_2 \sim \nu} \mathbb{E}|Z_1 - Z_2| = \int_{\mathbb{R}} |F_1(x) - F_2(x)| dx$$

with F_i being the distribution function of Z_i for $i = 1, 2$; see Definition 2.1 and Theorem 2.9 in Bobkov and Ledoux (2019). We will write $\mu * K$ for the convolution of a measure μ and a kernel K on $\mathbb{R}$; i.e., $(\mu * K)(z) = \int_{\mathbb{R}} K(z - x) d\mu(x)$. Finally, for $M, a > 0$ we shall consider the following set of probability measures μ on $\mathbb{R}$:

$$\mathcal{C}(M, a) := \Big\{\mu : \int |x|^{a+2} d\mu(x) \le M\Big\}.$$

1.4 Organization of the paper

The paper is organized as follows: Section 2 presents the primary contributions of this work. Specifically, we introduce an estimator for the distribution μ_0 and establish its parametric rate of convergence in the Wasserstein distance of order 1. Furthermore, we introduce nonparametric estimators for the unconditional density of Z and the conditional density of Z given an observed response. This allows us to introduce and study the problem of estimating the true realization of the linear predictor and show how our estimators and theoretical results can be employed in such a statistical task. Section 3 illustrates several numerical experiments where we show the practical relevance and accuracy of the theoretical findings in various simulation scenarios, providing thereby a clear empirical evidence of their effectiveness. Section 4 provides proofs for the obtained results in which the main arguments leverage concepts and tools from empirical process theory.

2. Main results

2.1 Distributional deconvolution in unlinked linear models

Consider the statistic

$$\widehat{Z}_n = \widehat{\beta}_n^\top X^*$$

where X^* follows the empirical distribution based on the sample $\mathcal{X}_n$ and $\widehat{\beta}_n$ is a deconvolution least squares estimator (DLSE) as defined in Azadkia and Balabdaoui (2024), see also Section 1.3. Our estimator for the distribution μ_0 of $\mathbb{E}(Y|X)$ is $\mu^*_{\widehat{\beta}_n}$, the conditional distribution of $\widehat{Z}_n$ given observations $\{\mathcal{X}_n, \mathcal{Y}_n\}$. In other words, $\mu^*_{\widehat{\beta}_n}$ is the empirical distribution of $\widehat{\beta}_n^\top X_1, \ldots, \widehat{\beta}_n^\top X_n$.

It follows from Proposition 1 in Azadkia and Balabdaoui (2024) that $\widehat{\beta}_n$ exists in the sense that (at least) a minimizer of $\mathbb{D}_n$ exists with probability 1 provided that $n/d > c$ for some fixed constant $c > 0$ and F_ϵ is continuous, an assumption that we will also make below. The assumption $n/d > c$ will be automatically satisfied for sufficiently large n since d is assumed to be fixed. Note that $\widehat{\beta}_n$ might not be unique in which case it will denote any such DLSE, that is any of the elements in $\mathcal{B}_n$ where

$$\mathcal{B}_n := \{\beta \in \mathbb{R}^d \text{ s.t. } \mathbb{D}_n(\beta) = \min_{\beta \in \mathbb{R}^d} \mathbb{D}_n(\beta)\}.$$

We show below that $\mu^*_{\widehat{\beta}_n}$ is consistent in the sense that $W_1(\mu^*_{\widehat{\beta}_n}, \mu_0)$ converges to zero in probability. In fact, it converges at the parametric rate $n^{-1/2}$ provided that some appropriate assumptions hold.

A necessary condition for the consistency of any estimator of μ_0 is identifiability of the latter. In our context, with $Z_0 \sim \mu_0$ independent of ϵ, identifiability of μ_0 means that we have the equivalence

$$\beta^\top X + \epsilon \overset{d}{=} Z_0 + \epsilon \iff \beta^\top X \sim \mu_0 \tag{2}$$

for $\beta \in \mathbb{R}^d$. Although we would like to state our results in great generality and hence only assume that the equivalence in (2) holds true for all $\beta \in \mathbb{R}^d$, we give in the following lemma a sufficient condition for (2) to be satisfied.

Lemma 1 *Let κ_ϵ denote the characteristic function of the noise ϵ; i.e.,*

$$\kappa_\epsilon(t) = \mathbb{E}\exp(it\epsilon), \quad t \in \mathbb{R}.$$

If the set of zeros of κ_ϵ does not contain any open and non-empty interval, then the equivalence in (2) holds (and whence μ_0 is identifiable).

The following theorems extend the results in Azadkia and Balabdaoui (2024) about the convergence of $\widehat{\beta}_n$ to β_0 when β_0 is unique to the case where this parameter is possibly non-unique. Also, the theorems establish convergence of the estimator of μ_0 and give sufficient conditions for this convergence to hold at the parametric rate.

Theorem 2 *Assume that ϵ admits a continuous distribution and that $\mathbb{P}(\alpha^\top X = 0) = 0$ for any $\alpha \in \mathbb{R}^d \setminus \{0\}$. Then, the following holds true.*

1. *For all $\widehat{\beta}_n \in \mathcal{B}_n$ we have that*

$$d(\widehat{\beta}_n, \mathcal{B}_0) = o_{\mathbb{P}}(1), \text{ as } n \to \infty.$$

2. *If in addition $\mu_0 \in \mathcal{C}(M, a)$ for some $M > 0, a > 0$, $\mathbb{E}\|X\| < \infty$ and the equivalence in (2) holds for all $\beta \in \mathbb{R}^d$, then*

$$W_1(\mu^*_{\widehat{\beta}_n}, \mu_0) = o_{\mathbb{P}}(1), \text{ as } n \to \infty.$$

Theorem 3 *Suppose that the distribution function of ϵ is twice continuously differentiable with bounded first and second derivatives. Moreover, suppose that $\mathbb{E}\|X\|^2 < \infty$ and $\mathbb{P}(\alpha^\top X = 0) = 0$ for any $\alpha \in \mathbb{R}^d \setminus \{0\}$. For all $\beta \in \mathbb{R}^d$, define*

$$U(\beta) = \int \Big(\int x f^\epsilon(y - \beta^\top x) dF^X(x) \Big) \Big(\int x^\top f^\epsilon(y - \beta^\top x) dF^X(x) \Big) dF^Y(y),$$

denote by $\lambda(\beta)$ the smallest eigenvalue of $U(\beta)$, and assume $\inf_{\beta_0 \in \mathcal{B}_0} \lambda(\beta_0) > 0$. Then, the following holds true.

1. *For all $\widehat{\beta}_n \in \mathcal{B}_n$, we have that*

$$d(\widehat{\beta}_n, \mathcal{B}_0) = O_{\mathbb{P}}(n^{-1/2}), \text{ as } n \to \infty.$$

2. *If in addition $\mathbb{E}\|X\|^p < \infty$ for some $p > 2$, there exists $A > 0$ such that for all intervals $[a, b]$ in $\mathbb{R}$, $\mu_0([a, b]) \leq A|b - a|$, and the equivalence in (2) holds for all $\beta \in \mathbb{R}^d$, then*

$$W_1(\mu^*_{\widehat{\beta}_n}, \mu_0) = O_{\mathbb{P}}(n^{-1/2}), \text{ as } n \to \infty.$$

Let $\lambda_0 := \inf_{\beta_0 \in \mathcal{B}_0} \lambda(\beta_0)$. The assumption $\lambda_0 > 0$ is a well-separability condition. In M-estimation, such a condition is rather classical and made to ensure convergence of the M-estimator to the true parameter in case of identifiability. Since the set $\mathcal{B}_0$ might contain more than one element, the condition $\lambda_0 > 0$ is a natural well-separability requirement in

this case as it ensures that $\mathcal{D}(\beta) > 0$ whenever $\beta \notin \mathcal{B}_0$, where $\mathcal{D}$ is the population criterion; i.e.,

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

The major role played by λ_0 can be seen in the proof of Theorem 3: In the first and second claims of the theorem, the $O_{\mathbb{P}}(n^{-1/2})$ takes in fact the form $\lambda_0^{-1/2} O_{\mathbb{P}}(n^{-1/2})$, respectively $(1+\lambda_0^{-1/2}) O_{\mathbb{P}}(n^{-1/2})$, where now, the $O_{\mathbb{P}}$-term does not depend on λ_0. Hence, both bounds on $d(\widehat{\beta}_n, \mathcal{B}_0)$ and $W_1(\mu^*_{\widehat{\beta}_n}, \mu_0)$ are of order $n^{-1/2}$ but (for fixed n) become larger for smaller values of λ_0.

Note that the assumption that $\mu_0 \in \mathcal{C}(M, a)$ stated in Theorem 2 is not needed in Theorem 3 because in that theorem we require the stronger condition $\mathbb{E}\|X\|^p < \infty$ for some $p > 2$. Indeed, note that $\mu_0 \in \mathcal{C}(M, a)$ is equivalent to $\mathbb{E}|\beta_0^\top X|^{a+2} \leq M$ for all $\beta_0 \in \mathcal{B}_0$. Using the Cauchy-Schwarz inequality implies that

$$\mathbb{E}|\beta_0^\top X|^{a+2} \leq \|\beta_0\|^{a+2} \mathbb{E}\|X\|^{a+2}.$$

Thanks to Lemma 7 in Section 4 below, there exists $B > 0$ such that $\|\beta_0\| \leq B$ for all $\beta_0 \in \mathcal{B}_0$, thus, taking $a := p - 2 > 0$ and $M := B^{a+2}\mathbb{E}\|X\|^p < \infty$, we get $\mathbb{E}|\beta_0^\top X|^{a+2} \leq M$. In other words, the assumption $\mathbb{E}\|X\|^p < \infty$ for some $p > 2$ implies that we can find $a > 0$ and $M > 0$ such that $\mu_0 \in \mathcal{C}(M, a)$.

2.2 Density deconvolution and inference on the latent predictor

A relevant inferential task involves the scenario where we observe a realization y_0 from the response variable Y such that $Y = \beta_0^\top X + \epsilon$. Suppose that the realization of the covariate X, which generated y_0, is non-observed. At the same time, suppose that we are provided with a sample of unlinked covariates and responses $\{\mathcal{X}_n, \mathcal{Y}_n\}$. Let $Z := \mathbb{E}(Y|X) = \beta_0^\top X$ and z_0 be the true value of Z which is linked to y_0 but not accessible. Then, it is of interest to make inference about z_0 conditionally on the event $\{Y = y_0\}$. One possible approach is based on estimating the conditional density of $Z|Y = y_0$, together with several summaries of the corresponding conditional distribution, including location measures such as the mean and mode, as well as quantiles.

In our previous results, we introduced an estimator for the distribution of Z based on the empirical distribution of $\widehat{\beta}_n^\top X_1, \ldots, \widehat{\beta}_n^\top X_n$. Assuming that the distribution of Z admits a density with respect to Lebesgue measure, f_Z say, we can apply a Kernel Density Estimator (KDE) with a given kernel and bandwidth and which is based on $\widehat{\beta}_n^\top X_i, i = 1, \ldots, n$. We denote the resulting density estimator by $\widehat{f}_Z$. Let f_Y, $f_{Y|Z}$ and $f_{Z|Y}$ denote the density of Y, and conditional densities of Y given Z and Z given Y, respectively. Using the Bayes rule, we have that

$$f_{Z|Y}(z|y) = \frac{f_{Y|Z}(y|z) f_Z(z)}{f_Y(y)}, \quad y, z \in \mathbb{R}. \tag{3}$$

Since $f_{Y|Z}(y|z) = f_\epsilon(y - z)$, where f_ϵ is the known density of the noise ϵ, the only missing element is an estimator for the (marginal) density f_Y, for which a simple estimator is given

by

$$\widehat{f}_Y(y) = \frac{1}{n}\sum_{i=1}^{n} f_\epsilon(y - \widehat{\beta}_n^\top X_i), \quad y \in \mathbb{R}. \tag{4}$$

Thus, we consider the following estimator of the conditional density $f_{Z|Y}$:

$$\widehat{f}_{Z|Y}(z|y) = \frac{f_\epsilon(y-z)\widehat{f}_Z(z)}{\widehat{f}_Y(y)}, \quad y, z \in \mathbb{R}. \tag{5}$$

In the following theorem we provide guarantees for both the unconditional and conditional density estimators.

Theorem 4 *Suppose μ_0 admits a density $f_Z \in C^2(\mathbb{R})$ with bounded first and second derivatives. Moreover, for some $L > 0$ consider an L-Lipschitz kernel $K : \mathbb{R} \to [0, \infty)$ such that $\int K(u)du = 1$, $\int uK(u)du = 0$ and $\int u^2K(u)du < \infty$. Let $K_h(z) := h^{-1}K(z/h)$ with $h = h_n \asymp n^{-1/8}$. Under the assumptions of Theorem 3, the following holds true.*

1. *The estimator $\widehat{f}_Z(z) := (\mu^*_{\widehat{\beta}_n} * K_h)(z) = n^{-1}\sum_{i=1}^n K_h(z - \widehat{\beta}_n^\top X_i)$ satisfies*

$$\|\widehat{f}_Z - f_Z\|_\infty = O_{\mathbb{P}}(n^{-1/4}).$$

2. *If, moreover, f_ϵ is bounded and M-Lipschitz for some $M > 0$, and $f_Y(y_0) \in (0, \infty)$, then the estimator $\widehat{f}_{Z|Y}(z|y_0) = f_\epsilon(y_0 - z)\widehat{f}_Z(z)/\widehat{f}_Y(y_0)$, with $\widehat{f}_Y$ given in (4), satisfies*

$$\|\widehat{f}_{Z|Y}(\cdot|y_0) - f_{Z|Y}(\cdot|y_0)\|_\infty = O_{\mathbb{P}}(n^{-1/4}).$$

It is worth remarking that in the setting of Theorem 4, under the additional assumptions $f_Z \in C^s(\mathbb{R})$, $\int u^j K(u)du = 0$ for $j = 1 \dots s-1$, $\int u^s K(u)du < \infty$, and $h_n \asymp n^{-1/(2s+4)}$, it is possible to obtain the rate of convergence $n^{-s/(2s+4)}$ in sup-norm for both density estimators $\widehat{f}_Z$ and $\widehat{f}_{Z|Y}(\cdot|y_0)$. Moreover, it is evident that the resulting rate of convergence is slower than the one achieved in classical KDE theory. The question of optimality lies beyond the scope of this work and is left for future research.

Now suppose that $Y = y_0$, a realization from the distribution of $\beta_0^T X + \epsilon = z_0 + \epsilon$. The true mean z_0 is unknown and the goal is to use the estimator of the conditional density given in (5) to make inference about z_0 given that $Y = y_0$ was observed. We consider

$$\widehat{E}(Z|Y = y_0) = \int z\, \widehat{f}_{Z|Y}(z|y_0)\, dz, \qquad \widehat{M}(Z|Y = y_0) = \arg\max_z\, f_\epsilon(y_0 - z)\widehat{f}_Z(z), \tag{6}$$

which can be computed via numerical methods. In a Bayesian modelling framework, interval estimators can also be considered. These are derived as credible intervals based on the quantiles of the distribution of $Z|Y = y_0$, which are in turn obtained via $\widehat{f}_{Z|Y=y_0}$. Let $Q_{Z|Y}(\cdot|y_0)$ be the quantile function of $Z|Y = y_0$. For a given probability $p \in (0, 1)$, consider the estimator

$$\widehat{Q}_{Z|Y}(p|y_0) := \inf\left\{x \in \mathbb{R} : \int_{-\infty}^{x} \widehat{f}_{Z|Y}(z|y_0)dz \geq p\right\} = \widehat{F}^{-1}_{Z|Y}(p|y_0). \tag{7}$$

where $\widehat{F}_{Z|Y}(\cdot|y_0) = \int_{-\infty}^{\cdot} \widehat{f}_{Z|Y}(z|y_0)dz$. The following Proposition provides rates of convergence for (6) and (7).

Proposition 5 *Suppose that the assumptions of Theorem 4 are satisfied and let* $y_0 \in \mathbb{R}$ *such that* $f_Y(y_0) \in (0, \infty)$.

1. *For the mean estimator it holds*

$$|\widehat{E}(Z|Y = y_0) - \mathbb{E}(Z|Y = y_0)| = O_{\mathbb{P}}(n^{-1/4}).$$

2. *Suppose that* $z \mapsto f_{Z|Y}(z|y_0)$ *admits a maximizer at* $z^* := \mathbb{M}(Z|Y = y_0)$ *such that if* $|z - z^*| > M$ *then for some* $\delta > 0$ *(which depends on* M*) it holds that*

$$f_{Z|Y}(z^*|y_0) - f_{Z|Y}(z|y_0) \geq \delta.$$

Then, $\widehat{M}(Z|Y = y_0) \xrightarrow{\mathbb{P}} z^*$. *Furthermore, if*

$$f_{Z|Y}(z^*|y_0) - f_{Z|Y}(z|y_0) \geq \lambda |z^* - z|^{\alpha},$$

on a small neighbourhood of z^* *for some constants* $\alpha, \lambda > 0$, *then*

$$|\widehat{M}(Z|Y = y_0) - \mathbb{M}(Z|Y = y_0)| = O_{\mathbb{P}}(n^{-1/(4\alpha)}).$$

3. *Let* $p \in (0, 1)$ *and suppose* $f_{Z|Y}(\cdot|y_0)$ *is strictly positive in a neighbourhood of* $Q_{Z|Y}(p|y_0)$. *Moreover, assume that there exists constants* $c_1, c_2, c_3 > 0$ *such that* $f_{Z|Y}(z|y_0) \leq c_2 e^{-c_3|z|}$ *for all* $|z| > c_1$. *Then,*

$$|\widehat{Q}_{Z|Y}(p|y_0) - Q_{Z|Y}(p|y_0)| = O_{\mathbb{P}}(n^{-1/4} \log n).$$

In the next section, we show numerically that when the noise ϵ has a small variance σ^2, the density estimator $\widehat{f}_{Z|Y=y_0}$ and the estimators in (6) perform well in recovering the true value z_0 linked to y_0.

3. Numerical experiments

In this section, we illustrate the theoretical results using simulated data. Our first goal is to show that $\mu^*_{\widehat{\beta}_n}$ is close to μ_0 in the sense of W_1-distance and that the estimation error decreases at a parametric rate. In addition, we aim to evaluate the performance of the introduced estimators both qualitatively and quantitatively. For all experiments, the estimator $\widehat{\beta}_n$ is calculated using the `optim` function from the R-package `stats`. The simulation study was conducted using 500 independent Monte Carlo replications, and the results were reported in the form of Monte Carlo averages of the relevant quantities. Let $\mathcal{N}(\mu, \Sigma)$ denote the multivariate Gaussian distribution with mean $\mu \in \mathbb{R}^d$ and positive definite covariance matrix $\Sigma \in \mathbb{R}^{d \times d}$. In addition, we denote by Gamma(a, b) the Gamma distribution with shape $a > 0$ and scale $b > 0$ respectively. In all simulations we took $\epsilon \sim \mathcal{N}(0, \sigma^2)$, for $\sigma > 0$, and considered four different settings:

(a) $d = 2$, $X \sim \mathcal{N}(0, I)$, $\beta_0^\top = (3, -5)$,

(b) $d = 3$, $X \sim \mathcal{N}(0, I)$, $\beta_0^\top = (-1.5, 2, 7)$,

(c) $d = 2$, $X = (X^{(1)}, X^{(2)})^\top$ with $X^{(1)} \sim \text{Gamma}(1,1)$, $X^{(2)} \sim \text{Gamma}(2,4)$, $\beta_0^\top = (1,2)$,

(d) $d = 3$, $X = (X^{(1)}, X^{(2)}, X^{(3)})^\top$ with $X^{(1)} \sim \text{Gamma}(1,1)$, $X^{(2)} \sim \text{Gamma}(2,4)$, $X^{(3)} \sim \text{Gamma}(1.5,3)$, $\beta_0^\top = (0.5,2,3)$.

Note that in the simulation settings (a) and (b) it holds that $\mathcal{B}_0 = \{\beta \in \mathbb{R}^2 : \|\beta\| = \|\beta_0\| = \sqrt{34}\}$ for (a) and $\mathcal{B}_0 = \{\beta \in \mathbb{R}^3 : \|\beta\| = \|\beta_0\| = \sqrt{55.25}\}$ for (b). Hence, β_0 is not identifiable in these settings. For the settings in (c) and (d), we can use Theorem 5 of Balabdaoui et al. (2025). This theorem implies that if $X = (X_1, \ldots, X_p)^\top$ such that $X_1, \ldots, X_p$ are independent and $X_i \sim \text{Gamma}(a_i, b_i), i = 1, \ldots, p$ with the property that

$$\sum_{i \in I} a_i = \sum_{j \in J} a_j \implies I = J \tag{8}$$

for $I, J \subset \{1, \ldots, p\}$, then $|\mathcal{B}_0| = 1$. In setting (c), we have that $a_1 = 1 \neq a_2 = 2$, and the condition in (8) is clearly satisfied. In setting (d), we have that $a_1 = 1, a_2 = 2, a_3 = 1.5$, and we can easily check that (8) holds. Thus, in settings (c) and (d), β_0 is identifiable.

Experiment 1. We fix $\sigma = 1$ and calculate $W_1(\mu^*_{\widehat{\beta}_n}, \mu_0)$ on the basis of i.i.d. samples with increasing size $n \in \{1000, 2000, \ldots, 5000\}$, drawn from the distribution of each of the four settings (a)-(d). For this purpose, a general method to practically compute $W_1(\mu^*_{\widehat{\beta}_n}, \mu_0)$ is the following: First, draw a large sample from μ_0 of size $m = 10^6$ and compute the associated empirical distribution function $F_{0,m}$. Successively, compute the empirical distribution function $F^*_{\widehat{\beta}_n}(x) = \frac{1}{n}\sum_{i=1}^n \mathbb{1}_{\widehat{\beta}_n X_i \leq x}$ based on the sample $\widehat{\beta}_n^\top X_1, \ldots, \widehat{\beta}_n^\top X_n$. Now, introduce the sequence $\{x_j\}_{j=1}^{n+m}$ in $\mathbb{R}$ obtained merging and sorting the two empirical samples and the sequence $\{u_j\}_{j=1}^{n+m+1} = \{0\} \cup \{v(x_j)\}_{j=1}^{n+m}$ in $[0,1]$ such that

$$v(x_j) = \begin{cases} F^*_{\widehat{\beta}_n}(x_j), & \text{if } x_j \text{ belongs to the sample from } \mu^*_{\widehat{\beta}_n}, \\ F_{0,m}(x_j), & \text{if } x_j \text{ belongs to the sample from } \mu_0, \end{cases}$$

namely, $\{u_j\}_{j=1}^{n+m+1}$ contains all the sorted values attained by the two empirical distribution functions. Finally, consider the following numerical approximation

$$W_1(\mu^*_{\widehat{\beta}_n}, \mu_0) \approx \sum_{j=1}^{n+m+1} |F^{*-1}_{\widehat{\beta}_n}(u_{j+1}) - F^{-1}_{0,m}(u_{j+1})|(u_{j+1} - u_j),$$

which is arbitrarily accurate by increasing m.

Now, denote by $W_1^{(j)}(\mu^*_{\widehat{\beta}_n}, \mu_0)$ the j-th Monte Carlo replication of $W_1(\mu^*_{\widehat{\beta}_n}, \mu_0)$ for $j = 1, \ldots, J = 500$. Based on these replications, we compute the slope of the log-linear fit (linear fit on the logarithmic scale) for

$$\overline{W}_1^{(k)} := \frac{1}{J}\sum_{j=1}^J \left(W_1^{(j)}(\mu^*_{\widehat{\beta}_n}, \mu_0)\right)^k, \quad k = 1, 2, 3,$$

Table 1: Estimated slopes of the log-linear fit for the first 3 moments and 99% quantile of the Monte Carlo distribution of $W_1(\mu^*_{\widehat{\beta}_n}, \mu_0)$ under the settings (a)-(d).

Setting	$\overline{W}_1^{(1)}$	$\overline{W}_1^{(2)}$	$\overline{W}_1^{(3)}$	$q_{W_1}^{0.99}$
(a)	-0.502	-1.007	-1.516	-0.567
(b)	-0.493	-0.987	-1.479	-0.514
(c)	-0.501	-1.000	-1.496	-0.464
(d)	-0.485	-0.953	-1.365	-0.477

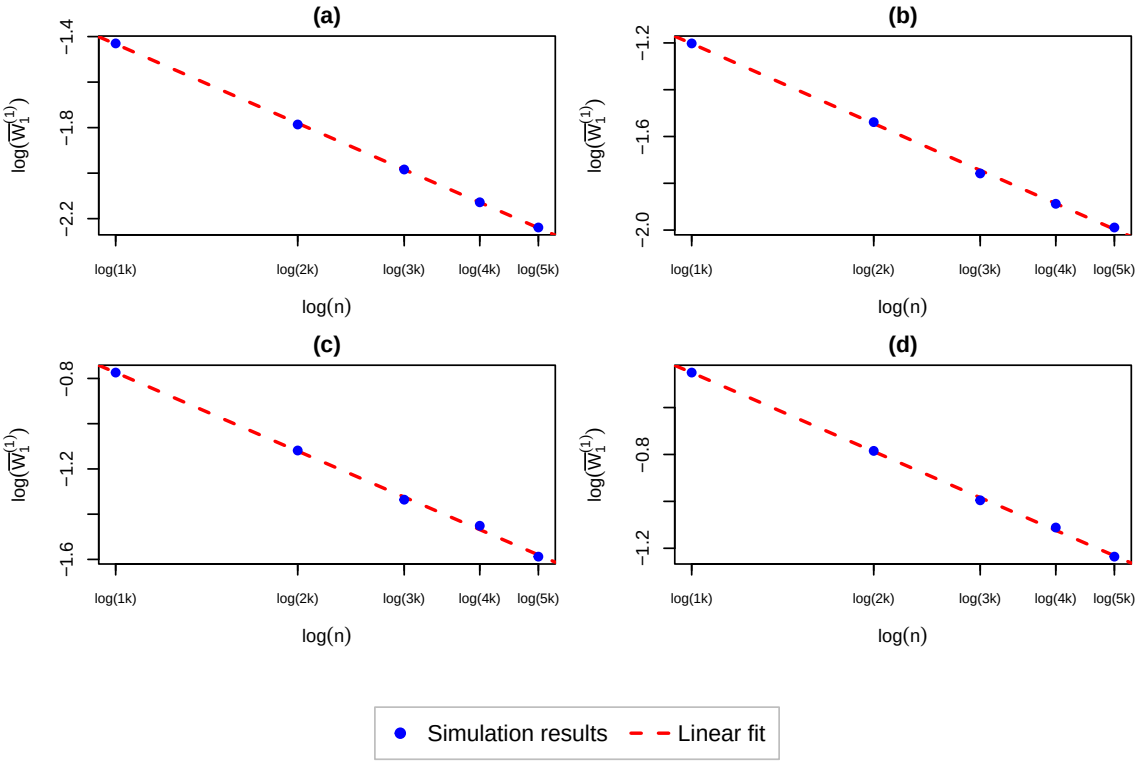

Figure 1: Log-linear fit for the first moment $\overline{W}_1^{(1)}$ of the Monte Carlo distribution of $W_1(\mu^*_{\widehat{\beta}_n}, \mu_0)$ under the settings (a)-(d) for varying sample size. The slopes of the linear fit are -0.502 in (a), -0.493 in (b), -0.501 in (c) and -0.485 in (d).

and also the 99% empirical quantile, which we denote by $q_{W_1}^{0.99}$. The results are reported in Table 1 and suggest that $\overline{W}_1^{(k)} \asymp n^{-k/2}$ and $q_{W_1}^{0.99} \asymp n^{-1/2}$, thus providing strong empirical evidence for the theory. Figure 1 also reports graphically the Monte Carlo average $\overline{W}_1^{(1)}$ of the Wasserstein distances on the logarithmic scale.

Experiment 2. In all the four settings, we again fix $\sigma = 1$. Here, a random sample of size $n = 500$ is drawn from each of the corresponding distributions and then used to calculate the KDE $\widehat{f}_Z$ introduced in Theorem 4 using a Gaussian Kernel with a bandwidth systematically chosen via the simple rule of thumb $h = n^{-1/8}\,\mathrm{sd}(\widehat{\beta}_n^\top X_1, \ldots, \widehat{\beta}_n^\top X_n)$ where sd denotes the sample standard deviation, in order to ensure that it is appropriately rescaled and matches the setting of Theorem 4. For a given value of z_0 randomly generated from Z, a new random response $Y \sim z_0 + \epsilon$ was generated. Based on the realization y_0 of Y, we compute the estimator $\widehat{f}_{Z|Y=y_0} := \widehat{f}_{Z|Y}(\cdot|y_0)$ for the density of the random variable $Z|Y = y_0$ given in (5), the resulting mean $\widehat{E}(Z|Y = y_0)$ and mode $\widehat{M}(Z|Y = y_0)$ given in (6), and the resulting 95% credible interval that we denote by $\mathrm{CI}_{0.95}(\widehat{f}_{Z|Y=y_0})$. We remark that when Z has a small variance, adaptive quadrature methods for computing $\widehat{E}(Z|Y = y)$ may perform poorly. For this reason, we resort to importance sampling to obtain a more robust and stable approximation of the integral in (6) and of the credible intervals based on $\widehat{f}_{Z|Y=y}$. Moreover, most density evaluations are performed on the logarithmic scale to enhance numerical stability by reducing errors in floating-point arithmetic. The results of the experiment are shown in Figure 2. In all settings, it is clear that $\widehat{f}_Z$ recovers well the true f_Z. However, as expected, the conditional density estimator $\widehat{f}_{Z|Y=y_0}$ yields a much better prediction for the true z_0.

Experiment 3. For the four settings, we fix $\sigma = 1$ and let $n \in \{50, 100, 500\}$. Here, we shall compare the performance at recovering the true z of the point estimators $\widehat{E}(Z|Y = y)$ and $\widehat{M}(Z|Y = y)$ with the estimated mean and mode of Z, respectively given by

$$\widehat{E}(Z) = \frac{1}{n}\sum_{i=1}^{n} \widehat{\beta}_n^\top X_i, \quad \widehat{M}(Z) = \arg\max_z \widehat{f}_Z(z).$$

The performance comparison is done by computing the performance ratios

$$R_{\widehat{E}} = \frac{\mathrm{MSE}(\widehat{E}(Z|Y = y))}{\mathrm{MSE}(\widehat{E}(Z))}, \quad R_{\widehat{M}} = \frac{\mathrm{MSE}(\widehat{M}(Z|Y = y))}{\mathrm{MSE}(\widehat{M}(Z))}.$$

To explain how the MSE is computed, let us focus on the conditional mode. For each Monte Carlo replication, we compute this mode using a sample of size n, thereby obtaining $y \mapsto \widehat{M}^{(j)}(Z|Y = y)$ for $j = 1, \ldots J = 500$. The corresponding MSE is then obtained as the average squared difference between the mode and the true z,

$$\mathrm{MSE}(\widehat{M}(Z|Y = y)) = \frac{1}{J}\sum_{j=1}^{J}\frac{1}{T}\sum_{t=1}^{T}\left(\widehat{M}^{(j)}(Z|Y = y_t) - z_t\right)^2 \tag{9}$$

where T is the size of test sample $\{(z_t, y_t)\}_{t=1}^T$, which is set to be equal to 100. A similar approach is used for the conditional mean and the unconditional estimators. In this experiment, we also compare the performance of two credible intervals, one obtained through the symmetric extreme empirical quantiles based on $\widehat{\beta}_n^\top X_1, \ldots, \widehat{\beta}_n^\top X_n$, and the other obtained through importance sampling based on $\widehat{f}_{Z|Y=y}$. For a given nominal $\alpha \in (0, 1)$, with a slight abuse of notation, we denote such intervals by $\mathrm{CI}_{1-\alpha}(\widehat{f}_Z)$ and $\mathrm{CI}_{1-\alpha}(\widehat{f}_{Z|Y=y})$, respectively.

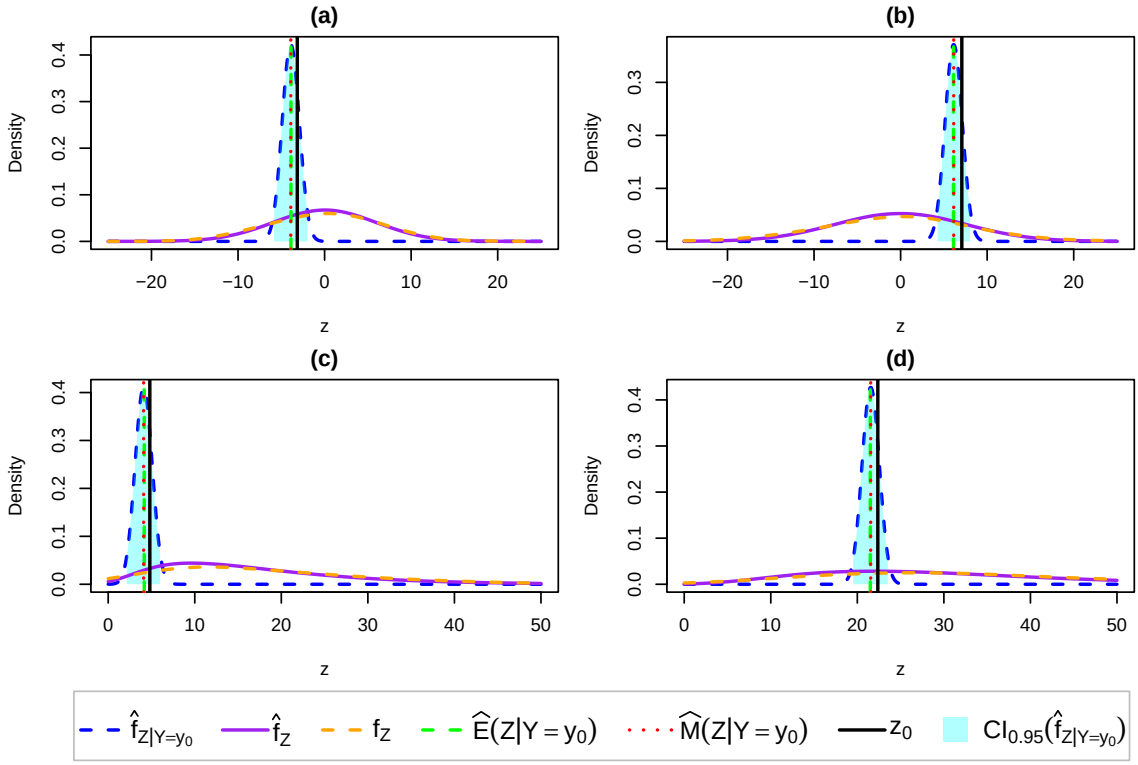

Figure 2: Graphical results of Experiment 2 under the settings (a)-(d). The true (z_0, y_0) are $(-3.155, -3.996)$ in (a), $(7.074, 6.233)$ in (b), $(4.809, 3.968)$ in (c) and $(22.367, 21.526)$ in (d).

For the experiment, we set $\alpha = 0.05$ and report the Monte Carlo empirical frequentist coverage of the credible intervals computed as

$$\mathcal{C}(\mathrm{CI}_{1-\alpha}(\widehat{f}_{Z|Y=y})) = \frac{1}{J}\sum_{j=1}^{J}\frac{1}{T}\sum_{t=1}^{T}\mathbb{1}\{z_t \in \mathrm{CI}_{1-\alpha}(\widehat{f}^{(j)}_{Z|Y=y_t})\},$$

and similarly for $\mathrm{CI}_{1-\alpha}(\widehat{f}_Z)$. Moreover, we report the length ratio

$$R_{\mathrm{len}(\mathrm{CI}_{1-\alpha})} = \frac{\frac{1}{J}\sum_{j=1}^{J}\frac{1}{T}\sum_{t=1}^{T}\mathrm{len}(\mathrm{CI}_{1-\alpha}(\widehat{f}^{(j)}_{Z|Y=y_t}))}{\frac{1}{J}\sum_{j=1}^{J}\frac{1}{T}\sum_{t=1}^{T}\mathrm{len}(\mathrm{CI}_{1-\alpha}(\widehat{f}^{(j)}_{Z}))},$$

where $\mathrm{len}([a,b]) = b-a$. The results are reported in Table 2. It is at once apparent that the proposed conditional estimators are preferable in terms of MSE. Moreover, for all scenarios, both interval estimators attain the correct frequentist coverage, with the conditional one being calibrated already for very small sample sizes and leading to much narrower intervals.

Table 2: Performance comparison under the settings (a)-(d) for varying sample size.

Setting	n	$R_{\widehat{E}}$	$R_{\widehat{M}}$	$\mathcal{C}(\mathrm{CI}_{0.95}(\widehat{f}_{Z\|Y=y}))$	$\mathcal{C}(\mathrm{CI}_{0.95}(\widehat{f}_Z))$	$R_{\mathrm{len}(\mathrm{CI}_{1-\alpha})}$
(a)	50	0.028	0.027	0.949	0.915	0.182
	100	0.029	0.028	0.949	0.933	0.175
	500	0.029	0.029	0.949	0.946	0.170
(b)	50	0.018	0.017	0.950	0.912	0.144
	100	0.018	0.017	0.950	0.934	0.138
	500	0.018	0.018	0.950	0.946	0.134
(c)	50	0.009	0.007	0.948	0.913	0.101
	100	0.008	0.007	0.948	0.931	0.096
	500	0.008	0.006	0.948	0.946	0.092
(d)	50	0.010	0.010	0.949	0.910	0.105
	100	0.010	0.010	0.949	0.929	0.102
	500	0.010	0.010	0.948	0.946	0.098

Experiment 4. Under the four settings, for a given linked realization (z_0, y_0) we use $\widehat{E}(Z|Y = y_0)$ and $\widehat{M}(Z|Y = y_0)$ in (5) and (6) for predicting z_0, and we evaluate again their predictive performance by estimating their associated Mean Squared Error (MSE) over Monte Carlo replications of (z, y). As opposed to the previous experiment we perform such evaluation for different values of n and also σ (recall that $\sigma = 1$ in Experiment 3). The MSE of the estimators is computed as in (9). Here, we take $n \in \{50, 100, 500\}$ and $\sigma^2 \in \{0.5, 1, 1.5, 2, 2.5\}$. The results are reported in Tables 3 and 4 where it is shown that the proposed estimators of the true value of the linear predictor are comparable in terms of their performance. In both tables, the MSE decreases as n increases under all the considered settings. However, it does not tend to 0 since the magnitude of the noise constraints the accuracy that any estimator can achieve. The MSE is comparable to σ^2, and it is interesting to note that noise introduces a form of aleatoric uncertainty that cannot be reduced as opposed to epistemic uncertainty that decreases with increasing sample size; see Hüllermeier and Waegeman (2021).

4. Proofs

4.1 Preparatory lemmas

The following lemma is an extension of Theorem 3.2.5 in van der Vaart and Wellner (2023).

Lemma 6 *Let Θ be equipped with a semi-metric d with $\mathcal{B} \subseteq \Theta$ and define $d(\theta, \mathcal{B}) = \inf_{\theta' \in \mathcal{B}} d(\theta, \theta')$. Let $\mathbb{M}_n$ be a stochastic process indexed by Θ and $\mathbb{M} : \Theta \mapsto \mathbb{R}$ a deterministic function such that for every $\theta_0 \in \mathcal{B}$ and $\theta \in \Theta$ satisfies*

$$\mathbb{M}(\theta) - \mathbb{M}(\theta_0) \leq -C_0^2 d^2(\theta, \mathcal{B})$$

Table 3: MSE for the mean estimator under the settings (a)-(d) for varying σ^2 and sample size.

Setting	n	$\sigma^2 = 0.5$	$\sigma^2 = 1$	$\sigma^2 = 1.5$	$\sigma^2 = 2$	$\sigma^2 = 2.5$
(a)	50	0.497	0.981	1.454	1.915	2.365
	100	0.496	0.980	1.452	1.911	2.359
	500	0.497	0.979	1.449	1.906	2.352
(b)	50	0.499	0.989	1.472	1.948	2.416
	100	0.498	0.988	1.469	1.942	2.408
	500	0.498	0.987	1.468	1.940	2.405
(c)	50	0.645	1.138	1.627	2.111	2.590
	100	0.551	1.044	1.532	2.015	2.493
	500	0.517	1.010	1.496	1.976	2.449
(d)	50	0.546	1.041	1.532	2.017	2.499
	100	0.512	1.006	1.496	1.980	2.460
	500	0.509	1.002	1.491	1.974	2.453

Table 4: MSE for the mode estimator under the settings (a)-(d) for varying σ^2 and sample size.

Setting	n	$\sigma^2 = 0.5$	$\sigma^2 = 1$	$\sigma^2 = 1.5$	$\sigma^2 = 2$	$\sigma^2 = 2.5$
(a)	50	0.496	0.980	1.453	1.914	2.364
	100	0.496	0.979	1.451	1.910	2.358
	500	0.496	0.978	1.447	1.905	2.350
(b)	50	0.498	0.988	1.471	1.947	2.415
	100	0.497	0.987	1.467	1.941	2.407
	500	0.497	0.986	1.466	1.938	2.402
(c)	50	0.643	1.134	1.622	2.105	2.584
	100	0.550	1.042	1.529	2.011	2.488
	500	0.515	1.007	1.492	1.972	2.446
(d)	50	0.545	1.039	1.530	2.015	2.497
	100	0.510	1.004	1.493	1.978	2.458
	500	0.507	1.001	1.489	1.971	2.449

for some $C_0 > 0$. Suppose that for all n and sufficiently small $\delta > 0$, the centred process $\mathbb{M}_n - \mathbb{M}$ satisfies

$$\mathbb{E} \sup_{(\theta,\theta_0)\in\Theta\times\mathcal{B}:\ d(\theta,\theta_0)<\delta} \left|(\mathbb{M}_n - \mathbb{M})(\theta) - (\mathbb{M}_n - \mathbb{M})(\theta_0)\right| \lesssim \frac{\phi_n(\delta)}{\sqrt{n}}, \tag{10}$$

for functions ϕ_n such that $\delta \mapsto \phi_n(\delta)/\delta^\alpha$ is decreasing for some $\alpha < 2$ (not depending on n). Moreover, suppose that for all n,

$$r_n^2 \phi_n\Big(\frac{1}{r_n}\Big) \leq \sqrt{n},$$

the sequence $\widehat{\theta}_n$ satisfies

$$\mathbb{M}_n(\widehat{\theta}_n) \geq \sup_{\theta_0 \in \mathcal{B}} \mathbb{M}_n(\theta_0) - O_{\mathbb{P}}(r_n^{-2}),$$

and $d(\widehat{\theta}_n, \mathcal{B}) \to 0$ in outer probability, then

$$r_n d(\widehat{\theta}_n, \mathcal{B}) = C_0^{-1} O_{\mathbb{P}}(1)$$

where the $O_{\mathbb{P}}$-term does not depend on C_0.

In the next lemma, we state the compactness of the set $\mathcal{B}_0$.

Lemma 7 *Suppose that $E\|X\|^2 < \infty$. Then, $\mathcal{B}_0$ is compact and for every $\beta \in \mathbb{R}^d$ there exists $\beta_0 \in \mathcal{B}_0$ such that*

$$\inf_{\beta' \in \mathcal{B}_0} \|\beta - \beta'\| = \|\beta - \beta_0\|.$$

The next lemma is a slight extension of Proposition 11 in Azadkia and Balabdaoui (2024) to the case where $|\mathcal{B}_0| > 1$, and we use the empirical process notation (van der Vaart and Wellner, 2023). With V denoting either X or Y, we introduce

$$\mathbb{G}_n^V := \sqrt{n}(\mathbb{P}_n^V - \mathbb{P}^V)$$

where $\mathbb{P}^V$ denotes the distribution of the random variable V and $\mathbb{P}_n$ denotes the empirical distribution of n independent random variables distributed as V, and for a class of functions $\mathcal{F}$ we let

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

Moreover, if $\mathcal{F}$ has envelop F we define

$$J(\delta, \mathcal{F}) := \sup_{\mathbb{Q}} \int_0^\delta \sqrt{1 + \log N(\eta \|F\|_{\mathbb{Q}}, \mathcal{F}, L^2(\mathbb{Q}))} d\eta$$

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

Lemma 8 *Suppose that $\mathbb{E}\|X\|^2 < \infty$, and that ϵ has a bounded density. Consider the classes of functions $\mathcal{M}$ and $\mathcal{G}_\delta$, where $\delta > 0$, defined as*

$$\mathcal{M} := \{m_{\beta_0, y},\ y \in \mathbb{R}, \beta_0 \in \mathcal{B}_0\},$$

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

$$\mathcal{G}_\delta := \{g_{\beta,\beta',y}(x) = m_{\beta,y}(x) - m_{\beta',y}(x),\ \|\beta - \beta'\| \leq \delta, y \in \mathbb{R}\}$$

Then, we have

$$\mathbb{E}\|\mathbb{G}_n^X\|_{\mathcal{M}}^2 \lesssim 1 \quad \text{and} \quad \mathbb{E}\|\mathbb{G}_n^X\|_{\mathcal{G}_\delta}^2 \lesssim \delta^2.$$

The next lemma is a slight extension of Proposition 14 in Azadkia and Balabdaoui (2024) to the case where $\mathcal{B}_0$ is not necessarily reduced to a single point. The proof is also similar except that it uses Lemma 8 above instead of Proposition 11 in Azadkia and Balabdaoui (2024). Hence, the proof will be omitted. Let us define the population criterion as

$$\mathcal{D}(\beta) := \int \Big(F^Y(y) - \int F^{\epsilon}(y - \beta^\top x) dF^X(x)\Big)^2 dF^Y(y) \tag{11}$$

where F^V is the distribution function of V, with V denoting either X, or Y, or ϵ.

Lemma 9 *Suppose that $\mathbb{E}\|X\|^2 < \infty$, and that ϵ has a bounded density. Then, there exists a constant $C > 0$, such that*

$$\sqrt{n}\mathbb{E}\sup_{\beta,\beta_0}\Big|\mathbb{D}_n(\beta) - \mathcal{D}(\beta) - (\mathbb{D}_n(\beta_0) - \mathcal{D}(\beta_0))\Big| \leq C\Big(\delta + \frac{\delta}{\sqrt{n}} + \frac{1}{\sqrt{n}}\Big),$$

where the supremum is extended to all $\beta \in \mathbb{R}^d$ and $\beta_0 \in \mathcal{B}_0$ such that $\|\beta - \beta_0\| \leq \delta$.

The next lemma will be useful to bound the Wasserstein distance between μ_0 and its estimator, it uses the theory of entropy with bracketing. For a class of functions $\mathcal{F}$ on $\mathbb{R}^d$ and $\varepsilon > 0$ we denote by $N_{[]}(\varepsilon, \mathcal{F})$ the bracketing number of $\mathcal{F}$ with radius ε, with respect to the $L^2(\mathbb{P}^X)$-norm, see Definition 2.1.6 in van der Vaart and Wellner (2023).

Lemma 10 *For every $\beta_0 \in \mathcal{B}_0$ and $t \in \mathbb{R}$, consider the function defined for $x \in \mathbb{R}^d$ by*

$$f_{\beta_0,t}(x) = \begin{cases} \mathbb{1}_{\beta_0^\top x > t} & \text{if } t > 0 \\ \mathbb{1}_{\beta_0^\top x \leq t} & \text{if } t < 0. \end{cases}$$

and let $\mathcal{F}_t$ be the class of functions $\mathcal{F}_t = \{f_{\beta_0,t}, \beta_0 \in \mathcal{B}_0\}$. Suppose that $\mathbb{E}\|X\|^p < \infty$ for some $p > 2$, and there exists $A > 0$ such that for all intervals $[a, b]$, $\mu_0([a, b]) \leq A|b - a|$.

1. *There exists $C > 0$ such that for all $t \in \mathbb{R}$ and $\delta > 0$,*

$$\log N_{[]}(\delta, \mathcal{F}_t) \leq C(1 + \log(\delta^{-1})).$$

2. *For arbitrary $q \in (0, 2)$, there exists $C > 0$ such that for all t such that $|t| > 1$, one has*

$$\sqrt{n}\mathbb{E}\sup_{f\in\mathcal{F}_t}\Big|\int f(d\mathbb{P}_n^X - d\mathbb{P}^X)\Big| \leq C\Big(|t|^{-\frac{p}{2}(1-\frac{q}{2})} + |t|^{\frac{pq}{2}} n^{-1/2}\Big).$$

3. *There exists $C > 0$ such that such that for all t such that $|t| \leq 1$, one has*

$$\sqrt{n}\mathbb{E}\sup_{f\in\mathcal{F}_t}\Big|\int f(d\mathbb{P}_n^X - d\mathbb{P}^X)\Big| \leq C.$$

4.2 Proof of Theorem 2

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

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

as the convolution distribution function and its empirical estimator, where F_n^X is the empirical distribution function in the sample $\mathcal{X}_n$ and F^X is the distribution function of X. Recall the definition of $\mathcal{D}$ from (11). It is proved in Azadkia and Balabdaoui (2024) that

$$\sup_{\beta \in \mathbb{R}^d} |\mathbb{D}_n(\beta) - \mathcal{D}(\beta)| \xrightarrow{\mathbb{P}} 0, \tag{12}$$

see the display before (A.10) in that paper. Note that although Theorem 5 in Azadkia and Balabdaoui (2024) assumes that $\mathcal{B}_0$ is reduced to a single point, this assumption is not used for the proof of the convergence in the last display, whence the convergence still holds under our assumptions. With similar arguments as for the proof of (A.10) in Azadkia and Balabdaoui (2024), where we just replace the set $\mathcal{O}$ by the closed ball of center β_0 and radius r, one obtains that for arbitrary $\beta_0 \in \mathcal{B}_0$ and $r > 0$,

$$0 = \mathcal{D}(\beta_0) < \inf_{\beta:\, d(\beta, \mathcal{B}_0) > r} \mathcal{D}(\beta). \tag{13}$$

Note that the corresponding proof in Azadkia and Balabdaoui (2024) uses the assumption that $\mathcal{B}_0$ contains a unique vector β_0 to claim that if

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

for some $\tilde{\beta}$, then we must have $\tilde{\beta} = \beta_0$. In our case, $\mathcal{B}_0$ may contain several vectors so we cannot use that argument. Instead, we use the fact that if the equality in the above display holds, then $\tilde{\beta}^\top X + \epsilon$ must have the same distribution as $\beta_0^\top X + \epsilon$, which by assumption (2) implies that $\tilde{\beta}^\top X$ has the same distribution as $\beta_0^\top X$, which in turn means that $\tilde{\beta} \in \mathcal{B}_0$ by definition of $\mathcal{B}_0$, and hence $d(\tilde{\beta}, \mathcal{B}_0) = 0$. The rest of the proof is similar to that in Azadkia and Balabdaoui (2024). Now, for arbitrary $r > 0$ it follows from (13) that we can find $\eta_r > 0$ such that

$$\inf_{\beta:\, d(\beta, \mathcal{B}_0) > r} \mathcal{D}(\beta) - \mathcal{D}(\beta_0) = \inf_{\beta:\, d(\beta, \mathcal{B}_0) > r} \mathcal{D}(\beta) > \eta_r.$$

Hence, it follows from (12), (13) and the definition of $\widehat{\beta}_n$, that for arbitrary $r > 0$, arbitrary $\beta_0 \in \mathcal{B}_0$, and $\widehat{\beta}_n \in \mathcal{B}_n$ one has

$$\begin{aligned} \limsup_{n \to \infty} \mathbb{P}\Big(d(\widehat{\beta}_n, \mathcal{B}_0) > r \Big) &\le \limsup_{n \to \infty} \mathbb{P}\Big(\mathbb{D}_n(\beta_0) \ge \inf_{\beta:\, d(\beta, \mathcal{B}_0) > r} \mathbb{D}_n(\beta) \Big) \\ &\le \mathbb{P}\Big(o_{\mathbb{P}}(1) > \eta_r \Big), \end{aligned}$$

which converges to zero as $n \to \infty$. Therefore, $d(\widehat{\beta}_n, \mathcal{B}_0)$ converges in probability to zero as $n \to \infty$. This proves the first claim.

We turn to the proof of the second claim. By the triangle inequality, it follows that for arbitrary $\beta_0 \in \mathcal{B}_0$ one has

$$W_1(\mu^*_{\widehat{\beta}_n}, \mu_0) \leq W_1(\mu^*_{\widehat{\beta}_n}, \mu^*_{\beta_0}) + W_1(\mu^*_{\beta_0}, \mu_0)$$

and therefore,

$$W_1(\mu^*_{\widehat{\beta}_n}, \mu_0) \leq \inf_{\beta_0 \in \mathcal{B}_0} W_1(\mu^*_{\widehat{\beta}_n}, \mu^*_{\beta_0}) + \sup_{\beta_0 \in \mathcal{B}_0} W_1(\mu^*_{\beta_0}, \mu_0). \tag{14}$$

We will consider both terms on the right-hand side separately, and show that they both converge in probability to zero as $n \to \infty$. To deal with the first term, note that by definition of the Wasserstein-1 distance, we have

$$W_1(\mu^*_{\widehat{\beta}_n}, \mu^*_{\beta_0}) = \inf_{Z_1, Z_2} E|Z_1 - Z_2|$$

where the infimum is taken over all pairs (Z_1, Z_2) such that $Z_1 \sim \mu^*_{\widehat{\beta}_n}$ and $Z_2 \sim \mu^*_{\beta_0}$. Taking the particular choice $Z_1 = \widehat{\beta}_n^\top X^*$ and $Z_2 = \beta_0^\top X^*$, we get

$$\begin{aligned} W_1(\mu^*_{\widehat{\beta}_n}, \mu^*_{\beta_0}) &\leq E|(\widehat{\beta}_n - \beta_0)^\top X^*| \\ &\leq \|\widehat{\beta}_n - \beta_0\| E\|X^*\|, \end{aligned} \tag{15}$$

where in the second inequality we use the Cauchy-Schwarz inequality. Taking the infimum over all possible β_0s on both sides yields

$$\inf_{\beta_0 \in \mathcal{B}_0} W_1(\mu^*_{\widehat{\beta}_n}, \mu^*_{\beta_0}) \leq d(\widehat{\beta}_n, \mathcal{B}_0) E\|X^*\|$$

The second term on the right-hand side is bounded in probability since, by definition of X^*, its expectation is finite:

$$\mathbb{E}E\|X^*\| = \mathbb{E}\Big(\frac{1}{n}\sum_{i=1}^n \|X_i\|\Big) = \mathbb{E}\|X\| < \infty$$

where we recall that E denotes the conditional expectation given $\{\mathcal{X}_n, \mathcal{Y}_n\}$. Since $d(\widehat{\beta}_n, \mathcal{B}_0)$ converges to zero in probability by the first claim of Theorem 2, we get

$$\inf_{\beta_0 \in \mathcal{B}_0} W_1(\mu^*_{\widehat{\beta}_n}, \mu^*_{\beta_0}) = o_{\mathbb{P}}(1).$$

It remains to show that for the second term on the right-hand side of (14), we have

$$\sup_{\beta_0 \in \mathcal{B}_0} W_1(\mu^*_{\beta_0}, \mu_0) = o_{\mathbb{P}}(1). \tag{16}$$

For all β_0, $\mu^*_{\beta_0}$ is the empirical distribution in the sample $\beta_0^\top X_1, \ldots, \beta_0^\top X_n$ whereas μ_0 is the common distribution of the i.i.d. variables $\beta_0^\top X_1, \ldots, \beta_0^\top X_n$. Moreover, it is assumed

that $\mu_0 \in \mathcal{C}(M, a)$ for some $M, a > 0$ so, with F_0 the distribution function corresponding to μ_0 and $Z \sim \mu_0$, we have

$$\begin{aligned}\int_{\mathbb{R}} \sqrt{F_0(t)(1-F_0(t))}dt &\le 2\int_0^\infty \sqrt{\mathbb{P}(|Z| \ge t)}dt \\ &\le 2\Big(1 + \int_1^\infty \sqrt{\frac{\mathbb{E}|Z|^{a+2}}{t^{a+2}}}dt\Big) \\ &\le 2\Big(1 + \int_1^\infty \frac{M^{1/2}}{t^{1+a/2}}dt\Big) \\ &\le C\end{aligned}$$

for some $C > 0$ that depends only on a and M. Hence, Theorem 3.2 in Bobkov and Ledoux (2019) implies that

$$\sup_{\beta_0 \in \mathcal{B}_0} \mathbb{E}W_1(\mu^*_{\beta_0}, \mu_0) \le Cn^{-1/2}.$$

Let $\eta \in (0,1)$ to be fixed later, N be the minimal number of balls in $\mathbb{R}^d$ with radius η that cover $\mathcal{B}_0$, and $\beta_1, \dots, \beta_N$ be the centers of such balls. Lemma 7 ensures that $\mathcal{B}_0$ is a compact subset of $\mathbb{R}^d$ so it follows from Lemma 2.5 in van de Geer (2000) that we can find $K > 0$ that depends only on the radius of $\mathcal{B}_0$ and d such that $N \le K\eta^{-d}$, and we have

$$\sup_{\beta_0 \in \mathcal{B}_0} W_1(\mu^*_{\beta_0}, \mu_0) \le \max_{1 \le j \le N} \Big(W_1(\mu^*_{\beta_j}, \mu_0) + \sup_{\beta \in \mathbb{R}^d:\ \|\beta - \beta_j\| \le \eta} W_1(\mu^*_\beta, \mu^*_{\beta_j})\Big).$$

Note that for given j, β_j need not to belong to $\mathcal{B}_0$. However, because the β_js define a minimal η-covering of $\mathcal{B}_0$, one can find $\theta_j \in \mathcal{B}_0$ such that $\|\beta_j - \theta_j\| \le \eta$ which implies that

$$\begin{aligned}\sup_{\beta_0 \in \mathcal{B}_0} W_1(\mu^*_{\beta_0}, \mu_0) &\le \max_{1 \le j \le N} \Big(W_1(\mu^*_{\theta_j}, \mu_0) + 2\sup_{\beta \in \mathbb{R}^d:\ \|\beta - \beta_j\| \le \eta} W_1(\mu^*_\beta, \mu^*_{\beta_j})\Big) \\ &\le \sum_{j=1}^N W_1(\mu^*_{\theta_j}, \mu_0) + 2\sup_{\beta, \beta' \in \mathbb{R}^d:\ \|\beta - \beta'\| \le \eta} W_1(\mu^*_\beta, \mu^*_{\beta'}).\end{aligned}$$

Similar to (15) one obtains

$$\sup_{\beta, \beta' \in \mathbb{R}^d:\ \|\beta - \beta'\| \le \eta} W_1(\mu^*_\beta, \mu^*_{\beta'}) \le \eta\|X^*\|$$

whence

$$\mathbb{E}\sup_{\beta, \beta' \in \mathbb{R}^d:\ \|\beta - \beta'\| \le \eta} W_1(\mu^*_\beta, \mu^*_{\beta'}) \le \eta\mathbb{E}\|X^*\| = \eta\mathbb{E}\|X\|.$$

Combining, we get

$$\begin{aligned}\mathbb{E}\sup_{\beta_0 \in \mathcal{B}_0} W_1(\mu^*_{\beta_0}, \mu_0) &\le N \sup_{\beta_0 \in \mathcal{B}_0} \mathbb{E}W_1(\mu^*_{\beta_0}, \mu_0) + 2\eta\mathbb{E}\|X\| \\ &\le C(\eta^{-d}n^{-1/2} + \eta)\end{aligned}$$

for some constant $C > 0$. We can choose η in such a way that $\eta \ll 1$ and $\eta \gg n^{-1/(2d)}$ so that the right-hand side converges to zero as $n \to \infty$. For such a choice of η we get (16), which concludes the proof of Theorem 2. ■

4.3 Proof of Theorem 3

Let us recall again the definition of $\mathcal{D}$ (aleady given in (11)):

$$\mathcal{D}(\beta) := \int \Big(F^Y(y) - \int F^\epsilon(y - \beta^\top x) dF^X(x)\Big)^2 dF^Y(y).$$

By Lemma 7, we know that $\mathcal{B}_0$ is compact. This implies that for arbitrary $\eta > 0$, the set $\mathcal{C}_\eta$ of all $\beta \in \mathbb{R}^d$ such that $d(\beta, \mathcal{B}_0) \leq \eta$ is a compact subset of $\mathbb{R}^d$. For all $\beta \in \mathcal{C}_\eta$, we denote by β_0 (which depends on β) an element of $\mathcal{B}_0$ that satisfies $\|\beta - \beta_0\| \leq \eta$. It is clear that there exists at least one such β_0 by compactness. In the identifiable case, it is proved in Azadkia and Balabdaoui (2024) that $\mathcal{D}$ is twice continuously differentiable on a small neighbourhood of (the unique) β_0, see the proof of Proposition 15 in that paper where $\mathcal{D}$ was denoted by ψ. A closer inspection of that proof shows that differentiability holds more generally on $\mathbb{R}^d$. Moreover, with similar arguments as in the proof of Proposition 15 in Azadkia and Balabdaoui (2024), we obtain that for all $\beta_0 \in \mathcal{B}_0$,

$$\frac{\partial^2 \mathcal{D}(\beta)}{\partial \beta \partial \beta^\top}\Big|_{\beta=\beta_0} = 2U(\beta_0),$$

and that second derivative is uniformly continuous on the compact set $\mathcal{C}_\eta$. Now, $\mathcal{D}(\beta) \geq 0$ for all $\beta \in \mathbb{R}^d$ with exact equality if $\beta \in \mathcal{B}_0$. Thus, the first derivative of $\mathcal{D}$ equals zero on $\mathcal{B}_0$. Using Taylor expansion, it follows that

$$\begin{aligned}
\mathcal{D}(\beta) &= (\beta - \beta_0)^\top U(\beta_0)(\beta - \beta_0) + o(\|\beta - \beta_0\|^2) \\
&\geq \|\beta - \beta_0\|^2 \lambda(\beta_0) + o(\|\beta - \beta_0\|^2) \\
&\geq \|\beta - \beta_0\|^2 \lambda_0 + o(\|\beta - \beta_0\|^2)
\end{aligned}$$

where

$$\lambda_0 = \inf_{\beta_0 \in \mathcal{B}_0} \lambda(\beta_0),$$

which is strictly positive by assumption. The small o-term is uniform in β since U is continuous, and whence uniformly continuous on the compact set $\mathcal{C}_\eta$. This implies that we can choose η small enough so that

$$o(\|\beta - \beta_0\|^2) \geq -\lambda_0 \|\beta - \beta_0\|^2/2$$

for all $\beta \in \mathcal{C}_\eta$. Hence,

$$\mathcal{D}(\beta) \geq \lambda_0 \|\beta - \beta_0\|^2/2$$

for all $\beta \in \mathcal{C}_\eta$, and

$$\mathcal{D}(\beta) \geq \lambda_0 \inf_{\beta_0 \in \mathcal{B}_0} \|\beta - \beta_0\|^2/2.$$

Next, we define $r_n = \sqrt{n}$ and

$$\phi_n(\delta) = \delta + \frac{\delta}{\sqrt{n}} + \frac{1}{\sqrt{n}}$$

for all $\delta > 0$. Thanks to Lemma 9 and Theorem 2, the conditions of Lemma 6 are satisfied with $\mathbb{M}_n = -\mathbb{D}_n$, $\mathbb{M} = -\mathcal{D}$, $C_0 = \sqrt{\lambda_0}$, $\alpha = 1$ and $\mathcal{B} = \mathcal{B}_0$, and the first claim in Theorem 3 follows.

We turn to the proof of the second claim. For this task, we fix $\varepsilon > 0$ arbitrarily small and we note that from the first claim in Theorem 3 we can find a constant $C_\varepsilon > 0$ such that

$$\mathbb{P}\Big(\sqrt{n\lambda_0}d(\widehat{\beta}_n, \mathcal{B}_0) \geq C_\varepsilon\Big) \leq \varepsilon/2.$$

Therefore, for every $C > 0$ we have

$$\begin{aligned}\mathbb{P}\Big(\sqrt{n}W_1(\mu^*_{\widehat{\beta}_n}, \mu_0) > C\Big) &\leq \mathbb{P}\Big(\sqrt{n}W_1(\mu^*_{\widehat{\beta}_n}, \mu_0) > C \text{ and } \sqrt{n\lambda_0}d(\widehat{\beta}_n, \mathcal{B}_0) < C_\varepsilon\Big) + \varepsilon/2 \\ &\leq \mathbb{P}\Big(\sqrt{n}\sup_{\beta:\, \sqrt{n\lambda_0}d(\beta,\mathcal{B}_0)<C_\varepsilon} W_1(\mu^*_\beta, \mu_0) > C\Big) + \varepsilon/2. \qquad (17)\end{aligned}$$

For all β such that $\sqrt{n\lambda_0}d(\beta, \mathcal{B}_0) < C_\varepsilon$ we denote by β_0 (which depends on β) an element of $\mathcal{B}_0$ that satisfies $\sqrt{n\lambda_0}\|\beta - \beta_0\| \leq C_\varepsilon$. It is clear that there exists at least one such β_0. The triangle inequality yields

$$W_1(\mu^*_\beta, \mu_0) \leq W_1(\mu^*_\beta, \mu^*_{\beta_0}) + W_1(\mu^*_{\beta_0}, \mu_0)$$

where we recall that $\mu^*_{\beta_0}$ is the distribution (given $X_1, \ldots, X_n$) of $\beta_0^\top X_{J_n}$ where J_n is independent of the observations and follows a uniform distribution on $\{1, \ldots, n\}$. Therefore,

$$\sup_\beta W_1(\mu^*_\beta, \mu_0) \leq \sup_\beta W_1(\mu^*_\beta, \mu^*_{\beta_0}) + \sup_\beta W_1(\mu^*_{\beta_0}, \mu_0)$$

where all suprema are taken over the set of βs such that $\sqrt{n\lambda_0}d(\beta, \mathcal{B}_0) < C_\varepsilon$. By definition of the Wasserstein distance, for the first term on the right-hand side, we have

$$\begin{aligned}\sup_\beta W_1(\mu^*_\beta, \mu^*_{\beta_0}) &\leq \sup_\beta E|\beta^\top X_{J_n} - \beta_0^\top X_{J_n}| \\ &\leq \sup_\beta \|\beta - \beta_0\| E\|X_{J_n}\|\end{aligned}$$

where E is the expectation with respect to J_n and using the Cauchy-Schwarz inequality for the second inequality. Letting

$$m_n = \frac{1}{n}\sum_{i=1}^n \|X_i\|,$$

we have $E\|X_{J_n}\| = m_n$ and therefore,

$$\sqrt{n}\sup_\beta W_1(\mu^*_\beta, \mu^*_{\beta_0}) \leq C_\varepsilon m_n \lambda_0^{-1/2}.$$

Note that $\mathbb{E}m_n = \mathbb{E}\|X\| < \infty$ by assumption, thus, m_n is bounded in probability and we get

$$\begin{aligned}\sqrt{n}\sup_\beta W_1(\mu^*_\beta, \mu_0) &\leq \lambda_0^{-1/2} O_{\mathbb{P}}(1) + \sqrt{n}\sup_\beta W_1(\mu^*_{\beta_0}, \mu_0) \\ &\leq \lambda_0^{-1/2} O_{\mathbb{P}}(1) + \sqrt{n}\sup_{\beta_0 \in \mathcal{B}_0} W_1(\mu^*_{\beta_0}, \mu_0). \qquad (18)\end{aligned}$$

Now, for the second term on the right-hand side, we have for arbitrary $\beta_0 \in \mathcal{B}_0$

$$W_1(\mu^*_{\beta_0}, \mu_0) = \int_{\mathbb{R}} |F^*_{\beta_0}(t) - F_0(t)| dt$$

where $F^*_{\beta_0}$ and F_0 are the distribution functions of $\mu^*_{\beta_0}$ and μ_0 respectively, see (Bobkov and Ledoux, 2019, Theorem 2.9). We have

$$F_0(t) = P(\beta_0^\top X \leq t) = 1 - P(\beta_0^\top X > t)$$

with X independent of the observations and

$$F^*_{\beta_0}(t) = \frac{1}{n}\sum_{i=1}^n \mathbb{1}_{\beta_0^\top X_i \leq t} = 1 - \frac{1}{n}\sum_{i=1}^n \mathbb{1}_{\beta_0^\top X_i > t}.$$

Let $\mathbb{P}_n^X$ be the empirical distribution of the sample $X_1, \ldots, X_n$ and $\mathbb{P}^X$ the distribution of X, for all β, x, $f_{\beta,t}(x) = \mathbb{1}_{\beta^\top x > t}$ if $t > 0$ and $f_{\beta,t}(x) = \mathbb{1}_{\beta^\top x \leq t}$ if $t \leq 0$, we have

$$|F^*_{\beta_0}(t) - F_0(t)| = \Big| \int f_{\beta_0,t} d\mathbb{P}_n^X - \int f_{\beta_0,t} d\mathbb{P}^X \Big|.$$

Combining yields

$$\sqrt{n} \sup_{\beta_0 \in \mathcal{B}_0} W_1(\mu^*_{\beta_0}, \mu_0) \leq \sqrt{n} \int_{\mathbb{R}} \sup_{\beta_0 \in \mathcal{B}_0} \Big| \int f_{\beta_0,t} (d\mathbb{P}_n^X - d\mathbb{P}^X) \Big| dt.$$

Hence, taking expectations

$$\sqrt{n}\mathbb{E} \sup_{\beta_0 \in \mathcal{B}_0} W_1(\mu^*_{\beta_0}, \mu_0) \leq E_1 + E_2 + E_3$$

where for $i \in \{1, 2, 3\}$ we define

$$E_i := \sqrt{n} \int_{D_i} \mathbb{E} \sup_{\beta_0 \in \mathcal{B}_0} \Big| \int f_{\beta_0,t} (d\mathbb{P}_n^X - d\mathbb{P}^X) \Big| dt$$

with the domains of integration being defined by

$$\begin{aligned} D_1 &:= [-1, 1] \\ D_2 &:= \{t : \ |t| \in (1, a_n]\} \\ D_3 &:= \{t : \ |t| > a_n\} \end{aligned}$$

where a_n is a sequence that tends to infinity as $n \to \infty$ and that will be choosen later. We deal separately with the three terms E_i. For E_1 we have

$$E_1 \leq 2\sqrt{n} \sup_{t \in [-1,1]} \mathbb{E} \sup_{\beta_0 \in \mathcal{B}_0} \Big| \int f_{\beta_0,t} (d\mathbb{P}_n^X - d\mathbb{P}^X) \Big|,$$

which is finite thanks to the third claim in Lemma 10. The second claim of the same lemma proves that for E_2 we have

$$E_2 \leq C \int_{|t|\in(1,a_n]} \Big(|t|^{-\frac{p}{2}(1-\frac{q}{2})} + |t|^{\frac{pq}{2}} n^{-1/2} \Big) dt.$$

Thus, choosing $q \in (0,2)$ small enough so that $\frac{p}{2}(1-\frac{q}{2}) > 1$ and taking $a_n = n^{1/(2+pq)}$ we get that

$$E_2 \leq 2C \Big(\frac{a_n^{1-\frac{p}{2}(1-\frac{q}{2})}}{1-\frac{p}{2}(1-\frac{q}{2})} + \frac{a_n^{1+\frac{pq}{2}}}{1+\frac{pq}{2}}\ n^{-1/2} \Big)$$

is also finite.

Finally, we deal with E_3 with the above choice for a_n. We recall that $\mathcal{B}_0$ is compact (see Lemma 7) and therefore, we can find $B > 0$ such that $\|\beta_0\| < B$ for all $\beta_0 \in \mathcal{B}_0$ and also, $|\beta_0^\top t| < B\|t\|$ for all t, using the Cauchy-Schwarz inequality. This implies that

$$\begin{aligned} E_3 &\leq \sqrt{n} \int_{|t|>a_n} \mathbb{E} \sup_{\beta_0\in\mathcal{B}_0} \Big(\int f_{\beta_0,t} d\mathbb{P}_n^X + \int f_{\beta_0,t} d\mathbb{P}^X \Big) dt \\ &\leq \sqrt{n} \int_{|t|>a_n} \mathbb{E} \Big(\int \mathbb{1}_{B\|x\|>|t|} d\mathbb{P}_n^X + \int \mathbb{1}_{B\|x\|>|t|} d\mathbb{P}^X \Big) dt \\ &= 2\sqrt{n} \int_{|t|>a_n} \mathbb{P}\Big(B\|X\| > |t| \Big) dt. \end{aligned}$$

For $p > 2$ such that $\mathbb{E}\|X\|^p < \infty$, by the Markov inequality it follows that there exists a constant $C > 0$ such that

$$E_3 \leq C\sqrt{n} \int_{|t|>a_n} |t|^{-p} dt = 2C\sqrt{n}\ \frac{a_n^{1-p}}{p-1}.$$

Since $a_n = n^{1/(2+pq)}$ for some $p > 2$ and a sufficiently small $q > 0$, we can choose $q > 0$ small enough so that $\sqrt{n} a_n^{1-p}$ tends to zero as $n \to \infty$. Hence, E_i is finite for all $i \in \{1,2,3\}$ and therefore, as a consequence of the Markov inequality we obtain

$$\sqrt{n} \sup_{\beta_0\in\mathcal{B}_0} W_1(\mu^*_{\beta_0}, \mu_0) = O_{\mathbb{P}}(1).$$

Hence, from (18) it follows that

$$\sqrt{n} \sup_{\beta} W_1(\mu^*_{\beta}, \mu_0) = O_{\mathbb{P}}(1).$$

Thus, we can find $C > 0$ large enough so that

$$\mathbb{P}\Big(\sqrt{n} \sup_{\beta} W_1(\mu^*_{\beta}, \mu_0) > C \Big) \leq \varepsilon/2,$$

and from (17) it follows that for arbitrary $\varepsilon > 0$ we can find $C > 0$ such that

$$\mathbb{P}\Big(\sqrt{n} W_1(\mu^*_{\widehat{\beta}_n}, \mu_0) > C \Big) \leq \varepsilon.$$

This completes the proof of Theorem 3. ■

4.4 Proof of Theorem 4

First note that

$$\begin{aligned}\|\widehat{f}_Z - f_Z\|_\infty &\leq \|\mu^*_{\widehat{\beta}_n} * K_h - \mu_0 * K_h\|_\infty + \|\mu_0 * K_h - f_Z\|_\infty \\ &=: A + B.\end{aligned}$$

The term A can be handled via Wasserstein duality. Indeed, setting $\phi_z(y) = K_h(z-y)$, one has

$$\begin{aligned}(\mu^*_{\widehat{\beta}_n} * K_h - \mu_0 * K_h)(z) &= \int \phi_z d(\mu^*_{\widehat{\beta}_n} - \mu_0) \\ &\leq \mathrm{Lip}(\phi_z) W_1(\mu^*_{\widehat{\beta}_n}, \mu_0) \\ &\leq \frac{L}{h^2} W_1(\mu^*_{\widehat{\beta}_n}, \mu_0) \\ &= O_{\mathbb{P}}(h^{-2} n^{-1/2}),\end{aligned}$$

where we used that $W_1(\mu^*_{\widehat{\beta}_n}, \mu_0) = O_{\mathbb{P}}(n^{-1/2})$ by Theorem 3, and that the kernel function is L-Lipschitz implying

$$|\phi_z(y_1) - \phi_z(y_2)| = \frac{1}{h}\Big|K\Big(\frac{z-y_1}{h}\Big) - K\Big(\frac{z-y_2}{h}\Big)\Big| \leq \frac{L}{h^2}|y_1 - y_2|$$

for all $y_1, y_2 \in \mathbb{R}$, thus, yielding $\mathrm{Lip}(\phi_z) = \sup_{y_1 \neq y_2} \frac{|\phi_z(y_1) - \phi_z(y_2)|}{|y_1 - y_2|} \leq \frac{L}{h^2}$. The term B is a bias term and to bound it we rewrite

$$\begin{aligned}(\mu_0 * K_h - f_Z)(z) &= \int K_h(z-y) f_Z(y) dy - f_Z(z) \\ &= \int \frac{1}{h} K\left(\frac{z-y}{h}\right)(f_Z(y) - f_Z(z)) dy \\ &= \int K(u)(f_Z(z - hu) - f_Z(z)) du\end{aligned}$$

where the third step is a change of variable $u = (z-y)/h$. Now, by the assumptions made on f_Z, Taylor's Theorem implies that for all z, u there exists some $\theta \in (0,1)$ such that

$$f_Z(z - hu) - f_Z(z) = -hu f'_Z(z) + \frac{(hu)^2}{2} f''_Z(z - \theta hu).$$

Since $\int uK(u)du = 0$, we obtain that for all $z \in \mathbb{R}$,

$$\begin{aligned}|(\mu_0 * K_h - f_Z)(z)| &= \left|\int K(u) \frac{(hu)^2}{2} f''_Z(z - \theta hu) du\right| \\ &\leq Ch^2 \int u^2 K(u) du = O(h^2),\end{aligned}$$

for some $C > 0$. Combining the bounds and since $h = h_n \asymp n^{-1/8}$ we conclude

$$\|\widehat{f}_Z - f_Z\|_\infty = O_{\mathbb{P}}(n^{-1/4}). \tag{19}$$

Now, we turn to the second claim. First, let us consider the decomposition

$$\begin{aligned}\widehat{f}_{Z|Y}(z|y_0) - f_{Z|Y}(z|y_0) &= \frac{f_\epsilon(y_0 - z)\widehat{f}_Z(z)}{\widehat{f}_Y(y_0)} - \frac{f_\epsilon(y_0 - z)f_Z(z)}{f_Y(y_0)} \\ &= f_\epsilon(y_0 - z)\Big(\frac{\widehat{f}_Z(z)}{\widehat{f}_Y(y_0)} - \frac{f_Z(z)}{\widehat{f}_Y(y_0)} + \frac{f_Z(z)}{\widehat{f}_Y(y_0)} - \frac{f_Z(z)}{f_Y(y_0)}\Big) \\ &= f_\epsilon(y_0 - z)\Big(\frac{\widehat{f}_Z(z) - f_Z(z)}{\widehat{f}_Y(y_0)} + f_Z(z)(\frac{1}{\widehat{f}_Y(y_0)} - \frac{1}{f_Y(y_0)})\Big) \\ &= f_\epsilon(y_0 - z)\Big(\frac{\widehat{f}_Z(z) - f_Z(z)}{\widehat{f}_Y(y_0)} - \frac{f_Z(z)}{f_Y(y_0)\widehat{f}_Y(y_0)}(\widehat{f}_Y(y_0) - f_Y(y_0))\Big).\end{aligned} \tag{20}$$

Now, recall that

$$\widehat{f}_Y(y) = \frac{1}{n}\sum_{i=1}^n f_\epsilon(y - \widehat{\beta}_n^\top X_i) = \int f_\epsilon(y - z)d\mu^*_{\widehat{\beta}_n}(z)$$

for $y \in \mathbb{R}$. Then,

$$\begin{aligned}|\widehat{f}_Y(y) - f_Y(y)| &= \left|\int f_\epsilon(y - z)d(\mu^*_{\widehat{\beta}_n}(z) - \mu_0(z))\right| \\ &\leq MW_1(\mu^*_{\widehat{\beta}_n}, \mu_0) = O_{\mathbb{P}}(n^{-1/2}),\end{aligned}$$

and hence

$$\|\widehat{f}_Y - f_Y\|_\infty = O_{\mathbb{P}}(n^{-1/2}). \tag{21}$$

Thus, with probability tending to 1, $\widehat{f}_Y(y_0) > f_Y(y_0)/2$, and it follows from (19) and (20) that

$$\begin{aligned}\|\widehat{f}_{Z|Y}(\cdot|y_0) - f_{Z|Y}(\cdot|y_0)\|_\infty &\leq \|f_\epsilon\|_\infty\Big(\frac{2\|\widehat{f}_Z - f_Z\|_\infty}{f_Y(y_0)} + \frac{2\|f_Z\|_\infty}{f_Y(y_0)^2}\|\widehat{f}_Y - f_Y\|_\infty\Big) \\ &= O_{\mathbb{P}}(n^{-1/4}) + O_{\mathbb{P}}(n^{-1/2}) = O_{\mathbb{P}}(n^{-1/4}),\end{aligned}$$

which concludes the proof. ■

4.5 Proof of Proposition 5

For the first part, it follows from the convergence rate in (21) that with probability tending to 1

$$\widehat{f}_Y(y_0) \geq f_Y(y_0)/2.$$

Now, using the identity in (20) we can write that

$$\begin{aligned}\big|\widehat{E}(Z|Y=y_0)-\mathbb{E}(Z|Y=y_0)\big| &= \Big|\frac{1}{\widehat{f}_Y(y_0)}\int z f_\epsilon(y_0-z)(\widehat{f}_Z(z)-f_Z(z))dz \\ &\quad - \frac{\widehat{f}_Y(y_0)-f_Y(y_0)}{\widehat{f}_Y(y_0)f_Y(y_0)}\int z f_\epsilon(y_0-z)f_Z(z)dz\Big| \\ &\le \frac{2\|\widehat{f}_Z-f_Z\|_\infty}{f_Y(y_0)}\int |z| f_\epsilon(y_0-z)dz \\ &\quad + \frac{2\|f_Z\|_\infty\|\widehat{f}_Y-f_Y\|_\infty}{f_Y(y_0)^2}\int |z| f_\epsilon(y_0-z)dz \\ &\le 2\Big(\frac{\|\widehat{f}_Z-f_Z\|_\infty}{f_Y(y_0)} + \frac{\|f_Z\|_\infty\|\widehat{f}_Y-f_Y\|_\infty}{f_Y(y_0)^2}\Big)\,(|y_0|+\mathbb{E}[|\epsilon|])\end{aligned}$$

with probability tending to 1. Note that above we used the fact that $\int |z| f_\epsilon(y_0-z)dz = \int |z-y_0+y_0| f_\epsilon(y_0-z)dz \le \mathbb{E}[|\epsilon|]+|y_0|$. Then, it follows from Theorem 4 that

$$\big|\widehat{E}(Z|Y=y_0)-\mathbb{E}(Z|Y=y_0)\big| = O_{\mathbb{P}}(n^{-1/4}).$$

For the second claim, let us write $\widehat{z}^* := \widehat{M}(Z|Y=y_0)$. We first prove that $\widehat{z}^* \xrightarrow{\mathbb{P}} z^*$. For some constant $K>0$, consider the event $\{|\widehat{z}^*-z^*|>K\}$. When this event occurs, we have by assumption that there exists some $\delta>0$ (depending on K) such that

$$f_{Z|Y}(z^*|y_0) - f_{Z|Y}(\widehat{z}^*|y_0) \ge \delta.$$

On the other hand, it holds that

$$|\sup_z \widehat{f}_{Z|Y}(z|y_0) - \sup_z f_{Z|Y}(z|y_0)| \le \|\widehat{f}_{Z|Y}(\cdot|y_0) - f_{Z|Y}(\cdot|y_0)\|_\infty \tag{22}$$

In fact, we have that for all z,

$$\begin{aligned}\widehat{f}_{Z|Y}(z|y_0) &\le f_{Z|Y}(z|y_0) + \|\widehat{f}_{Z|Y}(\cdot|y_0) - f_{Z|Y}(\cdot|y_0)\|_\infty \\ &\le \sup_z f_{Z|Y}(z|y_0) + \|\widehat{f}_{Z|Y}(\cdot|y_0) - f_{Z|Y}(\cdot|y_0)\|_\infty,\end{aligned}$$

implying that $\sup_z \widehat{f}_{Z|Y}(z|y_0) \le \sup_z f_{Z|Y}(z|y_0) + \|\widehat{f}_{Z|Y}(\cdot|y_0) - f_{Z|Y}(\cdot|y_0)\|_\infty$. The same argument can be used to obtain the inequality in the other direction. Then, Theorem 4 implies that

$$\widehat{f}_{Z|Y}(\widehat{z}^*|y_0) = \sup_z \widehat{f}_{Z|Y}(z|y_0) \xrightarrow{\mathbb{P}} \sup_z f_{Z|Y}(z|y_0) = f_{Z|Y}(z^*|y_0),$$

and hence with probability tending to 1,

$$\widehat{f}_{Z|Y}(\widehat{z}^*|y_0) + \frac{\delta}{2} \ge f_{Z|Y}(z^*|y_0) \ge f_{Z|Y}(\widehat{z}^*|y_0) + \delta,$$

from which we conclude that

$$\widehat{f}_{Z|Y}(\widehat{z}^*|y_0) \geq f_{Z|Y}(\widehat{z}^*|y_0) + \frac{\delta}{2}$$

with probability 1. Thus,

$$\begin{aligned}\mathbb{P}(|\widehat{z}^* - z^*| > K) &\leq \mathbb{P}\Big(\widehat{f}_{Z|Y}(\widehat{z}^*|y_0) - f_{Z|Y}(\widehat{z}^*|y_0) \geq \frac{\delta}{2}\Big) + o_{\mathbb{P}}(1) \\ &\leq \mathbb{P}\Big(\|\widehat{f}_{Z|Y}(\cdot|y_0) - f_{Z|Y}(\cdot|y_0)\|_\infty \geq \frac{\delta}{2}\Big) + o_{\mathbb{P}}(1) \to 0\end{aligned}$$

as $n \to \infty$. Since K was arbitrary, this proves that $\widehat{z}^*$ is consistent. This means that for $\eta > 0$ small,

$$\mathbb{P}(|\widehat{z}^* - z^*| \leq \eta) \to 1$$

as $n \to \infty$. When η is small enough, and the event $\{|\widehat{z}^* - z^*| \leq \eta\}$ occurs, then $\widehat{z}^*$ belongs to the small neighbourhood mentioned in the assumptions of the proposition. Fix $C > 0$ and note that

$$\mathbb{P}(|\widehat{z}^* - z^*| > Cn^{-1/(4\alpha)}) \leq \mathbb{P}(f_{Z|Y}(\widehat{z}^*|y_0) - f_{Z|Y}(z^*|y_0) \geq \lambda C^\alpha n^{-1/4}) + o_{\mathbb{P}}(1).$$

Moreover, it holds that

$$\begin{aligned}f_{Z|Y}(\widehat{z}^*|y_0) - f_{Z|Y}(z^*|y_0) &= f_{Z|Y}(\widehat{z}^*|y_0) - \widehat{f}_{Z|Y}(\widehat{z}^*|y_0) + \widehat{f}_{Z|Y}(\widehat{z}^*|y_0) - f_{Z|Y}(z^*|y_0) \\ &\leq \|\widehat{f}_{Z|Y}(\cdot|y_0) - f_{Z|Y}(\cdot|y_0)\|_\infty + |\sup_z \widehat{f}_{Z|Y}(\cdot|y_0) - \sup_z f_{Z|Y}(\cdot|y_0)| \\ &\leq 2\|\widehat{f}_{Z|Y}(\cdot|y_0) - f_{Z|Y}(\cdot|y_0)\|_\infty = O_{\mathbb{P}}(n^{-1/4}),\end{aligned}$$

where we used (22) in the last step. Therefore, it follows

$$\mathbb{P}(|\widehat{z}^* - z^*| > Cn^{-1/(4\alpha)}) \leq \mathbb{P}(\|\widehat{f}_{Z|Y}(\cdot|y_0) - f_{Z|Y}(\cdot|y_0)\|_\infty \geq \lambda C^\alpha n^{-1/4}/2) + o_{\mathbb{P}}(1) \to 0$$

as $C \to \infty$ and $n \to \infty$. This proves that $|\widehat{z}^* - z^*| = O_{\mathbb{P}}(n^{-1/(4\alpha)})$.

For the third claim, let us write again $\widehat{F}_{Z|Y}(z|y_0) := \int_{-\infty}^z \widehat{f}_{Z|Y}(x|y_0)dx$. Note that

$$\begin{aligned}\sup_{z\in\mathbb{R}} |\widehat{F}_{Z|Y}(z|y_0) - F_{Z|Y}(z|y_0)| &\leq \sup_{z\in\mathbb{R}} \int_{-\infty}^z |\widehat{f}_{Z|Y}(x|y_0) - f_{Z|Y}(x|y_0)|dx \\ &\leq \int_{\mathbb{R}} |\widehat{f}_{Z|Y}(z|y_0) - f_{Z|Y}(z|y_0)|dz \\ &= \int_{|z|\leq M} |\widehat{f}_{Z|Y}(z|y_0) - f_{Z|Y}(z|y_0)|dz \\ &\quad + \int_{|z|>M} |\widehat{f}_{Z|Y}(z|y_0) - f_{Z|Y}(z|y_0)|dz \\ &= I_1 + I_2\end{aligned}$$

For the first term I_1, we have that

$$I_1 \leq 2M\|\widehat{f}_{Z|Y}(\cdot|y_0) - f_{Z|Y}.(\cdot|y_0)\|_\infty.$$

For the second term, we have for $M > c_1$ that

$$
\begin{aligned}
I_2 &= \int_{|z|>M} |\widehat{f}_{Z|Y}(z|y_0) - f_{Z|Y}(z|y_0)|dz \\
&\leq \int_{|z|>M} \widehat{f}_{Z|Y}(z|y_0)dz + \int_{|z|>M} f_{Z|Y}(z|y_0)dz \\
&= 1 - \int_{|z|\leq M} \widehat{f}_{Z|Y}(z|y_0)dz + \int_{|z|>M} f_{Z|Y}(z|y_0)dz \\
&= 1 - \int_{|z|\leq M} \left(\widehat{f}_{Z|Y}(z|y_0) - f_{Z|Y}(z|y_0)dz\right) dz - \int_{|z|\leq M} f_{Z|Y}(z|y_0)dz \\
&\quad + \int_{|z|>M} f_{Z|Y}(z|y_0)dz \\
&\leq \int_{|z|\leq M} |\widehat{f}_{Z|Y}(z|y_0) - f_{Z|Y}(z|y_0)|dz + 2\int_{|z|>M} f_{Z|Y}(z|y_0)dz \\
&\leq 2M\|\widehat{f}_{Z|Y}(\cdot|y_0) - f_{Z|Y}(\cdot|y_0)\|_\infty + 2\int_{|z|>M} f_{Z|Y}(z|y_0)dz \\
&\leq 2M\|\widehat{f}_{Z|Y}(\cdot|y_0) - f_{Z|Y}(\cdot|y_0)\|_\infty + \frac{4c_2}{c_3}e^{-c_3M}
\end{aligned}
$$

where the last step follows from our assumption about the tail decay of $f_{Z|Y}(\cdot|y_0)$. Therefore, we have that

$$
\sup_{z\in\mathbb{R}} |\widehat{F}_{Z|Y}(z|y_0) - F_{Z|Y}(z|y_0)| \leq 4M\|\widehat{f}_{Z|Y}(\cdot|y_0) - f_{Z|Y}.(\cdot|y_0)\|_\infty + \frac{4c_2}{c_3}e^{-c_3M}.
$$

Now taking $M = M_n = \frac{1}{c_3}\log(\frac{4c_2}{c_3 n^{-1/4}}) = \frac{1}{4c_3}\log n + \frac{1}{c_3}\log(\frac{4c_2}{c_3})$ yields $\frac{4c_2}{c_3}e^{-c_3M} = n^{-1/4}$, and using the result of Theorem 4 it follows that

$$
\sup_{z\in\mathbb{R}} |\widehat{F}_{Z|Y}(z|y_0) - F_{Z|Y}(z|y_0)| = O_{\mathbb{P}}(n^{-1/4}\log n) + O_{\mathbb{P}}(n^{-1/4}) = O_{\mathbb{P}}(n^{-1/4}\log n). \tag{23}
$$

Now, consider a neighbourhood $N_\delta := [Q_{Z|Y}(p|y_0) - \delta, Q_{Z|Y}(p|y_0) + \delta]$ for some $\delta > 0$. By assumption, for any $z_1, z_2 \in N_\delta$, we have

$$
|F_{Z|Y}(z_2|y_0) - F_{Z|Y}(z_1|y_0)| = \Big|\int_{z_1}^{z_2} f_{Z|Y}(x|y_0)dx\Big| \geq |z_2 - z_1| \inf_{z\in N_\delta} f_{Z|Y}(z|y_0) > 0.
$$

Thus, since by (23) and local strict positivity of $f_{Z|Y}(\cdot|y_0)$ on N_δ we have $\widehat{Q}_{Z|Y}(p|y_0) \in N_\delta$ with probability tending to 1, we conclude that on this event

$$\begin{aligned}
|\widehat{Q}_{Z|Y}(p|y_0) - Q_{Z|Y}(p|y_0)| &\leq \frac{|F_{Z|Y}(\widehat{Q}_{Z|Y}(p|y_0)|y_0) - F_{Z|Y}(Q_{Z|Y}(p|y_0)|y_0)|}{\inf_{z\in N_\delta} f_{Z|Y}(z|y_0)} \\
&\leq \frac{|F_{Z|Y}(\widehat{Q}_{Z|Y}(p|y_0)|y_0) - \widehat{F}_{Z|Y}(\widehat{Q}_{Z|Y}(p|y_0)|y_0)|}{\inf_{z\in N_\delta} f_{Z|Y}(z|y_0)} \\
&\quad + \frac{|\widehat{F}_{Z|Y}(\widehat{Q}_{Z|Y}(p|y_0)|y_0) - F_{Z|Y}(Q_{Z|Y}(p|y_0)|y_0)|}{\inf_{z\in N_\delta} f_{Z|Y}(z|y_0)} \\
&\leq \frac{\sup_{z\in\mathbb{R}} |F_{Z|Y}(z|y_0) - \widehat{F}_{Z|Y}(z|y_0)|}{\inf_{z\in N_\delta} f_{Z|Y}(z|y_0)} \\
&= O_{\mathbb{P}}(n^{-1/4}\log n),
\end{aligned}$$

which concludes the proof. ■

4.6 Proof of the lemmas

We point out that our results generalize some results of Azadkia and Balabdaoui (2024) to the case of possibly non-identifiable β_0. Moreover, in the introduction of Azadkia and Balabdaoui (2024), it is assumed that the two samples $\mathcal{X}_n$ and $\mathcal{Y}_n$ are independent, an assumption that we do not make in the present paper. However, the proofs in Azadkia and Balabdaoui (2024), which we generalize here, do not use this assumption.

4.6.1 Proof of Lemma 1

For the equivalence in (2) to hold, it is sufficient to have the equivalence

$$Z + \epsilon \stackrel{d}{=} Z_0 + \epsilon \iff Z \sim \mu_0$$

for any real random variable Z which is independent of ϵ. Let

$$\kappa_Z(t) = \mathbb{E}\exp(itZ),\ t \in \mathbb{R}$$

denote the characteristic function of a real random variable Z. We recall that two given random variables have the same distribution if and only if their characteristic functions are equal. Moreover, if Z and ϵ are independent, then

$$\kappa_{Z+\epsilon} = \kappa_Z \kappa_\epsilon.$$

Hence, a sufficient condition for identifiability of μ_0 is that we have the equivalence

$$\kappa_Z \kappa_\epsilon = \kappa_{Z_0}\kappa_\epsilon \iff \kappa_Z = \kappa_{Z_0}$$

for all random variables Z that are independent of ϵ. It is proved in Meister (2009) (see the comments below display (2.23) on page 25) that a sufficient condition for the above equivalence to hold is that the set of zeros of κ_ϵ does not contain any open, nonempty interval as a subset. ■

4.6.2 Proof of Lemma 6

Assume for simplicity that $\widehat{\theta}_n$ truly maximizes the map $\theta \mapsto \mathbb{M}_n(\theta)$. For each n, $\Theta \setminus \mathcal{B}$ can be partitioned into shells

$$S_{j,n} = \{\theta : \ 2^{j-1} < C_0 r_n d(\theta, \mathcal{B}) \leq 2^j\}$$

with $j \in \mathbb{N}$. Similarly to the proof of Theorem 3.2.5 in van der Vaart and Wellner (2023) for every $\eta > 0$ we get

$$\mathbb{P}\Big(C_0 r_n d(\widehat{\theta}_n, \mathcal{B}) > 2^M\Big) \leq \sum_{j>M;\ 2^j \leq C_0 \eta r_n} \mathbb{P}\Big(\sup_{(\theta,\theta_0)\in S_{j,n}\times\mathcal{B}} (\mathbb{M}_n(\theta) - \mathbb{M}_n(\theta_0)) \geq 0\Big) + \mathbb{P}\Big(2d(\widehat{\theta}_n, \mathcal{B})) \geq \eta\Big).$$

Since $d(\widehat{\theta}_n, \mathcal{B})$ converges in outer probability to zero, the second probability on the right converges to zero as $n \to \infty$ for every $\eta > 0$. Thus, choose $\eta > 0$ small enough that the condition in (10) holds for every $\delta \leq \eta$. Then, for every $\theta_0 \in \mathcal{B}$, for every j involved in the sum and for every $\theta \in S_{j,n}$ we have

$$\mathbb{M}(\theta) - \mathbb{M}(\theta_0) \leq -C_0^2 d^2(\theta, \mathcal{B}) < \frac{-2^{2j-2}}{r_n^2}.$$

The rest of the proof is the same as that of van der Vaart and Wellner (2023) and is omitted. ■

4.6.3 Proof of Lemma 7

Let Σ denote the variance-covariance matrix of X and $\beta_0 \in \mathcal{B}_0$ be arbitrary. Then, $\mathrm{Var}(\beta_0^\top X) = \beta_0^\top \Sigma \beta_0 =: c$ with $c \in (0, +\infty)$, and because $\beta^\top X$ has the same distribution as $\beta_0^\top X$ for all $\beta \in \mathcal{B}_0$, we have $\|\beta\|_\Sigma = c$ for all $\beta \in \mathcal{B}_0$, where $\|\cdot\|_\Sigma$ is the norm defined by $\|\beta\|_\Sigma = \beta^\top \Sigma \beta$. Hence, $\mathcal{B}_0$ is a bounded set in $\mathbb{R}^d$.

Now, let (β_n) be a sequence in $\mathcal{B}_0$ that converges to some $\beta \in \mathbb{R}^d$ as $n \to \infty$. For every n, $\beta_n^\top X$ has the same distribution as $\beta_0^\top X$ where $\beta_0 \in \mathcal{B}_0$ is arbitrary, hence $\beta_n^\top X$ has the same characteristic function as $\beta_0^\top X$ for every n. Letting $n \to \infty$ shows that $\beta^\top X$ also has the same characteristic function as $\beta_0^\top X$, whence $\beta \in \mathcal{B}_0$. This means that $\mathcal{B}_0$ is a closed subset of $\mathbb{R}^d$, whence it is compact. The lemma follows. ■

4.6.4 Proof of Lemma 8

With similar arguments as in the proof of Proposition 11 in Azadkia and Balabdaoui (2024), we see that $\mathcal{G}$ is a subset of the convex hull of the class

$$\mathcal{H} := \{x \mapsto \phi_1(\beta^\top x + t - y) \times \phi_2(\beta_0^\top x + t - y),\ \beta, \beta_0 \in \mathbb{R}^d,\ t, y \in \mathbb{R}\},$$

and that $\mathcal{H}$ is a VC-class. Here, $\phi_1(u) = \mathbb{1}\{u \geq 0\}$ and $\phi_2(u) = \mathbb{1}\{u \leq 0\}$, $u \in \mathbb{R}$. Moreover, denoting by M a positive number such that the density f_ϵ of ϵ satisfies $f_\epsilon(t) \leq M$ for all $t \in \mathbb{R}$, for all β, β' such that $\|\beta - \beta'\| \leq \delta$ and all x, y we have

$$|g_{\beta,\beta',y}(x)| \leq M|(\beta - \beta')^\top x| \leq M\delta\|x\|$$

using the Cauchy-Schwarz inequality. This means that the function $G(x) := M\delta\|x\|$ is an envelope of $\mathcal{G}$. Since we have the same envelope as for their class $\mathcal{G}$, the second claim of the lemma follows from similar arguments as in the proof of Proposition 11 in Azadkia and Balabdaoui (2024), that uses finiteness of $J(1, \mathcal{G})$ and applies Theorem 2.14.1 of van der Vaart and Wellner (2023).

We turn to the proof of the first claim. It is proved in Section 4.6.3 that $\mathcal{B}_0$ is a compact set in $\mathbb{R}^d$. Hence, it is contained in the ball with center 0 and radius C, for some $C > 0$. For every $\beta_0 \in \mathcal{B}_0$ we can write

$$m_{\beta_0,y} = m_{0,y} + (m_{\beta_0,y} - m_{0,y})$$

so with

$$\mathcal{M}_0 := \{m_{0,y}, y \in \mathbb{R}\},$$

we have

$$\mathcal{M} \subset \mathcal{M}_0 + \mathcal{G}(C)$$

where $\mathcal{G}(C)$ is defined in the same way as $\mathcal{G}$ with δ replaced by C. Hence, we have

$$\mathbb{E}\|\mathbb{G}_n^X\|_{\mathcal{M}}^2 \leq \mathbb{E}\|\mathbb{G}_n^X\|_{\mathcal{M}_0}^2 + \mathbb{E}\|\mathbb{G}_n^X\|_{\mathcal{G}(C)}^2.$$

It follows from Proposition 11 in Azadkia and Balabdaoui (2024) that the first expectation on the right-hand side is finite, and we have shown above that the last expectation is also finite. This proves the first claim and completes the proof of Lemma 8. ■

4.6.5 Proof of Lemma 9

The proof is similar to the proof of Proposition 14 in Azadkia and Balabdaoui (2024), except that it uses Lemma 8 above, instead of Proposition 11 of the same paper, thus, we omit the details. ■

4.6.6 Proof of Lemma 10

We begin with the proof of the first claim. Lemma 7 ensures that $\mathcal{B}_0$ is a compact subset of $\mathbb{R}^d$ and therefore, we can find $B > 0$ such that $\|\beta_0\| < B$ for all $\beta_0 \in \mathcal{B}_0$. This means that the set $\mathcal{B}_0$ is included in the ball in $\mathbb{R}^d$ with diameter B in the Euclidean metric, so it follows from Lemma 2.5 in van de Geer (2000) that for all $\varepsilon > 0$, $\mathcal{B}_0$ can be covered by $N_\varepsilon := (4B + \varepsilon)^d \varepsilon^{-d}$ balls of radius ε. For a fixed $\varepsilon > 0$, we denote by B_j and β_j, with $j \in \{1, \ldots, N_\varepsilon\}$ the balls and their centers in such a covering, respectively. We then define the functions u_j and ℓ_j on $\mathbb{R}^d$ as follows:

$$u_j(x) = \sup_{\beta \in B_j} \beta^\top x \quad \text{and} \quad \ell_j(x) = \inf_{\beta \in B_j} \beta^\top x.$$

Then, for all $\beta \in B_j$, $t < 0$ and $x \in \mathbb{R}^d$ one has

$$\mathbb{1}_{u_j(x) \leq t} \leq f_{\beta,t}(x) \leq \mathbb{1}_{\ell_j(x) \leq t}.$$

This means that the brackets $[\mathbb{1}_{u_j\leq t}, \mathbb{1}_{\ell_j\leq t}]$, $j \in \{1,\dots,N_\varepsilon\}$, form a covering of the class $\mathcal{F}_t$. Similarly, if $t > 0$ then the brackets $[\mathbb{1}_{\ell_j\leq t}, \mathbb{1}_{u_j\leq t}]$, $j \in \{1,\dots,N_\varepsilon\}$, form a covering of the class $\mathcal{F}_t$. We now compute the length of the brackets in the $L_2(\mathbb{P}^X)$ norm. We have

$$\mathbb{E}|\mathbb{1}_{u_j(X)\leq t} - \mathbb{1}_{\ell_j(X)\leq t}|^2 = \mathbb{P}(\ell_j(X) \leq t < u_j(X)).$$

By the Cauchy-Schwarz inequality,

$$u_j(x) \leq \beta_j^\top x + \sup_{\beta\in B_j} (\beta-\beta_j)^\top x \leq \beta_j^\top x + \varepsilon\|x\|,$$

and similarly, $\ell_j(x) \geq \beta_j^\top x - \varepsilon\|x\|$ for all x. Hence,

$$\begin{aligned}\mathbb{E}|\mathbb{1}_{u_j(X)\leq t} - \mathbb{1}_{\ell_j(X)\leq t}|^2 &\leq \mathbb{P}(\beta_j^\top X - \varepsilon\|X\| \leq t < \beta_j^\top X + \varepsilon\|X\|)\\ &\leq \mathbb{P}(|\beta_j^\top X - t| \leq \varepsilon\|X\|).\end{aligned}$$

With $C_\varepsilon^p := \mathbb{E}\|X\|^p/\varepsilon$, it follows from the Markov inequality that

$$\mathbb{P}(\|X\| > C_\varepsilon) \leq \varepsilon$$

and therefore,

$$\begin{aligned}\mathbb{E}|\mathbb{1}_{u_j(X)\leq t} - \mathbb{1}_{\ell_j(X)\leq t}|^2 &\leq \mathbb{P}(|\beta_j^\top X - t| \leq \varepsilon C_\varepsilon) + \varepsilon\\ &= \mu_0([t-\varepsilon C_\varepsilon, t+\varepsilon C_\varepsilon]) + \varepsilon.\end{aligned}$$

By assumption there exists $A > 0$ such that for all intervals $[a,b]$ in $\mathbb{R}$, $\mu_0([a,b]) \leq A|b-a|$ so we conclude that

$$\begin{aligned}\mathbb{E}|\mathbb{1}_{u_j(X)\leq t} - \mathbb{1}_{\ell_j(X)\leq t}|^2 &\leq 2A\varepsilon C_\varepsilon + \varepsilon\\ &= 2A\varepsilon^{1-1/p}\mathbb{E}^{1/p}\|X\|^p + \varepsilon.\end{aligned}$$

This means that there exists $A' > 0$ (that depends only on A and $\mathbb{E}^{1/p}\|X\|^p$) such that

$$\mathbb{E}^{1/2}|\mathbb{1}_{u_j(X)\leq t} - \mathbb{1}_{\ell_j(X)\leq t}|^2 \leq \begin{cases} A'(\varepsilon^{1-1/p})^{1/2} & \text{for all } \varepsilon \in (0,1)\\ A'\varepsilon^{1/2} & \text{for all } \varepsilon > 1.\end{cases}$$

This means that the brackets $[\mathbb{1}_{u_j\leq t}, \mathbb{1}_{\ell_j\leq t}]$, $j \in \{1,\dots,N_\varepsilon\}$, form a covering of the class $\mathcal{F}_t$ with length $A'(\varepsilon^{1-1/p})^{1/2}$ if $\varepsilon \in (0,1)$ and $A'\varepsilon^{1/2}$ if $\varepsilon > 1$ and therefore,

$$N_{[]}((A'\varepsilon^{1-1/p})^{1/2}, \mathcal{F}_t) \leq N_\varepsilon \leq (4B+1)^d\varepsilon^{-d}$$

for all $\varepsilon \in (0,1)$. Taking $\delta = A'(\varepsilon^{1-1/p})^{1/2}$, this implies that for all $\delta \in (0, A')$,

$$N_{[]}(\delta, \mathcal{F}_t) \leq (4B+1)^d(A')^{2dp/(p-1)}\delta^{-2dp/(p-1)}$$

whence

$$\begin{aligned}\log N_{[]}(\delta, \mathcal{F}_t) &\leq \log\left((4B+1)^d(A')^{2dp/(p-1)}\right) + \frac{2dp}{p-1}\log\delta^{-1}\\ &\lesssim 1 + \log\delta^{-1}.\end{aligned}$$

Similarly, for $\delta > A'$ we take $\delta = A'\varepsilon^{1/2}$ and obtain that $N_{[]}(\delta, \mathcal{F}_t) \leq N_{(\delta/A')^2}$ so that $\log N_{[]}(\delta, \mathcal{F}_t) \lesssim 1 + \log \delta^{-1}$. The first claim of the lemma follows.

We turn to the proof of the second claim. It follows from the first claim of the lemma that for all $t \in \mathbb{R}$ and $\delta > 0$, the bracketing integral of the class $\mathcal{F}_t$ satisfies

$$\begin{aligned}\tilde{J}_{[]}(\delta, \mathcal{F}_t) &:= \int_0^\delta \sqrt{1 + \log N_{[]}(\varepsilon, \mathcal{F})} d\varepsilon \\ &\leq \int_0^\delta \sqrt{1 + C + C \log \varepsilon^{-1}} d\varepsilon.\end{aligned}$$

With arbitrary $q \in (0, 2)$, we have $\log \varepsilon^{-1} = (1/q) \log \varepsilon^{-q} \leq \varepsilon^{-q}/q$ and therefore,

$$\begin{aligned}\tilde{J}_{[]}(\delta, \mathcal{F}_t) &\leq \int_0^\delta \sqrt{1 + C} + \sqrt{C/q} \varepsilon^{-q/2} d\varepsilon \\ &= \delta \sqrt{1 + C} + \frac{\sqrt{C/q}}{1 - q/2} \delta^{1 - q/2} \\ &\lesssim \delta + \delta^{1 - q/2}. \end{aligned} \tag{24}$$

Now, it follows from the Cauchy-Schwarz inequality, that for all $t \in \mathbb{R}$ and $\beta_0 \in \mathcal{B}_0$ one has

$$0 \leq f_{\beta_0, t}(x) \leq \mathbb{1}_{B\|x\| > |t|}$$

and therefore, the Markov inequality yields that

$$\int f_{\beta_0, t}^2 d\mathbb{P}^X \leq \mathbb{P}(B\|X\| > |t|) \leq 1 \wedge \frac{B^p \mathbb{E}\|X\|^p}{|t|^p}$$

where $\mathbb{E}\|X\|^p$ is finite for $p > 2$ by assumption. Now we define $\delta^2 = C(1 \wedge |t|^{-p})$ for some $C > 1$ large enough so that the previous inequality implies that

$$\int f_{\beta_0, t}^2 d\mathbb{P}^X < \delta^2.$$

Combining with (24) and Theorem 2.14.17′ in van der Vaart and Wellner (2023) proves that

$$\sqrt{n} \mathbb{E} \sup_{f \in \mathcal{F}_t} \Big| \int f (d\mathbb{P}_n^X - d\mathbb{P}^X) \Big| \lesssim \tilde{J}_{[]}(\delta, \mathcal{F}_t) \Big(1 + \frac{\tilde{J}_{[]}(\delta, \mathcal{F}_t)}{\delta^2 \sqrt{n}} \Big) \tag{25}$$

for all $t \in \mathbb{R}$. Now, consider the case where $|t| > 1$. Then by definition, $\delta^2 = C|t|^{-p} < C$ and therefore, $\delta < C^{q/4} \delta^{1 - q/2}$ which implies that

$$\delta + \delta^{1 - q/2} \lesssim \delta^{1 - q/2} \lesssim |t|^{-\frac{p}{2}(1 - \frac{q}{2})}.$$

Hence, it follows from (24) combined with (25) that

$$\begin{aligned}\sqrt{n} \mathbb{E} \sup_{f \in \mathcal{F}_t} \Big| \int f (d\mathbb{P}_n^X - d\mathbb{P}^X) \Big| &\lesssim |t|^{-\frac{p}{2}(1 - \frac{q}{2})} \Big(1 + |t|^{\frac{p}{2}(1 + \frac{q}{2})} n^{-1/2} \Big) \\ &\lesssim |t|^{-\frac{p}{2}(1 - \frac{q}{2})} + |t|^{\frac{pq}{2}} n^{-1/2},\end{aligned}$$

which completes the proof of the second claim.

To deal with the case where $|t| \leq 1$, we note that by definition, $\delta^2 = C$ and $\delta + \delta^{1-q/2} \lesssim 1$. Hence, we can proceed exactly as above just replacing t by 1 and noting that $n^{-1/2} \lesssim 1$. This completes the proof of Lemma 10. ■

Acknowledgments

The research of the third author was supported by the FP2M federation (CNRS FR 2036).

References

Mona Azadkia and Fadoua Balabdaoui. Linear regression with unmatched data: A deconvolution perspective. *Journal of Machine Learning Research*, 25(197):1–55, 2024.

Fadoua Balabdaoui, Charles R Doss, and Cécile Durot. Unlinked monotone regression. *Journal of Machine Learning Research*, 22:172, 2021.

Fadoua Balabdaoui, Martin Slwaski, and Steffani Jonathan. Identifiability in unlinked linear regression: Some results and open problems. *arXiv:2507.14986*, 2025.

Sergey Bobkov and Michel Ledoux. *One-dimensional empirical measures, order statistics, and Kantorovich transport distances*, volume 261. American Mathematical Society, 2019.

Cristina Butucea and Catherine Matias. Minimax estimation of the noise level and of the deconvolution density in a semiparametric convolution model. *Bernoulli*, 11(2):309–340, 2005.

Claire Caillerie, Frédéric Chazal, Jérôme Dedecker, and Bertrand Michel. Deconvolution for the wasserstein metric and geometric inference. In *International Conference on Geometric Science of Information*, pages 561–568. Springer, 2013.

Alexandra Carpentier and Teresa Schlüter. Learning relationships between data obtained independently. In *Artificial Intelligence and Statistics*, pages 658–666. PMLR, 2016.

Marine Carrasco and Jean-Pierre Florens. A spectral method for deconvolving a density. *Econometric Theory*, 27(3):546–581, 2011.

Raymond J Carroll and Peter Hall. Optimal rates of convergence for deconvolving a density. *Journal of the American Statistical Association*, 83(404):1184–1186, 1988.

Raymond J Carroll and Peter Hall. Low order approximations in deconvolution and regression with errors in variables. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 66(1):31–46, 2004.

Fabienne Comte, Yves Rozenholc, and Marie-Luce Taupin. Penalized contrast estimator for adaptive density deconvolution. *Canadian Journal of Statistics*, 34(3):431–452, 2006.

Fabienne Comte, Yves Rozenholc, and M-L Taupin. Finite sample penalization in adaptive density deconvolution. *Journal of Statistical Computation and Simulation*, 77(11):977–1000, 2007.

Jérôme Dedecker and Bertrand Michel. Minimax rates of convergence for wasserstein deconvolution with supersmooth errors in any dimension. *Journal of Multivariate Analysis*, 122:278–291, 2013.

Jérôme Dedecker, Aurélie Fischer, and Bertrand Michel. Improved rates for Wasserstein deconvolution with ordinary smooth error in dimension one. *Electronic Journal of Statistics*, 9(1):234 – 265, 2015. doi: 10.1214/15-EJS997. URL `https://doi.org/10.1214/15-EJS997`.

Aurore Delaigle and Peter Hall. Parametrically assisted nonparametric estimation of a density in the deconvolution problem. *Journal of the American Statistical Association*, 109(506):717–729, 2014.

Cecile Durot and Debarghya Mukherjee. Minimax optimal rates of convergence in the shuffled regression, unlinked regression, and deconvolution under vanishing noise. *arXiv preprint arXiv:2404.09306*, 2024.

Bert Van Es and Hae-won Uh. Asymptotic normality of kernel-type deconvolution estimators. *Scandinavian Journal of Statistics*, 32(3):467–483, 2005.

Jianqing Fan. Asymptotic normality for deconvolution kernel density estimators. *Sankhyā Ser. A*, 53(1):97–110, 1991a. ISSN 0581-572X.

Jianqing Fan. On the optimal rates of convergence for nonparametric deconvolution problems. *The Annals of Statistics*, pages 1257–1272, 1991b.

Jianqing Fan. Deconvolution with supersmooth distributions. *Canadian Journal of Statistics*, 20(2):155–169, 1992.

Jianqing Fan. Adaptively local one-dimensional subproblems with application to a deconvolution problem. *The Annals of Statistics*, pages 600–610, 1993.

Piet Groeneboom and Geurt Jongbloed. Density estimation in the uniform deconvolution model. *Statistica Neerlandica*, 57(1):136–157, 2003.

Zhong Guan. Fast nonparametric maximum likelihood density deconvolution using bernstein polynomials. *Statistica Sinica*, 31(2):891–908, 2021.

Peter Hall and Peihua Qiu. Discrete-transform approach to deconvolution problems. *Biometrika*, 92(1):135–148, 2005.

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

Eyke Hüllermeier and Willem Waegeman. Aleatoric and epistemic uncertainty in machine learning: An introduction to concepts and methods. *Machine learning*, 110(3):457–506, 2021.

Ming Chung Liu and Robert L Taylor. A consistent nonparametric density estimator for the deconvolution problem. *Canadian Journal of Statistics*, 17(4):427–438, 1989.

Karim Lounici and Richard Nickl. Global uniform risk bounds for wavelet deconvolution estimators. *The Annals of Statistics*, 39(1):201 – 231, 2011. doi: 10.1214/10-AOS836. URL `https://doi.org/10.1214/10-AOS836`.

Jan Meis and Enno Mammen. Uncoupled isotonic regression with discrete errors. *Personal communication*, 2020.

Alexander Meister. Density estimation with normal measurement error with unknown variance. *Statistica Sinica*, pages 195–211, 2006.

Alexander Meister. *Deconvolution Problems in Nonparametric Statistics*. Springer Berlin Heidelberg, 2009.

Ashwin Pananjady, Martin J Wainwright, and Thomas A Courtade. Linear regression with shuffled data: Statistical and computational limits of permutation recovery. *IEEE Transactions on Information Theory*, 64(5):3286–3300, 2017.

Marianna Pensky and Brani Vidakovic. Adaptive wavelet estimator for nonparametric density deconvolution. *The Annals of Statistics*, 27(6):2033–2053, 1999.

Philippe Rigollet and Jonathan Weed. Uncoupled isotonic regression via minimum wasserstein deconvolution. *Information and Inference: A Journal of the IMA*, 8(4):691–717, 2019.

Martin Slawski and Emanuel Ben-David. Linear regression with sparsely permuted data. *Electronic Journal of Statististics*, 13(1):1–36, 2019.

Martin Slawski and Bodhisattva Sen. Permuted and unlinked monotone regression in r^ d: an approach based on mixture modeling and optimal transport. *Journal of Machine Learning Research*, 25(183):1–57, 2024.

Martin Slawski, Emanuel Ben-David, and Ping Li. Two-stage approach to multivariate linear regression with sparsely mismatched data. *The Journal of Machine Learning Research*, 21(1):8422–8463, 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.

Leonard A Stefanski and Raymond J Carroll. Deconvolving kernel density estimators. *Statistics*, 21(2):169–184, 1990.

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.

Sara van de Geer. *Empirical Processes in M-estimation*, volume 6. Cambridge university press, 2000.

Aad W. van der Vaart and Jon A. Wellner. *Weak Convergence and Empirical Processes*. Springer Series in Statistics. Springer Cham, 2023.

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*, 68(4):2509–2529, 2021.