Predicting Unobserved Individual-level Causal Effects

Christophe Gaillac*
University of Oxford

Job Market Paper, click here for the most recent version

December 19, 2023

Abstract

Measuring accurately heterogeneous effects is key for the design of efficient public policies. This paper focuses on predicting unobserved individual-level causal effects in linear random coefficients models, conditional on all the available data. In the application I consider, these "posterior effects" are the average effects of teachers' knowledge on their students' performance, conditional on both variables. I derive two nonparametric strategies for recovering these posterior effects, assuming independence between the effects and the covariates. The first strategy recovers the distribution of the random coefficients by a minimum distance approach, and then obtains the posterior effects from this distribution. The corresponding estimator can be computed using an optimal transport algorithm. The second approach, which is valid only for continuous regressors, directly expresses the posterior effects as a function of the data. The corresponding estimator is rate optimal. I discuss several extensions, in particular the relaxation of the independence condition. Finally, the application reveals large heterogeneity in the effect of teachers' knowledge, suggesting that we could substantially improve the cost-effectiveness of their training.

JEL codes: C14, H75, J24
Keywords: Empirical Bayes, teacher's value-added, random coefficients, optimal transport, generalized Tweedie's formula, voting analysis, inverse problem.

*Email: christophe.gaillac@economics.ox.ac.uk. Website: www.cgaillac.com. I am very grateful to Xavier D'Haultfœuille, Eric Gautier, Arnaud Maurel, Martin Weidner, and Bruno Crépon for their invaluable guidance and support. I also thank Steve Bond, Christian Bontemps, James Duffy, Jean-Pierre Florens, Emmanuel Guerre, Vishal Kamat, Maximilian Kasy, Pascal Lavergne, Michel Le Breton, Thierry Magnac, Sophocles Mavroeidis, Nour Meddahi, Vincent Pons and Karine Van der Straeten as well as seminar participants at Duke University, the University of Oxford, the Toulouse School of Economics, Queen Mary University of London, and at the Nuffield Postdoctoral seminar. The author acknowledges financial support from the grant ERC POEMH 337665. All errors are mine.

1 Introduction

Accurately estimating the impact of specific factors at the individual level is key to efficient microeconomic decision making. However, this heterogeneity is often unobserved. This paper therefore focuses on predicting the unobserved individual-level causal effects of specific covariates on an outcome within a linear random coefficients model, conditional on all available data. The latter includes the covariates but also the outcome. In the application I consider, these "posterior effects" (PE) are the average effects of teachers' knowledge of the program on their students' performance, conditional on both variables. The PE can be used to reveal important features of the heterogeneity of these effects. Importantly, they provide sufficient information to design efficient policies, such as identifying teachers who should be offered a training to improve their knowledge in order to maximize its impact on student scores.

My analysis focuses on the context in which covariates X_i are available to explain a (possibly noisy) measure of an outcome of interest Y_i. For example, among other teacher characteristics, their knowledge of the program might explain their value-added $\overline{Y}_i$, measured with noise from the average of student test scores Y_i. Leaving aside the noisy measurement problem for the moment, using a linear regression of Y_i on X_i will miss important heterogeneity in the effects. This paper therefore focuses on modeling the latter. It assumes that individual outcomes Y_i are explained by observed characteristics $X_i \in \mathbb{R}^p$ and a random vector Γ_i of unobserved heterogeneity, within the following linear random coefficients (RC) model,

$$Y_i = \Gamma_{1,i} + X_i^\top \Gamma_{-1,i}, \tag{1}$$

$$\Gamma_i \text{ and } X_i \text{ are independent}, \tag{2}$$

where $\Gamma_{1,i} \in \mathbb{R}$ is a random intercept and $\Gamma_{-1,i} \in \mathbb{R}^p$ is a random slope, which is the subvector of Γ_i without the first coordinate. In this context, there is little hope of recovering the individual heterogeneity Γ_i. However, the prediction of these effects conditional on the observed sample can be achieved using the posterior effects, which in the model (1)-(2) are defined as

$$\mathrm{PE}(x, y) := \mathbb{E}[\Gamma | (X, Y) = (x, y)]. \tag{3}$$

These are closely related to the Empirical Bayes (EB) framework, as $\widehat{\Gamma}_i^* := \mathrm{PE}(X_i, Y_i)$

can be interpreted as the mean squared error optimal estimates of the individual-level causal effects Γ_i.

The paper introduces new identification and tractable estimators for predicting these unobserved individual-level effects. The first method recovers the distribution of the coefficients and then uses Bayes' theorem to compute the posterior effects. Similar to Beran and Millar (1994), my method characterizes the RC distribution using minimum distance, but suggests the Wasserstein distance (see also Arellano and Bonhomme, 2023). This insight allows the reformulation of the target distribution as a barycenter of some observed distributions and the use of recent tools from optimal transport theory to solve this problem, known as the Generalized Wasserstein Barycenter (see Delon et al., 2022) (GWB hereafter). The second formulation, more efficiently expresses the PE directly as a function of the data, but it is only applicable when the covariates are continuous. I call this a *Generalized Tweedie's* formula (GT, hereafter), as it extends the original shrinkage (see Robbins, 1956; Efron, 2011) to this context with covariates.

In addition to its relevance in this context, model (1)-(2) serves as a foundation for more complex panel data models used in the literature on Teachers' Value Added (TVA) or economic mobility. Importantly, it also provides a new perspective and solution to the *ecological inference* problem (see, *e.g.*, King, 1997), one simple but striking illustration being the prediction of the probability of voting by race for a given precinct or county, using only census and election results data.

I explore alternatives to the independence assumption (2) that provide some robustness to it. When additional covariates are available, a middle ground between the model (1)-(2) and the one describing the heterogeneity of the effects using only covariates (see, *e.g*, Athey and Imbens, 2016; Athey et al., 2019), is to assume that the effects are determined by nonlinear functions of these observed covariates and additively separable unobservables (see, *e.g*, Breunig, 2021). An alternative is to use additional known controls, conditioning on which the independence between RCs and regressors holds. When instruments are available, a more involved alternative is to identify a control variable (Florens et al., 2008; Masten and Torgovitsky, 2016; Newey and Stouli, 2020), conditioning on which this independence holds. The final extension relaxes the baseline assumption to allow for distributions that are in a neighborhood

of the independent joint distribution, using the conditional partial independence introduced in Masten and Poirier (2018), and provide bounds on the PE.

The identification results are constructive, providing practical estimators for the PE in both the GWB and GT formulations. Under classical assumptions on the smoothness of the underlying RCs distribution, I show that the latter estimator is optimal in the minimax sense, up to logarithmic factors in the rates of convergence, and its tuning parameters are selected from the data (see, *e.g.*, Tsybakov, 2008; Giné and Nickl, 2016). I also show asymptotic normality. I provide a simple estimator for the GWB formulation and prove its consistency. Using discretization, it is possible to use this estimator even when the distribution of the regressors is continuous.

I apply these methods to study the sensitivity of TVA to the teachers' knowledge of the program using data from Pakistan, extending the analysis of Bau and Das (2020). In this panel data context, explaining TVA in terms of its time-invariant characteristics is usually done in a second step using linear regression of the fixed effects. They show that teachers' program knowledge is predictive of their performance, but this analysis lacks a description of the heterogeneity of these effects. In particular, my methods show that those teachers who have less to gain from increasing their knowledge are also those who have relatively important value added from other sources. Importantly, my method also identifies teachers with low knowledge and for whom an increase in knowledge is predicted to have a large impact on performance. As a consequence, I show that a personal development policy that takes this heterogeneity into account and targets this latter population would yield important efficiency gains relative to one that neglects it (up to about 31% when treating 10% of the sample), which is all the more important because school systems typically allocate 3% to 5% of their total budget to support training programs.

Related literature

The EB literature (see, *e.g.*, Robbins, 1964; Efron, 2012) is now large on settings without covariates (see, *e.g.*, James and Stein, 1992; Jiang and Zhang, 2009; Brown and Greenshtein, 2009; Johnstone and Silverman, 2004; Efron, 2011; Ignatiadis and Wager, 2022). This has found many policy-relevant applications in economics, where my methods also apply: on the value added of teachers, schools, and services (Rock-

off, 2004; Jacob and Lefgren, 2008; Rothstein, 2010; Chetty et al., 2014a; Angrist et al., 2017; Gilraine et al., 2020), neighborhood effects on intergenerational mobility or mortality (Chetty and Hendren, 2018; Finkelstein et al., 2021; Bonhomme and Weidner, 2022), or the study of discrimination (Kline et al., 2022). However, EB analysis of the case with covariates has been less explored (see, *e.g.*, Fay and Herriot, 1979; Cohen et al., 2013; Ignatiadis and Wager, 2019; Montiel Olea et al., 2021; Armstrong et al., 2022). Both Ignatiadis and Wager (2019) and Armstrong et al. (2022) focus on posterior estimation of the individual parameters Y_i in a setting with noise, but do not consider the heterogeneous individual effects of the covariates. Ignatiadis and Wager (2019) use covariates that enter flexibly and nonlinearly, at the cost of imposing strong restrictions on the unobserved heterogeneity which I do not make. This paper thus takes a complementary view, where I rely on the linear structure but focus on the complex heterogeneity and dependence between the different RCs. This paper is also related to Bonhomme and Weidner (2022), as they consider the average of these posterior effects, although not in a RCs model, allowing for misspecification and searching for estimators that have the least amount of bias.

We can think of the setup associated to (1)-(2) as repeatedly sampling the RCs from an unknown distribution F_Γ. Each them, combined with X_i, then generates an observation Y_i following a distribution $F_{Y|X}$. We then want to make inference that would be direct if F_Γ were known. In this context, the EB literature distinguishes between strategies based on modeling F_Γ, called G-modeling (Jiang and Zhang, 2009; Koenker and Mizera, 2014; Gu and Koenker, 2017; Gilraine et al., 2020), and those based on directly modeling the observed distribution $F_{Y|X}$, called F-modeling (Brown and Greenshtein, 2009; Efron, 2011, 2014). On the one hand, my GWB approach innovates by bringing tools from the optimal transport literature into the G-modeling strategy when covariates are available. On the other hand, my GT formulation uses F-modeling and relates it to the literature on RC models in econometrics.

RCs models and specifically linear ones have a long tradition in econometrics (see, *e.g.*, Beran and Hall, 1992; Beran and Millar, 1994; Beran et al., 1996; Masten, 2017; Hoderlein et al., 2017; Newey and Stouli, 2018; Dunker et al., 2019; Breunig, 2021; Gaillac and Gautier, 2022). Hoderlein and Mammen (2007, 2009); Hoderlein and Sasaki (2013); Chernozhukov et al. (2015) are the closest in terms of object interest,

providing identification and estimation for posterior marginal effects in nonseparable models when X is continuous and for individuals with $X = x$ and Y being a conditional quantile of Y given $X = x$, which are the derivatives of this quantile. An important feature of RCs models like (1)-(2), is that the variation of the regressors X_i is key, as it limits the size of the class and type of distributions Γ_i that can be identified (see, *e.g.*, Gaillac and Gautier, 2021b). In Gaillac and Gautier (2022), we study the minimax rates of convergence for estimating the density of the coefficients, in the difficult case where the regressors are bounded but continuous. Estimating PE is simpler but yields faster rates of convergence. Appendix G.3 studies identification in the linear *system* of RCs equations model (see Masten, 2017; Kasy, 2022).

Given the importance of the variation of the regressors in RC models, an important point is thus that the GWB formulation also allows more generally to estimate the distribution F_Γ with discrete covariates. I exploit advances in the so-called Wasserstein barycenters (Agueh and Carlier, 2011; Cuturi and Doucet, 2014; Delon et al., 2022; Carlier et al., 2022) by linking them to the inverse problem implied by model (1)-(2). The optimal transport literature has now found many applications in economics and econometrics (see, *e.g.*, Galichon and Henry, 2011; Chernozhukov et al., 2017; Galichon, 2018; D'Haultfoeuille et al., 2021; D'Haultfœuille et al., 2022; Gunsilius, 2023). Recent advances have made optimal transport problems computationally tractable (Cuturi, 2013; Peyré et al., 2019).

Finally, my approach allows to revisit the description of the impact of covariates on the population of interest in the applications. Extending the analysis of the impact of teachers' knowledge of the program on their performance in developing countries (see, *e.g.*, Bold et al., 2017; Bau and Das, 2020), my study of the heterogeneity of these effects shows how to gain efficiency in the allocation of on-the-job training. This makes it a possible alternative to the dismissal and retention policies discussed in the literature (see, *e.g.*, Hanushek et al., 2009; Chetty et al., 2014b; Gilraine et al., 2020). It also provides new fully nonparametric tools to perform ecological inference, extending a large literature in political science (see, *e.g.*, Goodman, 1959; King, 1997; Rosen et al., 2001; Imai et al., 2008; Frogner and Poggio, 2019). Our robustness analyses also provide some new answers to the criticisms that have been formulated (Gelman et al., 2001; Tam Cho, 1998; Tam Cho and Gaines, 2004; Wakefield, 2004).

Organization of the paper

The rest of the paper is organized as follows. In Section 2, I first describe contexts where posterior effects provide useful and sufficient information for the decision making. Then, in Section 3, I show how to identify these effects, considering various extensions. Section 4 details the different inference results for the two estimators and describes implementation. Section 5 then shows that my methods are empirically relevant for estimating the heterogeneity of TVAs with respect to teachers' own knowledge, and discusses the implications for policy learning. Section 6 concludes. The Appendix contains the main proofs, additional Monte Carlo simulation results, and the Tweedie's formula slightly extended for completeness. Specifically, Section G extends identification and estimation to the *ecological inference*. Section G.7 provides a real dataset validation of my method using a comparison with ground truth, focusing on the estimation of turnout by race. Finally, my methods will soon be compiled into a companion R package, `RegPE`, interfacing the Python library `POT` Flamary et al. (2021) for the optimal transport part, and will be available on CRAN-R.

2 Why considering posterior effects?

The posterior effects are only one feature of the unobserved heterogeneity in the model. However, there are relevant frameworks where the PE are *sufficient* to derive optimal decisions. All variables in this section are individual i specific, hence the index i is omitted hereafter.

2.1 Estimation under mean squared error

The simplest framework is the estimation of individual Γ effects. Consider sampling a Γ from the true unknown distribution of unobserved effects F_Γ. Combined with X, it generates the outcome Y according to the model (1)-(2). A standard goal in prediction is to find a *estimator* $p(X, Y)$ that depends on the data (X, Y) and minimizes the mean squared error:

$$R(p, \Gamma) = E_\Gamma \left[(\Gamma - p(X, Y))^2 \right].$$

Since Γ is actually random, it is more relevant to find p^* that minimizes the expected average (Bayes) risk under F_Γ,

$$R(p, F_\Gamma) = \int E_g[(g - p(X,Y))^2] dF_\Gamma(g). \tag{4}$$

In a Bayesian context, F_Γ would be the prior distribution. With this standard mean-squared error objective (4), the posterior effects $p^*(X,Y) = \text{PE}(X,Y) = E[\Gamma|(X,Y)]$ are optimal decisions, *i.e.*, minimizing $R(p, F_\Gamma)$.

This motivates the analysis of the MSE in Section 4. For this inference part, the goal is then to find a procedure $\widehat{p}$ that comes as close as possible to the oracle estimator p^* when $n \to \infty$, while being robust to the worst possible case of the distribution F_Γ. This is the minimax approach that I follow in Section 4.

2.2 Treatment allocation on X but targeting Y

In the second framework, I consider two time periods. In the first one, t, the policy maker collects some data (X_t, Y_t) about how a scalar variable X affects an outcome of interest Y. In the second period $t+1$, the policy maker can assign individuals to a treatment D that directly affects the variable X. However, this treatment is only one way of indirectly affecting the outcome of interest Y_{t+1}. In this context, the goal is to *ex-ante* design this allocation D to maximize the average impact on Y_{t+1}, under some capacity constraints.[1]

The planner's problem is then to find a function of the data $p(x,y) = E(D|(X_t, Y_t) = (x,y))$ which is the probability that an individual with the characteristics $(X_t, Y_t) = (x,y)$ is assigned to the treatment $D = 1$. Let $X_{t+1}(1)$ and $X_{t+1}(0)$ (respectively $Y_{t+1}(X_{t+1}(1))$ and $Y_{t+1}(X_{t+1}(0))$) denote the potential outcomes X (respectively Y) with and without the treatment: $X_{t+1} = DX_{t+1}(1) + (1-D)X_{t+1}(0)$.

This problem can be rewritten

$$\max_{r.v.\ D} \mathbb{E}(Y_{t+1}(X_{t+1}(D))) \quad \text{s.t.} \quad c = \mathbb{E}(D),$$

[1] By *ex-ante*, I mean before observing the outcome Y_{t+1} of this experiment performed on X_{t+1}, otherwise the optimal policy rule could be obtained using the literature on optimal policy learning with Y_{t+1} (see, *e.g.*, Manski, 2004; Stoye, 2009; Kitagawa and Tetenov, 2018).

where c is a capacity constraint that limits the number of people treated. This is equivalent to finding a measurable function $p : \text{Supp}(X_t, Y_t) \to [0, 1]$, where $\text{Supp}(X_t, Y_t)$ is the support of (X_t, Y_t), that maximizes

$$\begin{aligned} \max_{p(\cdot)} \ & \mathbb{E}\left[Y_{t+1}(X_{t+1}(0)) + p(X_t, Y_t)\left(Y_{t+1}(X_{t+1}(1)) - Y_{t+1}(X_{t+1}(0))\right)\right] \\ \text{s.t.} \quad & c = \mathbb{E}(p(X_t, Y_t)), \end{aligned} \tag{5}$$

Let precise the model in this context.

Assumption 1 *Consider the model*

$$Y_t = \Gamma_{1,t} + \Gamma_{2,t} X_t,$$

where Γ_t *and* X_t *are independent, and*

1. *the individual-level causal effects of* X_{t+1} *on* Y_{t+1} *are independent of the treatment effects on* X_{t+1}*, conditionally on the past values of* X_t *and* Y_t*, i.e,*

$$\Gamma_{t+1} \perp\!\!\!\perp (X_{t+1}(1) - X_{t+1}(0)) \mid X_t, Y_t,$$

2. *the individual-level causal effects* $\Gamma_{2,t}$ *is mean time invariant conditional on the data* $\mathbb{E}(\Gamma_{2,t+1}|X_t, Y_t) = \mathbb{E}(\Gamma_{2,t}|X_t, Y_t)$.

Under this Assumption 1, where the assumptions 1.1 and 1.2 seem reasonable in some contexts,[2] the problem (5) can be rewritten as a function of the conditional average treatment effect on X_{t+1}, $\text{CATE}(X_t, Y_t) = \mathbb{E}(X_{t+1}(1) - X_{t+1}(0)|X_t, Y_t)$, and the posterior effects

$$\begin{aligned} \max_{p(\cdot)} \quad & \mathbb{E}\left[p(X_t, Y_t)\ \text{PE}(X_t, Y_t)\ \text{CATE}(X_t, Y_t)\right] \\ \text{s.t.} \quad & c = \mathbb{E}(p(X_t, Y_t)). \end{aligned} \tag{6}$$

The optimal decision rule thus takes the form

$$p(X_t, Y_t) = \mathbb{1}\left\{\text{PE}(X_t, Y_t)\ \text{CATE}(X_t, Y_t) \geq \gamma\right\}, \tag{7}$$

[2]Note that if additional variables Z are available, these independence restrictions can be relaxed, in the spirit of conditional unconfoundedness, see Section 3.3.

where γ is such that $c = \mathbb{E}(p(X_t, Y_t))$.

In words and given a sample, this optimal rule simply translates into assigning to the treatment the $c\%$ of individuals with the best predicted effects on Y using past data: $\text{PE}(X_t, Y_t)$ $\text{CATE}(X_t, Y_t)$ in our sample.

In this decision, $\text{CATE}(X_t, Y_t)$ can be estimated from an auxiliary experiment using standard techniques, *e.g.*, from the machine learning literature for estimating heterogeneous treatment effects (see, *e.g.*, Athey et al., 2019), where Y_t is potentially included in the set of covariates. Importantly, equation (7) shows that the PE are necessary but also sufficient for the decision making in this context. This important motivation is further developed in the application in Section 5.3.

Other decisions where my characterization of F_Γ can be used. Of course, there are other decision problems where PE are not sufficient. However, in this paper I impose assumptions such as the distribution F_Γ is identified and my characterization of it in Proposition 1 and Theorem 1 using the Wasserstein minimum distance can then be used in these more general contexts to compute other functionals.

Specifically, this is the case for analogues of all the compound decision problems described in Gilraine et al. (2020); Gu and Koenker (2023) that require more than the PE. My strategy based on Wasserstein barycenter computation and optimal transport tools is an alternative to NPMLE estimation (see, *e.g.*, Jiang and Zhang, 2009; Koenker and Mizera, 2014). In the multivariate case, which is my context by definition, the latter might not be *a priori* as tractable (see Soloff et al., 2021), even if in some cases the geometry of the problem can be used to gain tractability, as done in Gu and Koenker (2022) for the binary outcome case. This allows to consider different objectives other than average welfare in the second decision framework above.

3 Identification of posterior effects

Before presenting my main identification results, I introduce some notation that will be used throughout the paper. Let $\cdot$, $\star$ denote a variable in a function. For two random vectors X and Y, $\mathbb{P}_{Y|X=x}$, $f_{Y|X=x}$, and $F_{Y|X=x}$ denote respectively the conditional probability, density, and cumulative distributions. For a random vector X,

I let $\varphi_X : t \mapsto E(e^{it^\top X})$ denote its characteristic function and $\text{Supp}(X)$ its support. For a measurable set $\mathcal{S} \subset \mathbb{R}^p$ and a function μ from $\mathcal{S}$ to $[0, \infty]$, $L^2(\mu)$ is the space of complex-valued square integrable functions equipped with the scalar product $\langle f, g \rangle_{L^2(\mu)} = \int_{\mathcal{S}} f(x)\overline{g}(x)\mu(x)dx$. This is denoted by $L^2(\mathcal{S})$ when $\mu = 1$. For $d \geq 1$, denote the Fourier transform of $f \in L^1\left(\mathbb{R}^d\right) \cap L^2\left(\mathbb{R}^d\right)$ by $\mathcal{F}[f](x) = \int_{\mathbb{R}^d} e^{ib^\top x} f(b)db$. Let $\otimes$ denote the product of functions (*e.g.*, $W^{\otimes d}(b) = \prod_{j=1}^d W(b_j)$) or measures. I also denote by $\mathcal{P}_d(\mathcal{S})$ the set of Borel probability measures on $\mathcal{S}$ with finite d first moments, and by $\mathcal{P}_{a.c.}(\mathcal{S})$ the one that are absolutely continuous with respect to the Lebesgue measure. I assimilate hereafter probability measures on $\mathbb{R}^p$ with their cdf, so I may write for instance $F \in \mathcal{P}_d(\mathcal{S})$.

3.1 In the baseline cross-section linear RC model

I first consider the baseline equation

$$Y_i = \Gamma_{1,i} + X_i^\top \Gamma_{-1,i}, \tag{8}$$

and maintain the following assumption, discussing relaxations in Section 3.3.

Assumption 2 $\Gamma_i \perp\!\!\!\perp X_i$.

In this context, the objects of interest are the posterior effects defined in (3), which are specific nonparametric regression functions of an *unobserved* variable Γ_i. I provide conditions under which the distribution of F_Γ is identified, which are stronger than necessary for the identification of the posterior effects, but which allow to compute more general posterior moments or functional that I also discuss, as well as to obtain an expression valid for discrete X.

The identification of the distribution F_Γ relies on a trade-off between the assumptions made about the support of the regressors X and those made about the distribution of Γ. I provide two constructive characterizations depending on these assumptions, and refer to Gaillac and Gautier (2021b) for sharper conditions than these ones under which F_Γ is identified.

Assumption 3 *Assume either that*

(A) *the distribution of* Γ *belongs to* $\mathcal{P}_d(\mathbb{R}^{p+1})$ *and is identified from the knowledge of its first* $d < \infty$ *moments, while the support of* X *contains the product* $\prod_{k=1}^{p} V_k$*, where* V_k *contains* $\kappa_k \geq d+1$ *points;*

(B) *or the distribution of* Γ *admits a density* $f_\Gamma \in L^2(W^{\otimes(p+1)})$*, where* $W := e^{|\cdot|/R}$*,* $R > 0$*, while the support of* X *contains a nonempty interior.*

The case (A) includes the empirically relevant case where f_Γ is continuous but parametric and identified from its first d moments. Common examples are finite Gaussian mixtures (see, *e.g.*, Améndola et al., 2015, for precise values of d ensuring identification). The case (B) means that the tails of f_Γ are not heavier than those of the exponential distribution. Indeed, we have, for all $\epsilon \in (0,1)$ and $k = 1, \ldots, p+1$, for $\lambda = (1-\epsilon)/(2R)$, by the Cauchy-Schwarz inequality,

$$E(e^{\lambda \Gamma_k}) \leq E(e^{\lambda |\Gamma_k|}) \leq \|f_\Gamma\|_{L^2(W^{\otimes(p+1)})} (2R/\epsilon)^{(p+1)/2} < \infty. \tag{9}$$

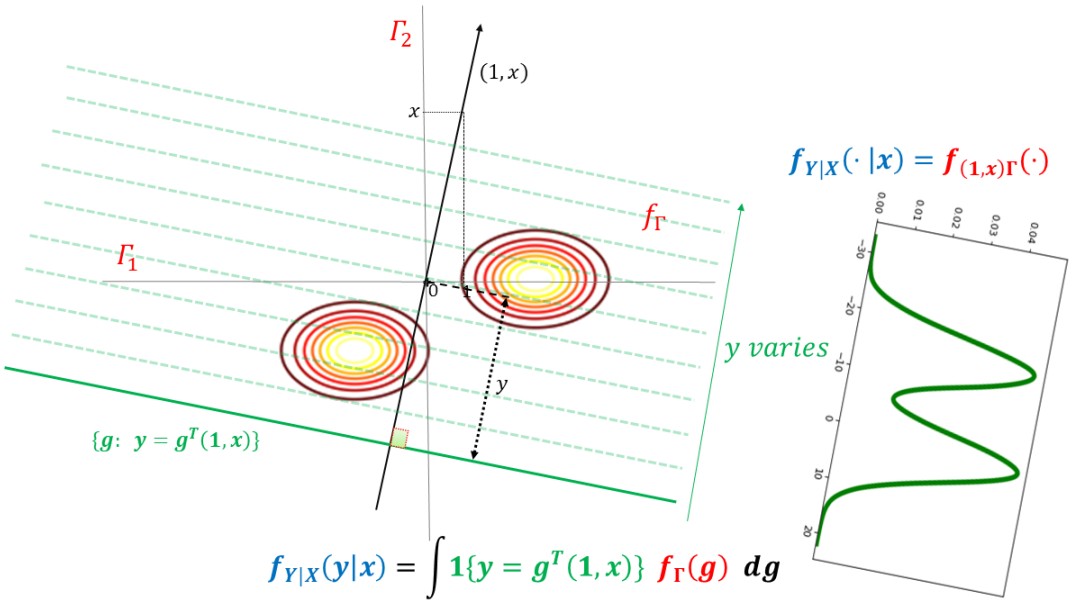
Our first characterization in Theorem 1.2 below is based on two intuitions: 1) recovering F_Γ, 2) using Bayes' Theorem, which expresses the PE directly as a function of this distribution, as in (10). In fact, as described in the Figure 1, the problem can be viewed as recovering the multivariate distribution F_Γ from its one-dimensional projections $F_{(1,x^\top)\Gamma}$, one for each point x of the support of X.[3]

Assumption 4 *The conditional density* $f_{Y|X}$ *exists and, for all* $l = 1, \ldots, p$ *and* x *in the support of* X*, its partial derivatives* $\partial_{x_l} f_{Y|X}(\cdot|x)$ *are integrable and square integrable on* $\mathbb{R}$*.*

I need the Assumption (4) only for the second characterization. In Section 4, I give sufficient conditions for Assumption 4 in terms of minimal smoothness of the density of Γ directly, rather than the one of (Y, X). Note that Assumption 4 holds for many classical parametric distributions of Γ. Let $\mathcal{S}_\Gamma$ denote a possible *a priori* on the support of Γ, *i.e.*, be such that $\mathrm{Supp}(\Gamma) \subseteq \mathcal{S}_\Gamma \subseteq \mathbb{R}^{p+1}$. I also denote by $\mathcal{I}(x,y) := \{g \in \mathcal{S}_\Gamma : \ y = (1, x^\top)g\}$.

Proposition 1 *In equation* (8) *together with Assumption* 2*, and*

[3] See also the operator formulation of this inverse problem using the Radon transform in Hoderlein et al. (2010), or using the partial Fourier transform in Gaillac and Gautier (2022).



Notes: Following the model $Y = \Gamma_1 + \Gamma_2 X = (1, X)\Gamma$, with $\Gamma \perp\!\!\!\perp X$. The red contour plot represents the *unobserved* density f_Γ. Let us fix a value of $X = x$. For a fixed value of y, the green line is the set of values that satisfy the model $\{g : y = g^\top(1, x)\}$. Making y vary, we thus identify one observed projection of this probability distribution f_Γ, as $f_{Y|X=x}(\cdot) = f_{(1,x)\Gamma}(\cdot)$, pictured in green on the right. Then, from observing different values of X yielding several observed one-dimensional projections of f_Γ, our aim with the first estimator is to recover the latter, then build a estimate of the functional $E(\Gamma|X, Y)$ of f_Γ using Bayes' theorem.

Figure 1: Illustration of the inverse problem with a bimodal density f_Γ

1. *under Assumption 3, the distribution F_Γ is identified, hence also the PE_k for $k = 1, \ldots, p+1$;*

2. *under assumption 3-(A), we have, for all $(x, y) \in \text{Supp}(X, Y)$,*

$$\mathbb{E}[\Gamma_k^d | (X, Y) = (x, y)] = \frac{\int_{\mathcal{I}(x,y)} g_k^d dF_\Gamma^*(g)}{\int_{\mathcal{I}(x,y)} dF_\Gamma^*(g)}, \ k = 1, \ldots, p+1, d \in \mathbb{N}, \tag{10}$$

where F_Γ^* is the unique solution of

$$\min_{F_\Gamma \in \mathcal{P}_d(\mathbb{R}^{p+1})} \int \left(\min_{F_{\widetilde{Y}_x, Y_x}:\ \widetilde{Y}_x \sim F_{(1,x)\Gamma},\ Y_x \sim F_{Y|X=x}} \mathbb{E}\left[(\widetilde{Y}_x - Y_x)^2 \right] \right) dF_X(x). \tag{11}$$

3. under assumptions 3-(B) and 4, we have, for all $(x, y) \in \text{Supp}(X, Y)$,

$$PE_{-1}(x, y) = \frac{-\partial_x F_{Y|X}(y|x)}{f_{Y|X}(y|x)}, \tag{12}$$

$$PE_1(x, y) = y - x^\top PE_{-1}(x, y). \tag{13}$$

Since PE_k is a linear functional of F_Γ in the linear RC model (8), Proposition 1.1 is a direct consequence of the results in Gaillac and Gautier (2021b).

Similar to Beran and Millar (1994); Arellano and Bonhomme (2023), Proposition 1.2 proposes a type of minimum distance formulation, using the Wasserstein distance. This simply consists in restating that for each value x of the support of X, the true distribution F_Γ minimizes the mean squared error between a variable distributed as the projection $F_{(1,x)\Gamma}$ and one distributed according to the observed $F_{Y|X=x}$. The new important point is that it can also be seen as a reformulation of the problem of finding the Generalized Wasserstein Barycenter introduced in Delon et al. (2022). This constructive reformulation opens the way to handle discrete support of the regressors in a nonparametric framework more generally for estimating the density f_Γ or handling varying coefficients as in Breunig (2021). This approach is closer in spirit to the Empirical Bayes modeling developed in Gu and Koenker (2017); Gilraine et al. (2020). Note that Proposition 1.2 holds even if we use a uniform measure on $\text{Supp}(X)$ instead of $\mathbb{P}_X$ in (11), and it seems to give empirically better results.

Proposition 1.3 is a constructive identification result, key to generalize the so-called Tweedie formula (see Robbins, 1956; Efron, 2011) in Section 3.2. Indeed, it allows one to estimate the individual effects directly, using features of the conditional distribution of the outcome Y on the regressors X. This simple closed-form expression allows nonparametric frequentist estimation of PE_k. This result is close to Hoderlein and Mammen (2007, 2009), where they consider more general nonseparable models than (8), but express the average effects as a function of the quantiles. Proposition 1.3 might also be deduced under different assumptions from Lemma 1 in Chernozhukov et al. (2015) and, under some conditions, holds for posterior marginal effects in more general models than the linear one. However, when particularized to the linear model, I provide a complete alternative proof based on Fourier analysis, allowing to compute other posterior moments as in Proposition 2, as well as the extensions to more elabo-

rate models developed in Section 3.2 or Appendix G.3 (the latter being also related to the multivariate outcome extension of Hoderlein and Mammen, 2007 in Kasy, 2022).

On the posterior variance. As discussed in Section 2, it may be also useful to recover higher moments of Γ. Specifically the posterior variance is a important feature to assess the information provided by the PE. If this is straightforward for the first characterization, as done in (10), this is less so for the second one. I can prove the following proposition.

Proposition 2 *Consider* (8)*, Assumptions 2-3-*(B)*, 4, and assume that the partial derivatives* $\partial_{x_k}\partial_{x_l} f_{Y|X}(\cdot|x)$ *are integrable and square integrable on* $\mathbb{R}$*. Then we have, for all* $(x, y) \in \text{Supp}(X, Y)$ *and* $k, l \in \{1, \ldots, p\}$*,*

$$\mathbb{E}[\Gamma_{k+1}\Gamma_{l+1}|(X, Y) = (x, y)] = \frac{\partial_{x_k}\partial_{x_l}\int_{-\infty}^{y} F_{Y|X}(v|x)dv}{f_{Y|X}(y|x)}. \tag{14}$$

3.2 Extension to a panel data model with individual effects

One important extension of the baseline model that I consider is

$$\widetilde{Y}_{i,t} = \Gamma_{1,i} + X_i^\top \Gamma_{-1,i} + W_{i,t}^\top \delta + \widetilde{\varepsilon}_{i,t}, \tag{15}$$

where $\overline{Y}_i := \Gamma_{1,i} + X_i^\top \Gamma_{-1,i}$ is the usual individual effect, X_i is a time-invariant covariate, Γ_i being an individual heterogeneity in the effect of X_i, while $\widetilde{\varepsilon}_{i,t}$ is an error term. I allow Γ_i to be correlated with time varying regressors $W_{i,t}$.

A standard approach is to start from a regression, removing out the effect of observed covariates $W_{i,t}^\top \delta$ (see, *e.g.*, Gilraine et al., 2020). Thus, I consider hereafter (15) with $\delta = 0$,

$$Y_i = (1, X_i^\top)\Gamma_i + \varepsilon_i, \tag{16}$$

under Assumption 2, where $Y_i = \sum_t \widetilde{Y}_{i,t}/n_i$ and $\varepsilon_i = \sum_t \widetilde{\varepsilon}_{i,t}/n_i$, n_i being the number of observations associated with individual i, considered as fixed. I consider the following assumption on the noise ε_i.

Assumption 5 *Assume that* ε_i *is independent of* (Γ_i, X_i) *and has a known distribution with density* f_ε *and characteristic function* φ_ε *which is nonvanishing on* $\mathbb{R}$*.*

Note that I maintain the Assumption 5 for simplicity. Assuming that f_ε is known can be relaxed using the Kotlarski lemma (see Kotlarski, 1967; Evdokimov and White, 2012; or Theorem 3 in Gaillac and Gautier (2021b) for weaker assumptions that do not require analyticity). The independence assumption could also be relaxed in this panel setting using Arellano and Bonhomme (2012). A last relaxation of Assumption 5 can be deduced from deconvolution results in Gaillac and Gautier (2021b), allowing φ_ε to have zeros on an open set at the cost of stronger assumptions on $\mathbb{P}_\Gamma$.

Motivated by the central limit theorem, the common assumption $\widetilde{\varepsilon}_{i,t} \sim_{iid} \mathcal{N}(0, \sigma_\varepsilon^2)$ yields that under Assumption 5, the noisy measure of the outcome is distributed as

$$Y_i \sim \mathcal{N}\left(\overline{Y}_i, \frac{\sigma_\varepsilon^2}{n_i}\right), \quad \overline{Y}_i = (1, X_i^\top)\Gamma_i, \tag{17}$$

where we are interested in decomposing the individual mean. I denote by $\mathcal{I}(x,y) = \{(g,e) \in \mathcal{S}_\Gamma \times \mathbb{R} : \ y = (1, x^\top)g + e\}$.

Theorem 1 *In model (15), under Assumption 5, and for all $(x,y) \in \text{Supp}(X,Y)$,*

1. *under Assumption 3, the distribution F_Γ is identified;*

2. *[Generalized Wasserstein barycenter formulation, GWB] under Assumption 3-(A), we have*

$$\mathbb{E}[\Gamma_k^d | (X,Y) = (x,y)] = \frac{\int_{\mathcal{I}(x,y)} g_k^d f_\varepsilon(e) dF_\Gamma^*(g) de}{\int_{\mathcal{I}(x,y)} f_\varepsilon(e) dF_\Gamma^*(g) de}, \ k = 1, \dots, p+1, \ d \in \mathbb{N} \tag{18}$$

where F_Γ^* is the unique solution of

$$\min_{F_\Gamma \in \mathcal{P}_d(\mathbb{R}^{p+1})} \int \left(\min_{F_{\widetilde{Y}_x, Y_x}:\ \widetilde{Y}_x \sim F_{(1,x)\Gamma},\ Y_x \sim h(\mathbb{P}_{Y|X=x})} \mathbb{E}\left[(\widetilde{Y}_x - Y_x)^2\right] \right) dF_X(x). \tag{19}$$

where

$$h(\mathbb{P}_{Y|X=x})(\cdot) := \mathcal{F}^{-1}\left[\frac{\varphi_{Y|X}(\star|x)}{\varphi_\varepsilon(\star)}\right](\cdot). \tag{20}$$

3. *[Generalized Tweedie formula, GT] under assumptions 3-(B), 4, and when $f_\varepsilon \in L^2(W)$, we have*

$$PE_{-1}(x,y) = \frac{-\partial_x F_{Y|X}(y|x)}{f_{Y|X}(y|x)},$$
$$PE_1(x,y) = y + \frac{x^\top \partial_x F_{Y|X}(y|x)}{f_{Y|X}(y|x)} + \frac{\mathcal{F}^{-1}\left[i\varphi_{Y|X}(\cdot, x)\varphi_\varepsilon' / \varphi_\varepsilon\right](y)}{f_{Y|X}(y|x)}.$$

Importantly, in the specific case where $F_\varepsilon \sim \mathcal{N}(0, \sigma_\varepsilon^2/n_i)$, this simplifies

$$PE_1(x, y) = y + \frac{x^\top \partial_x F_{Y|X}(y|x)}{f_{Y|X}(y|x)} + \frac{\sigma_\varepsilon^2}{n_i} \frac{\partial_y f_{Y|X}(y|x)}{f_{Y|X}(y|x)}. \tag{21}$$

Compared to Proposition 1.2, Theorem 1.2 contains an additional preliminary deconvolution step before considering the generalized Wasserstein barycenter.

Theorem 1.3 is a *Generalized Tweedie formula* to this context with covariates, and combines insights from the classical Tweedie formula, which I first present under an extended form in Appendix E for completeness, with those of Theorem 1. Equation (21) clearly shows that in this context what would correspond to the usual shrinkage in the parametric context is complemented by the predicted effect of X on the outcome.

3.3 Identification with additional variables or instruments

Let us focus on the baseline model of Section 3.1, as the relaxations of the baseline independence assumption developed below extend directly to the other contexts.

Using a varying coefficients approach. Without efficiently extending the tools developed in Breunig (2021) here, let us describe how available covariates can be combined with the previous approach to obtain a better description of the heterogeneity and to relax the independence assumption in one direction.

Suppose that additional covariates Z are available, where Z and X can have elements in common without X being a subset of Z. Consider the following model, which specifies the previous random coefficients as the sum of a nonlinear function $g(Z)$ and an unobserved random vector, denoted Γ for simplicity:

$$Y = (g_1(Z) + \Gamma_1) + (g_{-1}(Z) + \Gamma_{-1})^\top X, \tag{22}$$

$$\Gamma \perp\!\!\!\perp X, \quad \mathbb{E}(\Gamma|X, Z) = 0. \tag{23}$$

Under the conditional mean independence assumption (23), this implies that the function g is identified through the regression

$$\mathbb{E}(Y|X, Z) = g_1(Z) + g_{-1}^\top(Z)X. \tag{24}$$

Then, with the knowledge of g, we are back to the baseline model, using for the G-modeling strategy, that for $(x, y, z) \in \text{Supp}(X, Y, Z)$,

$$F_{Y|X,W}(y|x, z) = F_{(1,X)\Gamma}(y - (1, x)g(z)),$$

and for the F-modeling one, that for $t \in \mathbb{R}$ and $(x, z) \in \text{Supp}(X, Z)$,

$$\varphi_{\widetilde{Y}|X,Z}(t|x, z) = \mathcal{F}[f_\Gamma](t, tx), \quad \widetilde{Y} := Y - (1, X)g(Z).$$

I am not going to develop inference in this context, but one way is to use 1) sample-splitting for the estimation of g with well chosen machine learning estimators, then 2) my estimators of Section 4. This is implemented in my package `RegPE`.

Using conditioning. Additional variables Z of dimension p_Z can be used to relax the baseline independence assumption, performing the analysis conditional on Z.

Assumption 6 $\Gamma \perp\!\!\!\perp X|Z$.

Under Assumption 6, the parameter of interest becomes the expectation of Γ conditional on the observed quantities, *i.e.*, given values of the margins X, Y and the additional variables Z:

$$\text{PE}_k : (x, y, z) \mapsto \mathbb{E}\left[\Gamma_k|(X, Y, Z) = (x, y, z)\right], \quad k = 1, \ldots, p + 1. \tag{25}$$

Identification under Assumption 6 is the parallel of Proposition 1 and states that the same type of formula can be obtained for (25), simply conditioning on Z.

Proposition 3 *In equation (8) together with Assumption 6, and*

1. *under Assumption 3-(A) we have, for all $(x, y, z) \in \text{Supp}(X, Y, Z)$,*

$$\mathbb{E}[\Gamma_k^d|(X, Y, Z) = (x, y, z)] = \frac{\int_{\mathcal{I}(x,y)} g_k^d dF_\Gamma^{*,z}(g)}{\int_{\mathcal{I}(x,y)} dF_\Gamma^{*,z}(g)}, \ k = 1, \ldots, p + 1, \ d \in \mathbb{N} \tag{26}$$

where $F_\Gamma^{*,z}$ is the unique solution of

$$\min_{F_\Gamma^z \in \mathcal{P}_d(\mathbb{R}^{p+1})} \int \left(\min_{F_{\widetilde{Y}_{x,z}, Y_{x,z}}:\ \widetilde{Y}_{x,z} \sim F^z_{(1,x)\Gamma},\ Y_{x,z} \sim F_{Y|X=x,Z=z}} \mathbb{E}\left[(\widetilde{Y}_{x,z} - Y_{x,z})^2\right] \right) dF_{X,Z}(x, z).$$

2. *under Assumptions 3-(B) and 4, we have, for all $(x, y, z) \in \text{Supp}(X, Y, Z)$,*

$$PE_{-1}(x, y, z) = \frac{-\partial_x F_{Y|X,Z}(y|x, z)}{f_{Y|X,Z}(y|x, z)}, \tag{27}$$

$$PE_1(x, y, z) = y - x^\top PE_{-1}(x, y, z). \tag{28}$$

Using the control function approach. An alternative is to use the control function approach used in, *e.g.*, Florens et al. (2008); Imbens and Newey (2009); Masten and Torgovitsky (2016), when an instrument W is available.

Assumption 7 *1.* **(First stage equation)** *For each $k = 1, \dots, p$, there exists a scalar random variable V_k and a possibly unknown function h_k that is strictly increasing in its second argument, for which $X_k = h_k(W, V_k)$. The vector $V = (V_1, \dots, V_p)$ is continuously distributed.*

2. **(Instrument exogeneity)** $(\Gamma, V) \perp\!\!\!\perp W$.

Assumption 7 is another alternative to the independence Assumption 2. It restricts the dependence between X and Γ. Namely, it implies that most of the correlation between X and Γ occurs through V. This can be structurally motivated in some applications. Define $Z_k := F_{X_k|W}(X_k|W)$ for $k = 1, \dots, p$. Proposition 1 in Masten and Torgovitsky (2016) ensures that $(Z, \Gamma) \perp\!\!\!\perp W$ and that $X \perp\!\!\!\perp \Gamma | Z$, which gives identification in Proposition 3 under Assumption 7 rather than Assumption 2.

Proposition 4 (Identification using the control function) *Let the distribution of (Γ, X, Y, V, W) satisfy the assumptions 3-(B) and 7. The identified set of*

$$PE : (x, y, z) \mapsto \mathbb{E}\left[\Gamma | (X, Y, Z) = (x, y, z)\right]$$

is the same as in Proposition 3 conditioning on Z.

3.4 Assessing the sensitivity to the independence assumption

Finally, I provide tools for assessing sensitivity to the assumption 2. This section follows the findings from Masten and Poirier (2018); Masten et al. (2019). Let us define *conditional δ-dependence* (or *conditional partial independence*):

Definition 1 *Let δ be a nonnegative scalar. Say that Γ is conditional δ-dependent with X if*

$$\sup_{(x,y,z) \in \text{Supp}(X,Y,Z)} \sup_{g \in \mathcal{I}(x,y)} |f_{\Gamma|X,Z}(g|x,z) - f_{\Gamma|Z}(g|z)| \leq \delta$$

holds for all $z \in \text{Supp}(Z)$.

For $\delta = 0$, conditional δ dependence is equivalent to conditional Z independence and the independent joint distribution of (Γ, X) ensures point identification. For $\delta > 0$, I allow some deviations from the latter assumption, in a nonparametric neighborhood of this independent joint distribution of (Γ, X). Thus, I replace the conditional independence assumption by Assumption 8.

Assumption 8 *Let δ be a nonnegative scalar. Γ is conditional δ-dependent with X given Z.*

Under this assumption, the following theorem gives bounds on the PE, which can easily be computed.

Proposition 5 *Let the distribution of (Γ, X, Y, Z) satisfy (8) and make assumptions 3, 4 and 8. Then, for all $k = 1, \ldots, p+1$ and $(x, y, z) \in \text{Supp}(X, Y, Z)$,*

$$PE_k(x,y,z) \in \left[PE_k^*(x,y,z) - \frac{\delta \int_{\mathcal{I}(x,y)} |g_k| dg}{f_{Y|X,Z}(y|x,z)}, PE_k^*(x,y,z) + \frac{\delta \int_{\mathcal{I}(x,y)} |g_k| dg}{f_{Y|X,Z}(y|x,z)} \right],$$

where PE^* is the PE defined in (13)-(12) or (10) under Assumption 6.

4 Estimation of posterior effects

4.1 Using the generalized Tweedie formula (GT)

Asymptotic analysis with the minimax risk. This section characterizes the asymptotic properties of estimators of the PE in the minimax context, which I explain here. Based on a sample $(X_i, Y_i)_{i=1}^n$, let us define the expected error of an estimator $\widetilde{\text{PE}}_k$ of PE_k, for $k = 1, \ldots, p+1$,

$$\mathcal{R}\left(\widetilde{\text{PE}}_k, \text{PE}_k\right) := \mathbb{E}\left[\left\| \widetilde{\text{PE}}_k - \text{PE}_k \right\|_{L^2_\mu(\mathcal{S})} \right]$$

in $L^2_\mu(\mathcal{S})$, which is a L^2 norm on $\mathcal{S}$ possibly weighted by μ, $\mathcal{S}$ being a subset of $\mathbb{R}^{p+1}$ defined later in Assumption (Est.3).

First, for a specific estimator $\widetilde{\text{PE}}_k^{j_0}$, where j_0 is the tuning parameter, I show an upper bound on the maximum risk, which the worst error estimating PE associated

to a density f_Γ – assuming that it exists – in the space $\mathcal{H}^\sigma(l)$ defined later, for $k = 1, \ldots, p+1$,

$$\frac{1}{r(n)} \underbrace{\sup_{f_\Gamma \in \mathcal{H}^\sigma(l)} \mathcal{R}\left(\widetilde{\mathrm{PE}}_k^{j_0}, \mathrm{PE}_k\right)}_{\text{Maximum risk}} = O(1), \tag{29}$$

where $r(n)$ is thus a rate of convergence for this estimator. $\mathcal{H}^\sigma(l)$ characterizes the smoothness of the distributions f_Γ and is indexed by two parameters σ and l. Thus, controlling the maximum risk for an estimator shows the uniformity of its performance with respect to all distributions in the class $\mathcal{H}^\sigma(l)$.

Second, I turn to the question of the optimality of this estimator. The performance measure I consider is the minimax risk, *i.e.*, the minimum of the maximum risk that an estimator $\widetilde{\mathrm{PE}}_k$ can achieve,

$$\mathcal{R}_n^* := \inf_{\widetilde{\mathrm{PE}}_k} \sup_{f_\Gamma \in \mathcal{H}^\sigma(l)} \mathcal{R}\left(\widetilde{\mathrm{PE}}_k, \mathrm{PE}_k\right). \tag{30}$$

I show a lower bound $r(n)$ on the latter which takes the form, for all $k = 1, \ldots, p+1$,

$$\exists \nu > 0 : \ \underline{\lim}_{n\to\infty} \frac{1}{r(n)} \mathcal{R}_n^* \geq \nu. \tag{31}$$

Obviously, the goal is to get as sharp a lower bound as possible, and to get a rate for my estimator in (29) that is as close as possible to the rate achievable for this statistical problem in (31). Note that (29) also gives an upper bound on the minimax risk (30), since we are considering a specific estimator. Our estimator in this paper is based on Legendre polynomials, and Proposition 6 below shows that it achieves the best rate.

However, the tuning parameter j_0 must be chosen as a function of the smoothness parameter σ, which is unobserved. Therefore, the last step is to choose the tuning parameter $\widehat{j}_0$ using only the data, while keeping a rate close to the case where the smoothness parameter is known. In fact, I show that my estimator is adaptive, namely satisfies

$$\frac{1}{r(n)} \sup_{f_\Gamma \in \mathcal{H}^\sigma(l)} \mathcal{R}\left(\widetilde{\mathrm{PE}}_k^{\widehat{j}_0}, \mathrm{PE}_k\right) = O(1), \tag{32}$$

where the rate $r(n)$ is the one in (29) up to a logarithmic term. Table (1) below presents a summary of the rates obtained with my estimator in L^2_μ norm. Data-driven rule for selecting the tuning parameters is given in the Appendix B and the asymptotic normality results in Section C.

Smoothness and sampling assumptions.

Assumption 9 (Assumption on the supports) *Assume*

1. $\text{Supp}(X) := \prod_{l=1}^{p}[\widetilde{x}_l - x_0, \widetilde{x}_l + x_0] \subseteq \text{Supp}(X)$, *where* $\widetilde{x} \in \mathbb{R}^p$ *and* $x_0 > 0$;
2. $\text{Supp}(\Gamma) \subseteq \mathcal{S}_\Gamma := \prod_{l=1}^{p+1}[-g_0, g_0]$, *where* $g_0 > 0$.

I denote by $\omega := x_0 g_0 e/2$ *and assume that* $\omega e^{\omega(p-1)/e} < 1$.[4]

I maintain Assumption 9 for simplicity. Assumption 9-1 can be relaxed considering $\mathcal{S}_X := \prod_{l=1}^{p}[\widetilde{x}_l - x_0, \widetilde{x}_l + x_0] \subseteq \text{Supp}(X)$. In the following and similarly to what is done in Gaillac and Gautier (2022), this would imply conditioning all the estimated quantities by $X \in \mathcal{S}_X$, in particular using the truncated densities $f_{X|\mathcal{S}_X}$ and $f_{Y|X,\mathcal{S}_X}$. This would weaken Assumption 11-(Est.3) below. One can remove the condition $\omega e^{\omega(p-1)/e} < 1$ at the cost of a slightly sub optimal rate with the estimator I consider.

Assumption 10 (Smoothness assumption, Sobolev ellipsoid) *Let* $l \in (0, \infty)$, $\sigma > p/2$, *and assume that* f_Γ *exists and belongs to*

$$\mathcal{H}^\sigma(l) := \left\{ f_\Gamma : \quad \int_{\mathbb{R}^{p+1}} (1 \vee |\xi|_2)^{2\sigma} \left|\mathcal{F}[f_\Gamma](\xi)\right|^2 d\xi \leq l^2 \right\}.$$

The key proposition linking this Sobolev-type smoothness to the regularity of $\partial F_{Y|X}$ is Proposition 8, which is of independent interest. Note that, contrary to Assumption 3, the uniform distribution or truncated normal used by King (1997) does not satisfy Assumption 10. This is due to the discontinuity at the boundary of the support. Therefore, smooth approximations of the uniform distribution or the truncated normal at the boundary satisfy the Assumption 10. More importantly, the beta and Dirichlet distributions with parameter strictly greater than one, or the logit-normal distribution, which are common parametric distributions to represent probabilities hence used for ecological inference (see, *e.g.*, Katz and King, 1999; Imai et al., 2008), satisfy the Assumption 10.

The following assumptions are introduced to be able to derive convergence rates.

Assumption 11 *Assume that:*

[4] For $p > 1$, this means $\omega < \mathcal{W}((p-1)/e)e/(p-1)$, where $\mathcal{W}$ is the Lambert W function leading to a bound of 0.75, 0.62, 0.33 for $p = 2$, 3, and 10.

(*Est.1*) *we observe an i.i.d sample* $(X_i, Y_i)_{i=1}^n$*;*

(*Est.2*) *there exist densities* f_X *and* $f_{Y|X}$ *which are considered known for simplicity in the body of this paper and estimated under Assumption 12 in the Appendix;*

(*Est.3*) *For* $c_X, c_{X,Y} \in (0,\infty)$, $\|1/f_X\|_{L^\infty(\text{Supp}(X))} \leq c_X$, $\|f_X\|_{L^\infty(\text{Supp}(X))} \leq C_X$, *and there exists a bounded subset* $\mathcal{S} = \mathcal{S}_Y \times \text{Supp}(X)$ *of* $\text{Supp}(X,Y)$ *such that* $\|1/f_{Y|X}\|_{L^\infty(\mathcal{S})} \leq c_{X,Y}$.

I denote by $\mu = 1 \otimes_{l=1}^p \widetilde{\mu}_l^2$, where $\widetilde{\mu}_l(\cdot) = (1 - ((\cdot - \widetilde{x}_l)/x_0)^2)^{1/2}$. The use of weight μ means that we do not weight loss on the boundaries of $\text{Supp}(X)$ in the asymptotic analysis when using the $L^2_\mu(\mathcal{S})$ risk. I refer to Gaillac (2021) for a more complicated approach baesd on vaguelets-wavelets, but with uniform weight.

Table 1: Minimax $L^2_\mu(\mathcal{S})$ risk rates of convergence in $\mathcal{H}^\sigma(l)$, $\sigma = s + 1 - p/2$

	Lower bound, "best" est.	Est. (36), s known	Est. (36)-(96), data-driven
Rate, $r(n)$	$n^{-\frac{2s}{2s+p+2}}$	$n^{-\frac{2s}{2s+p+2}}$	$\left(\frac{n}{\ln(n)}\right)^{-\frac{2s}{2s+p+2}}$
This paper	(31), Proposition 7	(29), Proposition 6	(32), Proposition 10

Notes: The asymptotic is in n, "est." means estimator.

4.1.1 Series estimator.

The proof of Theorem 1 is constructive and my estimator $\widetilde{\text{PE}}_k^{j_0}$ is based on a plug-in approach of an estimator of $\left(\partial_{x_l} F_{Y|X}\right)_{l=1}^p$. Let $y \in \mathcal{S}_Y$. I focus here on the estimation of the derivatives $\partial_x F_{Y|X}(y|\cdot)$, which are key elements entering the PE formulation (12). I use a truncation of its decomposition on normalized Legendre polynomials $(L_k)_{k\in\mathbb{N}_0^p}$ in $L^2(\text{Supp}(X))$. The complete strategy implies also having first-step estimators of $f_{Y|X}$ and f_X, which I describe in the Appendix for simplicity of exposition.

Assuming that $F_{Y|X}(y|\cdot) \in L^2(\text{Supp}(X))$, we have the expansion

$$F_{Y|X}(y|\cdot) = \sum_{k\in\mathbb{N}_0^p} d_k(y) L_k(\cdot), \tag{33}$$

where $d_k(y) := \langle \mathbb{E}\left[\mathbb{1}\{Y \leq y\}|X = \cdot\right], L_k\rangle_{L^2(\text{Supp}(X))}$. When $F_{Y|X}(y|\cdot)$ also admits a square integrable derivative with respect to the $l \in \{1, \ldots, p\}$ variable such that $\widetilde{\mu}(\cdot)\partial_l F_{Y|X}(y|\cdot) \in L^2(\text{Supp}(X))$, a valid decomposition of $\partial_l F_{Y|X}(y|\cdot)$ in the space $L^2_{\widetilde{\mu}^2}(\text{Supp}(X))$ is simply[5]

$$\partial_l F_{Y|X}(y|\cdot) = \sum_{k \in \mathbb{N}_0^p} d_k(y)\partial_l L_k(\cdot). \tag{34}$$

Let $j_0 \geq 0$ be a parameter chosen *a posteriori* as a function of the sample size n. To deal with the approximation and statistical problems, I use

$$\widetilde{\partial_l F_{Y|X}}^{j_0}(\star|\cdot) := \sum_{|k|_\infty \leq j_0} \widetilde{d}_k(\star)\partial_l L_k(\cdot), \tag{35}$$

where, for all $y \in \mathcal{S}_Y$,

$$\widetilde{d}_k(y) := \frac{1}{n}\sum_{i=1}^{n} \frac{\mathbb{1}\{Y_i \leq y\}}{f_X(X_i)} L_k\left(X_i\right), \tag{36}$$

and replace $\partial_l F_{Y|X}$ by $\widetilde{\partial_l F_{Y|X}}^{j_0}$ in (13)-(12) to obtain the estimator $\widetilde{\text{PE}}_k^{j_0,GT}$. Note that there is no regularization with respect to the first variable Y. An intuitive explanation is that the estimation of the unconditional cdf can be done at parametric rate (see, *e.g.*, Brunel et al., 2010, for more details).

4.1.2 Upper and lower bounds

Proposition 6 (L^2_μ convergence rate) *Let $\sigma = s + 1 - p/2$, $s > p - 1/2$, and $j_0 = \lfloor \widetilde{j} \rfloor$, $\widetilde{j} = n^{1/(2s+p+2)}$. Make assumptions 2, 3-(B), 9 and 11, then (29) holds with $r(n) = n^{-s/(2s+p+2)}$ for the estimator $\widetilde{PE}_k^{j_0,GT}$.*

[5] This holds because the functions $\Omega_{k,l}(\cdot) = \partial_l L_k(\cdot)\widetilde{\mu}(\cdot)/\sqrt{k_l(k_l+1)}$ are tensor products of associated Legendre functions and Legendre polynomials. $(\Omega_{k,l})_{k\in\mathbb{N}_0^p}$ constitute also an orthonormal basis of $L^2(\text{Supp}(X))$ using, *e.g.*, 14.17.6 in Olver et al. (2010) and as they are solutions of the Sturm-Liouville equation 14.2.2 in Olver et al. (2010). Note that we have

$$\widetilde{\mu}_l(\cdot)\partial_l F_{Y|X}(y|\cdot) = \sum_{k \in \mathbb{N}_0^p} d_k(y)\sqrt{k_l(k_l+1)}\Omega_{k,l}(\cdot),$$

hence the link with the vaguelet-wavelet formulation of this inverse problem in Cai (2002) and that I use in Appendix. The vaguelet-wavelet formulation is more complex but allows to handle more general geometry of $\text{Supp}(X)$ and without the weight μ. This approach is similar in spirit to the vaguelet-wavelet decomposition (see, *e.g.*, Section 2.2 in Cai, 2002).

Proposition 6 shows that my main estimator based on Legendre polynomials admits a polynomial-weighted L^2 convergence rate.

Proposition 7 (Minimax lower bounds) *Make assumption 2. Let* $\sigma = s+1-p/2$ *and for* $0 < l < \infty$, *assume* $s \geq p - 1/2$, $\|f_X\|_{L^\infty(\text{Supp}(X))} \leq C_X < \infty$. *Then (31) holds with* $r(n) = n^{-s/(2s+p+2)}$.

To comment on Propositions 7 and 6, let us give more background and compare two related inverse problems where regressors have limited variation, in the case $p = 1$:

1. estimation of the density f_Γ,
2. estimation of the PE, which are functionals of f_Γ.

Estimating the density f_Γ when the regressors have compact support is an inverse problem treated in Gaillac and Gautier (2022). There, we decompose the problem using the truncated Fourier operator $\mathcal{F}_c : L^2\left(W_{[-1,1]}\right) \to L^2\left([-1,1]\right)$, where $W_{[-1,1]} = \mathbb{1}\{[-1,1]\} + \infty\ \mathbb{1}\{[-1,1]^c\}$ and $L^2\left(W_{[-1,1]}\right) = \left\{f \in L^2\left(\mathbb{R}^d\right) :\ \text{Supp}(f) \subseteq [-1,1]\right\}$, $\mathcal{F}_c[f] = \mathcal{F}\left[f\right](c\,\cdot)$ and show that for all $t \in \mathbb{R}$, in $L^2\left([-1,1]\right)$,

$$\mathcal{F}_{tx_0}\left[\mathcal{F}_{1\text{st}}\left[f_\Gamma\right](t,\cdot_2)\right](\star) = \mathbb{E}\left[e^{itY}|X = x_0\star\right],$$

where $\mathcal{F}_{1\text{st}}$ is the Fourier transform with respect to the first variable. We show that the operator $\mathcal{F}_c$ admits a singular value decomposition, and that the singular values decay sub-exponentially with k as $e^{-2k\ln(7e\pi(k+1)/c)}$ (see, *e.g.*, Lemma B.5. in Gaillac and Gautier, 2022). This is a severely ill-posed problem and lower bounds for the L^2 risk in Theorem 1 in Gaillac and Gautier (2022) give logarithmic rates of convergence $(\ln(n)/\ln_2(n))^{-\sigma}$, where σ is a Sobolev-type regularity of the same type as the Assumption 10 (see, *e.g.*, Appendix B.5. in Gaillac and Gautier, 2022). A plug-in approach of this density to estimate the PE leads to slower convergence rates than my direct approach in this paper.

Estimating the posterior effects PE is a simpler problem and thus achieves faster rates. Minimax convergence rates for the L^2 risk in nonparametric estimation of the k-th derivative of a regression function with p dimensional covariates, assuming it belongs to a classical Sobolev space indexed by s, are $(n/\ln(n))^{-s/(2s+d+2k)}$ (see, *e. g.*, Theorem 6.3.7 in Giné and Nickl, 2016). The difficulty of the problem amounts to

estimating a first derivative $\partial_l F_{Y|X}$, hence $k = 1$ in Table 1. Since Γ *is not observed*, Proposition 8 importantly relates the regularity of $\partial_l F_{Y|X}$ (indexed by s) to that of f_Γ (indexed by σ).

4.1.3 Practical implementation in the panel context

I focus here on inference in the context of Section 3.2 under the normality assumption of the error term, which yields (21). In addition to the estimation procedure developed in the previous sections, we first need a preliminary nonparametric estimator $\widehat{\sigma}^2_\varepsilon$ of σ^2_ε, which, similarly to Gilraine et al. (2020), is taken as

$$\widehat{\sigma}^2_\varepsilon = \frac{\sum_j \sum_i (\widetilde{Y}_{i,j} - \widetilde{Y}_j)}{\sum_j (n_j - 1)}. \tag{37}$$

Then, we also need an estimator of $\partial_y f_{Y|X}(y|x)$ for $(x, y) \in \text{Supp}(X, Y)$. I use that $\partial_y f_{Y|X}(y|x) = \partial_y f_{Y,X}(y, x)/f_X(x)$ and one approach is to use the same preliminary estimator as when estimating $f_{Y|X}(y|x)$, based on series estimators using Legendre polynomials and their first derivative (see Section A.2.1 and Ullah and Pagan, 1999; Giné and Nickl, 2016 for many other examples of such estimators).

4.2 Using the optimal transport based estimator (GWB)

In this section, I assume that X has discrete support $\text{Supp}(X) := \{x_j\}_{j=1}^\kappa$ and that assumptions for identification are satisfied. I denote by $\overline{p}_j := \mathbb{P}(X = x_j)$ and the p-simplex by Σ_p. Assume we have $n = \sum_{j=1}^\kappa n_j$ i.i.d. observations $(Y_{i,j})_{i=1,\dots,n_j;j=1,\dots,\kappa}$ from the marginals $\{F_{Y|X=x_j}\}_{j=1,\dots,\kappa}$, used in empirical estimators of the conditional distributions of $Y|X = x_j$, namely $\widehat{F}_{Y|X=x_j}$.

Denote by $A = \sum_{j=1}^\kappa P^\top_{(1,x_j)} P_{(1,x_j)}$, where $P_{(1,x_j)} \in \mathcal{M}_{p+1,1}(\mathbb{R})$ is a projection matrix onto $\text{span}(1, x_j)$. As the support points are distinct, the matrix A is invertible. In this section, I maintain Assumption 3-(A), so that there indeed exists a unique solution to the Wasserstein barycenter problem if at least one the marginals is absolutely continuous (see Proposition 6 in Le Gouic and Loubes, 2017).

4.2.1 Estimator in the cross-section linear RC model

Introducing, for $k \in \{1, \dots, p+1\}$ and $(x, y) \in \text{Supp}(X, Y)$, the function $\underline{m}_{k,x,y}$ which to $(p, G) \in \Sigma_p \times \mathcal{P}_{a.c.}(\mathcal{S})^\kappa$ associates the solution of

$$\min_{F_\Gamma \in \mathcal{P}_d(\mathbb{R}^{p+1})} \sum_{j=1}^{\kappa} p_j \left(\min_{F_{\widetilde{Y}_x, Y_x}:\ \widetilde{Y}_x \sim F_{(1,x)\Gamma},\ Y_x \sim G_j} \mathbb{E}\left[(\widetilde{Y}_x - Y_x)^2 \right] \right),$$

we can write the posterior as

$$\text{PE}_k(x, y) = \underline{m}_{k,x,y}(\underline{p}, F_{Y|X=x_1}, \dots, F_{Y|X=x_\kappa}).$$

Then, let us introduce an estimator $\widehat{\text{PE}}_k(x, y)$ of $\text{PE}_k(x, y)$ based on the plug-in

$$\widehat{\text{PE}}_k^{GWB}(x, y) = \underline{m}_{k,x,y}\left(\widehat{p}, \widehat{F}_{Y|X=x_1,n_1}, \dots, \widehat{F}_{Y|X=x_\kappa,n_\kappa} \right). \tag{38}$$

4.2.2 Consistency

I show a consistency results for $\widehat{\text{PE}}_k^{GWB}$.

Theorem 2 (Consistency of $\widehat{\text{PE}}^{GWB}$) *Make Assumption 3-(A). Then, for $k = 1, \dots, p+1$ and $(x, y) \in \text{Supp}(X, Y)$, we have*

$$\widehat{PE}_k^{GWB}(x, y) \xrightarrow{\mathbb{P}} PE_k(x, y)$$

as n goes to infinity.

In order to prove consistency, I actually decompose the function $\underline{m}_{k,x,y}$ in the proof, and rewrite the problem as a classical Wasserstein barycenter problem, using Proposition 3.1 in Delon et al. (2022). Theorem 2 builds on results from Le Gouic and Loubes (2017) that ensure the continuity of the Wasserstein barycenter map and the consistency of the empirical conditional cdf $\widehat{F}_{Y|X=x}$. This result is similar in spirit to the consistency results of Arellano and Bonhomme (2023) for deconvolution models, but focuses on the consistency of the conditional expectations.

Unfortunately, in general the estimator $\widehat{\text{PE}}_k^{GWB}$ introduced above is not a smooth function of the conditional distributions $Y|X_j$ (see, *e.g.*, Agueh and Carlier, 2017). Similar to what is done in the literature related to the classical OT problem (see, *e.g.*, Goldfeld et al., 2022), one solution is to regularize the problem to obtain confidence bounds for the predictions, at the cost of having to deal with a bias term. As this would introduce additional complications, I leave this for further research.

4.2.3 Practical implementation

I implement this estimator adapting the algorithm of Delon et al. (2022), leveraging the free support approach of Cuturi and Doucet (2014). The output of this algorithm is a discrete uniform distribution with a prespecified N_g number of points of support that I denote $\widehat{g}$, which can be considered as an approximation parameter. In order to compute the estimator $\widehat{\mathrm{PE}}^{GWB}$ in (10), (18), and (26), we then have to integrate this distribution and a type of smoothing of the indicator function $\mathbb{1}\{g \in \mathcal{I}(x,y)\}$ or $\mathbb{1}\{(g,e) \in \mathcal{I}(x,y)\}$ is thus needed. More precisely, in the case (10), I use

$$\widehat{\mathrm{PE}}_k^{GWB}(x,y) = \frac{\sum_{l=1}^{N_g} \widehat{g}_{k,l}\phi\left(\frac{y-(\widehat{g}_{1,l}+\widehat{g}_{-1,l}^{\top}x)}{h}\right)}{\sum_{l=1}^{N_g}\phi\left(\frac{y-(\widehat{g}_{1,l}+\widehat{g}_{-1,l}^{\top}x)}{h}\right)}, \tag{39}$$

where $h := 10/N_g$ is a smoothing parameter, ϕ is the standard normal density kernel, and proceed similarly for (18) and (26).

4.2.4 Estimator in the panel data model with individual effects

In the case of Section 3.2, we first need a nonparametric estimator of f_ε. Under the assumption of normality, $F_\varepsilon \sim \mathcal{N}(0, \sigma_\varepsilon^2/n_i)$, I use (37). The second step is to estimate the marginals $h(\mathbb{P}_{Y|X=x})$ in (20). This requires a Gaussian error deconvolution step. This problem of recovering a density when it is measured with additive noise of known density is a classical problem in the statistical literature (see, *e.g.*, Carroll and Hall, 1988; Delaigle and Meister, 2008; Comte and Lacour, 2013; Giné and Nickl, 2016, and references therein). I opt for the kernel-type density estimator of Delaigle and Meister (2008), which allows for heteroscedastic errors, which is empirically relevant in our case. It uses the bootstrap bandwidth selector without resampling as implemented in the R package `decon` by Wang and Wang (2011). I then sample N observations with estimated measures $h(\mathbb{P}_{Y|X=x})$, where $N = 300$ is an approximation parameter. The results do not appear to be sensitive to taking larger values of N.

4.2.5 Case where X is continuous

If X is continuous, one strategy is to use optimal quantization (see, *e.g.*, Graf and Luschgy, 2007; Pagès, 2015; Mérigot et al., 2021), which is the problem of finding a discrete distribution that is as close as possible to the target distribution with

respect to the 2-Wasserstein distance. Thus, we search for a point cloud $X^d = (x_1^d, \ldots, x_K^d) \in (\mathbb{R}^p)^K$ such that the uniform measure with support $\text{Supp}(X^d)$, denoted by δ_{X^d}, minimizes the 2-Wasserstein distance between δ_{X^d} and F_X. If this problem is non-convex, in practice and for well chosen initial supports, simple algorithms exist for solving it. There are also guarantees for the quality of the approximation in terms of Wasserstein distance, which decays at a rate of $K^{-1/p}$.

In practice, I use the stochastic gradient algorithm called Competitive Learning Vector Quantization, implemented in the R package `QuantifQuantile` by Charlier et al. (2015). An alternative strategy when $p = 1$ is to first discretize X using a grid of K empirical quantiles $X_{\lfloor l/n \rfloor,1}$ for $l = 0, n/K, \ldots, n$, where $K \to \infty$ as $n \to \infty$. I take the rule $K = \max(3, \lfloor 1.5(n/p)^{0.25} \rfloor)$. I provide robustness checks for this discretization in the relevant cases and leave a theoretical discussion of this choice and its implications for further research.

4.3 Monte-Carlo simulations

I provide several validations of my methods in finite samples. This section presents Monte-Carlo simulations with the baseline independence assumption (Section 4.3.1) and with conditional independence with a discrete covariate Z (Section 4.3.2). The appendix collects additional Monte Carlo simulations when Z is continuous in the latter case (Section F.1), or in the panel data model (Section F.2). Finally, Section G.7 considers an application to ecological inference where the *true value* of the parameters is known using specific register data.

An alternative to the series estimator theoretically analyzed in Section 4.1.1 is to use a kernel-based estimator for both $\partial_l F_{Y|X}$ and $f_{Y|X}$, then similarly plug in (13)-(12). (see Chapter 4 in Ullah and Pagan, 1999). I consider the estimators of Hall et al. (2004), where the bands are selected based on cross-validation. Both the Legendre and kernel-based GT estimators are implemented in my R package `RegPE`, using the package `np` for the latter, but the kernel-based estimator seems to be much more stable in practice to the different distributions of the regressors X. The results below for the GT estimator are thus based on this kernel-based implementation.

4.3.1 With independence

Consider the baseline model of Section 3.1, when $p = 1$ and in two setups where the independence assumption 2 holds. I want to compare to Bayesian estimators now standard in the ecological inference literature, see Section G, hence the choice of a setup where a direct comparison is possible, namely $\Gamma = (\Gamma_2^*, \Gamma_1^* - \Gamma_2^*)$ where Γ^* is compactly supported in $[0,1]^2$. More precisely, I take:

1. Γ^* is distributed according to $C(F_{\Gamma_1^*}, F_{\Gamma_2^*})$ where C is a Gaussian copula of parameter Σ and the marginals $F_{\Gamma_1^*}$, $F_{\Gamma_2^*}$ follow a Beta$(4, 1.5)$ distribution. I take $\Sigma_{11} = 0.2$, $\Sigma_{2,2} = 0.1$, and $\Sigma_{2,1} = 0.1$. $X \sim U(0,1)$ is uniformly distributed.

2. Γ^* is a logit mixture of normal distributions with mixing probability $(0.6, 0.4)$. The first distribution is normal with mean $(-0.4, 1.4)$, variance $(0.2, 0.1)$, and covariance 0. The second is normal with mean $(-0.4, -1.4)$ and same covariance matrix.[6] I take X following a truncated normal to $[0,1]$ with untruncated mean 0.6 and variance 0.05.

I compare the estimators with the true value of $\mathrm{PE}(X_i, Y_i)$ rather than the value of Γ_i. Indeed, the other part of the error is not varying with the type of estimator of the PE, only reflecting the information contained in the PE. Table 2 shows the results. It compares my two estimators with the Bayesian parametric method of King (1997) which has a multivariate truncated normal prior, implemented in the `R` package `ei`. Using the Bayesian parametric method with a different prior developed in Imai et al. (2011) and implemented in the `R` package `eco` gives very similar results as King (1997) (see also Appendix G.7 for a comparison with the hierarchical Dirichlet model of Rosen et al., 2001, when $p = 2$).[7]

The results are presented in Table 2. A first point is that the error of the Bayesian method in these two particular contexts does not really shrink with the sample size, probably due to the misspecification. On the contrary, the errors for my two non-parametric methods are well reduced when the sample size goes from 1000 to 5000 (*e.g.* from 0.06 to 0.04 (respectively 0.045 to 0.026) for the l^1 error on Γ_1 of the GT

[6] This is similar to Simulation II in Imai et al. (2008).

[7] The computational cost of the nonparametric Bayesian method of Imai et al. (2011) with the `R` package `eco` for these simulations with sample sizes 1000 and 5000 is prohibitive.

methods in case 1). However, it is interesting to note that for the smallest sample sizes (1000), the Bayesian method performs well: it is the best in case 1 with a uniform regressor, and better than the GT method in Case 2, despite being dominated by the GWB estimator.

Importantly, my methods seem to be more robust to the type of regressor, as they both perform well in the two different contexts. A final point to emphasize is that even in this context, with a continuous regressor X which discretized to fit its theoretical setting, the GWB method performs dramatically well in the two scenarios: it is the best performing method for all sample sizes in Case 2, and very close to being the best for all sample sizes in Case 1.

Table 2: In-sample errors with independence

Case 1 (Γ Beta distribution, X Uniform)

	l^1 error				l^2 error			
	Γ_1		Γ_2		Γ_1		Γ_2	
Sample size	1000	5000	1000	5000	1000	5000	1000	5000
Bayesian parametric	**0.038**	0.038	**0.083**	0.084	**0.058**	0.059	**0.116**	0.118
GT	0.04	0.033	0.096	0.079	0.061	**0.052**	0.136	0.114
GWB (disc. X)	0.043	**0.036**	0.084	**0.073**	0.068	0.056	0.119	**0.105**

Case 2 (Γ logit-mixture of normals, X truncated normal)

	l^1 error				l^2 error			
	Γ_1		Γ_2		Γ_1		Γ_2	
Sample size	1000	5000	1000	5000	1000	5000	1000	5000
Bayesian parametric	0.051	0.05	0.091	0.089	0.064	0.063	0.117	0.116
GT	0.06	0.041	0.106	0.071	0.081	0.058	0.145	0.106
GWB (disc. X)	**0.045**	**0.026**	**0.075**	**0.046**	**0.062**	**0.038**	**0.098**	**0.064**

Notes: in this 2 dimensional case, the in-sampled l^1 error is computed as $\sum_{i=1}^{n} |\widehat{\text{PE}}_k(X_i, Y_i) - \text{PE}(X_i, Y_i)|/n$ and the l^2 error as $(\sum_{i=1}^{n}(\widehat{\text{PE}}_k(X_i, Y_i) - \text{PE}(X_i, Y_i))^2/n)^{1/2}$, where $\widehat{\text{PE}}_k(X_i, Y_i)$ are the different estimators."GWB (disc. X)" refers to the GWB estimator where the distribution of X has been discretized using the rule of Section 4.2.5. "Bayesian parametric" refers to King (1997) method with bivariate truncated normal prior, implemented in the R package ei. The Monte-Carlo experiment uses 250 simulations.

4.3.2 With conditional independence

I consider a DGP that allows us to demonstrate the use of the various ways to relax the independence assumption with my estimators. Consider a DGP where Γ and X are related by an additional variable Z. Here I take Z as the discretized version with 3 points of support of a variable Z^* that is Beta$(2, 1.3)$ distributed (cutoffs at 0.3 and 0.8). Then I consider ϵ distributed as Γ^* in Case 1 of Section 4.3 and

$$\Gamma^* = \begin{pmatrix} 0.2Z^* \\ 0.1Z^* \end{pmatrix} + \begin{pmatrix} 0.6\epsilon_1 \\ 0.7\epsilon_2 \end{pmatrix}$$
$$X^* = 0.2(Z^*)^2 + 0.8\eta, \ \ \eta \sim \text{Beta}(4, 2).$$

Table 3: In-sample errors with conditional independence

	l^1 error				l^2 error			
	Γ_1		Γ_2		Γ_1		Γ_2	
	1000	5000	1000	5000	1000	5000	1000	5000
				Without Z				
Bayesian parametric	0.09	0.09	0.143	0.142	0.103	0.104	0.16	0.16
GT	0.099	0.094	0.149	0.138	0.137	0.133	0.197	0.183
GWB (disc. (X))	**0.085**	**0.078**	**0.087**	**0.087**	**0.102**	**0.094**	**0.132**	**0.132**
				With Z				
Bayesian parametric	**0.047**	0.049	**0.073**	0.076	**0.062**	0.065	**0.095**	0.099
GT varying	0.062	0.042	0.098	0.066	0.099	0.07	0.156	0.106
GT	0.048	**0.036**	0.075	**0.055**	0.07	**0.056**	0.105	**0.082**
GWB (disc. (X, Z))	0.075	0.059	0.148	0.124	0.104	0.081	0.208	0.173

Notes: in this 2 dimensional case, the in-sampled l^1 error is computed as $\sum_{i=1}^n |\widehat{\text{PE}}_k(X_i, Y_i, Z_i) - \text{PE}(X_i, Y_i, Z_i)|/n$ and the l^2 error as $(\sum_{i=1}^n (\widehat{\text{PE}}_k(X_i, Y_i, Z_i) - \text{PE}(X_i, Y_i, Z_i))^2/n)^{1/2}$, where $\widehat{\text{PE}}_k(X_i, Y_i, Z_i)$ are the different estimators. See the Appendix for non-sampled results and comparison to the true value of Γ. "Bayesian parametric" refers to King (1997) method with bivariate truncated normal prior, implemented in the R package ei. "GWB (disc. (X, Z))" refers to the GWB estimator where the distribution of (X, Z) has been discretized using the rule of Section 4.2.5. "GT varying" corresponds to the varying coefficients approach described in (23). The Monte-Carlo experiment uses 250 simulations.

The results are shown in Table 3. A first point is that *without* using the variable Z, the GWB method again performs better than the Bayesian method of King (1997) and my

GT method for all sample sizes. The latter two remain close. Due to misspecification, errors for all methods without Z do not really shrink with the sample size.

Then, I compare different estimators with this additional variable Z: 1) the same Bayesian method of where Z can be introduced, 2) the GT method with varying coefficients ("GT varying") corresponding to (23), 3) the GT method where Z enters fully nonparametrically as in (27)-(28), and finally 4) the GWB method where both (X, Z) are discretized. Again, an important point is that although the parametric Bayesian method actually performs better for a sample size of 1000 when Z is used, its errors remain nearly constant. On the contrary, the errors shrink for all my methods. Specifically, the GT method without constraint on Z performs best at sample size 5000 and is really close to the Bayesian method for $n = 1000$ (0.048 and 0.075 (0.047 and 0.073, respectively) for the l^1 norm of Γ_1 and Γ_2). Finally, if including Z in the GWB method helps reducing the errors, it does not compete well with the others in this setting, probably due to the discretization. One might prefer to use a GWB method with varying random coefficients approach.

5 Individual level effect of teachers' knowledge on their performance

I apply my method to predict how each teacher's value added is affected by his or her knowledge, extending the work of Bau and Das (2020). Similarly, I focus not only on estimating the TVA using data from Pakistan, but also on explaining its variation with respect to observed teacher characteristics. The innovation is that my method allows to describe the heterogeneity of this variation and to use it for policy design. Our estimates are based on the same data collected between 2003 and 2007 from 112 villages in Punjab province, Pakistan, as part of the Learning and Educational Achievement in Pakistan Schools (LEAPS) project.

5.1 Context and OLS/IV estimations

Importantly, these data include test scores for matched student-teacher pairs as well as a rich set of teacher characteristics that can explain the TVA. Test scores are estimated with item response theory (IRT, see, *e.g.*, Das and Zajonc, 2010) and

measured in standard deviations. Bau and Das (2020) perform this analysis by first estimating the TVA using a teacher-year fixed effects model of student test scores. Then, in a second step, they regress this estimated TVA on several characteristics listed below.

There are a few peculiarities that we need to take into account when replicating the analysis of Bau and Das (2020). The first is that estimating teacher effects with observational data requires controlling for sorting between students and teachers. They use lagged test scores, which may affect students differently in different grades, as well as year-specific and grade-specific shocks as controls.[8] Therefore, I first consider a similar model for estimating the TVA (40) and then use linear regression to explain it using the characteristics (41),

$$\overline{Y}_{j,i,g,t} = \delta_0 + \sum_a \delta_a \overline{Y}_{j,t-1} \mathbb{1}\{\text{grade} = a\} + \mu_g + \widetilde{Y}_{j,i,t} + \widetilde{\varepsilon}_{j,i,t} \tag{40}$$

$$\widetilde{Y}_{j,i,t} = \nu_{1,i} + \gamma_1 X_{1,i} + X_{-1,i}^\top \gamma_{-1}, \tag{41}$$

where $\overline{Y}_{j,t-1}$ are past students' tests scores, μ_g are the grades fixed effects, $X_i = (X_{1,i}, X_{-1,i})$ contains mean teacher's knowledge $X_{1,i}$ as well as $X_{-1,i}$ which includes district fixed effects, gender, being a local, whether teachers received some training, have at least a bachelor's degree, more than 3 years of experience, whether the school is public or private, and having a temporary contract, and $\nu_{1,i}$ is teacher-specific error term in the value added, containing the unobserved effects and independent of the noise $\widetilde{\varepsilon}_{j,i,t}$. Similar to Bau and Das (2020), I also consider an IV strategy and instrument for the teacher's mean score in the first year tested, $X_{1,i}$, with the mean score of the second year, denoted by W_i.

The results are presented in Table 5. The first important finding is that higher teacher knowledge of the program, as measured by the same average test scores on the same tests as students, is significantly associated with higher TVA. These effects are similar in magnitude to those estimated in other developing countries (see Bau and Das, 2020, for more details). The results in column (3) indicate that a 1 SD increase in teacher knowledge increases TVA by 0.24 SD on average. The second result is that these

[8] Bau and Das (2020) performs several checks that I do not replicate here as they are not the focus of this paper. These suggest that there is little systematic sorting here. Similar to Chetty et al. (2014a), one of them is the use of students who change schools.

observed teacher characteristics explain only a small percentage of the variation in TVA, highlighting the importance of modeling the unobserved heterogeneity.

5.2 Estimation of the individual-level causal effect of knowledge on performance

Let's take the analysis a step further and consider the heterogeneity of this effect of knowledge on performance. Thus, instead of the main equation (41), I consider

$$\widetilde{Y}_{j,i,t} = \Gamma_{1,i} + \Gamma_{2,i} X_{1,i} + X_{-1,i}^{\top}\gamma + \widetilde{\varepsilon}_{j,i,t}, \tag{42}$$

where $\Gamma_{2,i}$ is the individual causal effect of knowledge on TVA, $\widetilde{Y}_{j,i,t}$, while $\Gamma_{1,i}$ captures the unobserved effects not explained by X_i. We are only interested in the homogeneous effect with respect to $X_{-1,i}$, so γ is kept as a deterministic vector.[9] Tests of the linearity of (42) with respect to teachers' knowledge do not reject this assumption.[10] I keep the normality assumption of $\widetilde{\varepsilon}_{j,i,t}$ and denote by by $h_{it} = n_{i,t}/\sigma_{\varepsilon}^2$, $\widetilde{Y}_{i,t} = \sum_{j=1}^{n_{j,t}} \widetilde{Y}_{j,i,t}/n_{i,t}$, and $Y_i = \sum_t h_{i,t}\widetilde{Y}_{i,t}/\sum_t h_{i,t}$, which yields the analog of (17) in this context:

$$Y_i \sim \mathcal{N}\left(\Gamma_{1,i} + \Gamma_{2,i} X_{1,i} + X_{-1,i}^{\top}\gamma, \frac{\sigma_{\varepsilon}^2}{\sum_t n_{i,t}}\right).$$

To handle the potential endogeneity of $X_{1,i}$ in a simple way, I consider an additional first-stage homogeneous equation $X_{1,i} = a_0 + a_1 W_i + \eta_i$, where η_i is mean independent of W_i.[11] I present results for this IV specification below. Finally, I assume either:

A1. Independence $\Gamma_i \perp\!\!\!\perp W_i$,

A2. Independence in a varying coefficients approach (22)-(23) with a linear specification for $g(Z)$, where Z_i is either i) some teacher's training, or ii) some teacher's training, public school, and experience.

[9] See Breunig and Hoderlein (2018) for a test of wheter a coefficient is fixed or random in the context without noise.

[10] Specifically, the Ramsey RESET test (see, *e.g.*, Wooldridge, 2010) does not reject the OLS model (p-values 0.54 and 0.72 with 4th and 3rd order polynomials, respectively).

[11] Being less parsimonious and considering a full triangular model with random coefficients approach like Hoderlein et al. (2017) is possible but complicated in this context with somewhat limited sample size. Alternatively, one can use the control function approach of Section 3.3.

A3. Conditional independence $\Gamma_i \perp\!\!\!\perp W_i | Z_i$, where Z_i is experience.

This allows us to check the robustness to the baseline assumption of full independence in A1. The use of teacher training or experience in A2 and A3 is motivated by the fact that it can strengthen pedagogy and thus the transfer of knowledge to students. I also use a dummy for public school in A2, as differences in funding may affect both the effect of knowledge on performance and the knowledge of teachers through greater access to personal development.

I proceed according to the following steps:

1. I first estimate the parameters in (40) using linear regression with fixed effects, then form the quantities $\widetilde{Y}_{j,it}$.

2. I then use the following nonparametric estimator $\widehat{\sigma}_\varepsilon^2$ of σ_ε^2:

$$\widehat{\sigma}_\varepsilon^2 = \frac{\sum_i \sum_t \sum_j (\widetilde{Y}_{j,i,t} - \widetilde{Y}_{i,t})}{\sum_j \sum_t (n_{i,t} - 1)}. \tag{43}$$

3. The third step uses linear regression to estimate the coefficients β in (42) and bring the model back to (16), using $p = 1$ regressor.

4. I then either use:

   - the $\widehat{\text{PE}}^{GT}$ estimator; Under the normality assumption for the distribution of ε, (21) bypasses the need to perform a deconvolution step. I choose the adaptive choice (96) for the tuning parameters.
   - the $\widehat{\text{PE}}^{GWB}$ estimator; Since mean teacher knowledge can be considered as a continuous random variable, I use the discretization procedure described in Section 4.2.4. The results are not sensitive to taking large values of N, and I provide a robustness result with respect to the discretization parameter K in the Appendix.

My preferred specification is the varying coefficients one A2 (ii), and the results for the GT estimator are shown in Figure 2, while Figure 5 in appendix presents the one for GWB. Appendix D gathers the alternative estimation procedures with A1 and A3, and I discuss robustness below.

Specifically, Figure 2(a) shows the joint individual-level distribution in the sample of the predicted fraction of TVA that cannot be explained by teacher knowledge (*i.e.*, $\Gamma_1 + X_{-1}^\top \gamma$) and the estimated effect of teacher knowledge on TVA (Γ_2), conditional on the observed information about TVA and knowledge. Figure 2(b) shows the joint individual-level distribution in the sample of the estimated TVA and the estimated effect of teacher knowledge on TVA (Γ_2).

A first conclusion from Figure 2(a) is that there is important heterogeneity in the sample, with the effect of teacher knowledge on student test scores ranging from barely positive to 0.75 SD. This can be compared to the average annual test score gain of 0.33 SD for the student cohort over all four years of the sample. The distribution appears to be unimodal.

A second interesting fact is that this effect seems to be strongly negatively correlated with the unexplained part of the TVA (-0.75 and -0.90 for the GT and GWB estimators with IV and A2, respectively). This means that teachers for whom the effect of knowledge on performance is weaker are also those for whom a relatively large share of their value added comes from other sources. It is possible to identify these individuals using past scores, and this correlation is not necessarily expected. On Figure 2(b), we can see a positive correlation between the predicted individual effect of knowledge on performance and that performance. This also shows some important heterogeneity, with a significant proportion of individuals having average value added, but also with a very weak predicted effect of knowledge on the latter.

On the technical side, it is reassuring that both the GT and the GWB reach similar conclusions. The predictions of Γ_1 and Γ_2 are also strongly correlated (0.45 and 0.57 for the varying coefficients specification, respectively). Both estimators are fast to compute at these sample sizes, with the GWB and GT estimators taking 2 min and 1 min, respectively, to perform the estimation and generate the predictions.[12] Finally, I check the sensitivity of the different relaxations of independence assumption A1. Table 6 in the appendix presents characteristics of the individual level differences in the predictions between the different specifications, for both GT and GW. If the GT estimator seems to be only slightly affected by the use of additional variables Z

[12]These CPU times are obtained using R and Python code, parallelized on 4 CPUs on an Intel(R) Core i7-9850H CPU 2.60GHz with 16Gb of RAM.

(median change of -0.01), this is less true for the GWB estimator (0.07). However, compared to the case of A2 (ii), the other two specifications using Z yield only limited differences. This motivates the choice of A2 (ii) as my preferred specification.

5.3 Using PE matters empirically for policy design

Because of the potential heterogeneity in Γ, there could be important gains from targeting some on-the-job training policies to the population that would benefit the most in terms of performance. This is an application of the motivation developed in Section 2.2.

Specifically, consider the assignment to an on-the-job training program D based on initially available data on teacher knowledge and performance $(X_{i,t}, Y_{i,t})$, and with a conditional average treatment effect $\text{CATE}(X_{i,t}, Y_{i,t})$ on the average content knowledge of teachers $X_{i,t+1}$ in $t+1$, under capacity constraints.

The decision problem is then to select teachers into this training so as to maximize average utility, taken as TVA. According to Section 2.2 and under Assumption 1, this yields the optimal allocation decision rule p as the product of the CATE and the PE given in equation (7). A natural alternative empirical policy without considering this heterogeneity, given that the effect of teacher knowledge on performance ($\mathbb{E}(\Gamma_{2,i})$) is significantly positive (0.239), would be to allocate individuals to D to maximize the increase in knowledge, *i.e.*, based only on the CATE:

$$p(X_{i,t}, Y_{i,t}) = \mathbb{1}\{\mathbb{E}(\Gamma_{2,i})\ \text{CATE}(X_{i,t}, Y_{i,t}) \geq \gamma\}. \tag{44}$$

such that $\gamma = E(p(X_{i,t}, Y_{i,t}))$. I now compare the effects of such policies.

Unfortunately, to my knowledge, there is no randomized experiment evaluating the effect of such a policy in Pakistan. The closest I have found is that of Jakob et al. (2023) in Tanzania, where the estimated heterogeneous treatment effect in SD of standardized test scores measuring teacher knowledge is of the form[13]

$$\text{CATE}(X_{i,t}) = \max(0, 0.131 - 0.475X_{i,t}).$$

[13]For simplicity, I keep the same notation for standardized and nonstandardized test scores, but treat this in the estimation.

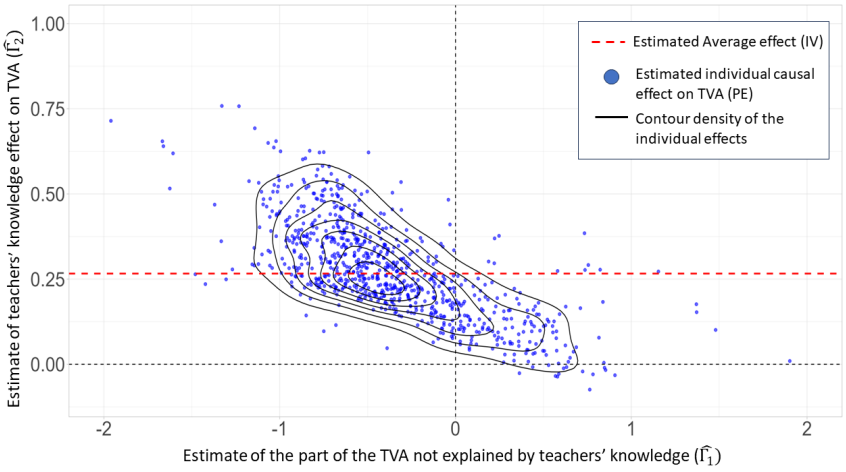
This policy thus has an effect on teachers who know less.[14] Here, the treatment is most effective on those with little prior knowledge and has no effect on those who know more.

However, this might affect the teachers' performance $Y_{i,t+1}$ differently among them, since the individual effects Γ_2 are very heterogeneous. To illustrate this while completing my estimation analysis, Figure 3 is the analog of Figure 2(b), showing the estimated effect (with GWB A2 (ii)) of teacher knowledge on TVA as a function of the latter, but only for the 20% of individuals at the bottom of the knowledge distribution, *i.e.*, those where $X_{i,t+1}$ will be more affected by the treatment. It also presents confidence intervals that identify the individuals in this population for whom an increase in their knowledge following the treatment on X is predicted to have a significant impact on Y. This appears to be the case for individuals with very different estimated TVA.

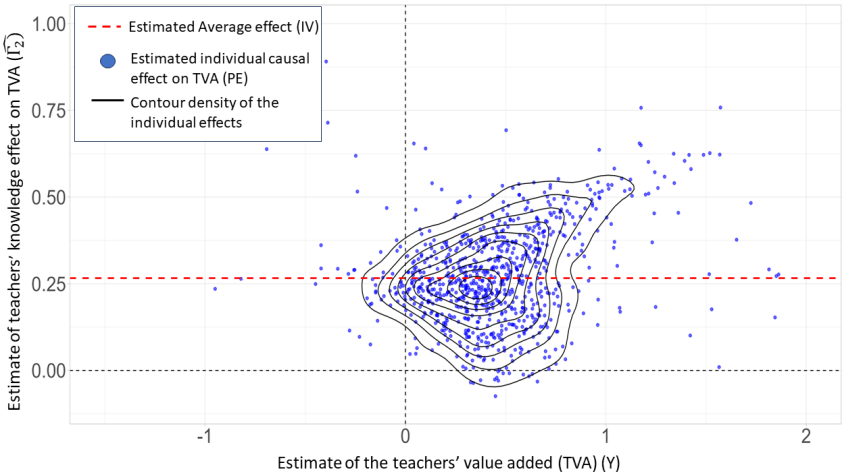
Finally, Figure 4 compares the allocation based on $\mathbb{E}(\Gamma_{2,i})\text{CATE}(X_{i,t})$ (x-axis) with the one based on $\text{PE}(X_{i,t}, Y_{i,t})$ $\text{CATE}(X_{i,t})$ (y-axis), as in (44). It represents the joint distributions of these predicted effects, where both scales can be interpreted directly as the impact of this personal development program on the student's test scores in SD. The two plain black lines represent the thresholds above which teachers would be assigned to such a program when treating 20% of the population. In this experiment, individuals shown in green (or red) would be treated (or not treated) by both selection rules. However, the optimal policy would treat those individuals shown in blue, as they have a strong predicted effect of knowledge on their performance. It would not treat those individuals shown in purple, who have low levels of knowledge but for whom such treatment would also be inefficient.

Table 4 summarizes the estimated welfare gains associated with Figure 4. It shows that the average gains from informing the decision with the PE would vary up to 31.1% (resp. 22.4%) treating 10% (resp. 20%) of the population. Table 4 also shows that the policy based on PE tends to select less systematically individuals with less knowledge (average of 2.16 compared to 2.02 when selecting 20%).

[14] Unfortunately, I cannot also model the heterogeneity of the treatment effect on X_{t+1} with respect to past values of students' test scores $Y_{i,t}$. It is reasonable to think that this heterogeneity is limited.



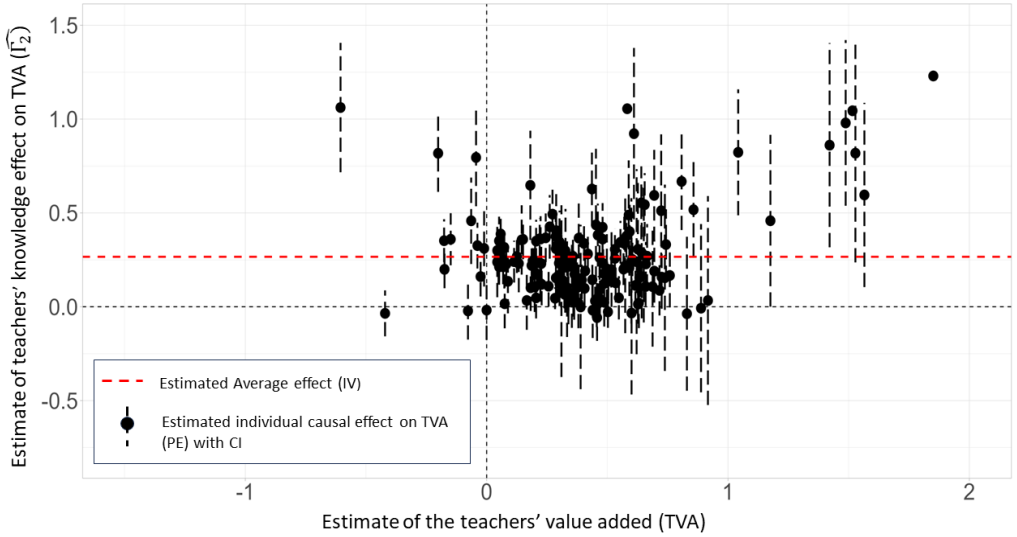
(a) Joint distribution of posterior estimates of $(\Gamma_{1,i} + X_{-1,i}^{\top}\gamma, \Gamma_{2,i})$



(b) Joint distribution of estimates of TVA_i and posterior estimates of $\Gamma_{2,i}$

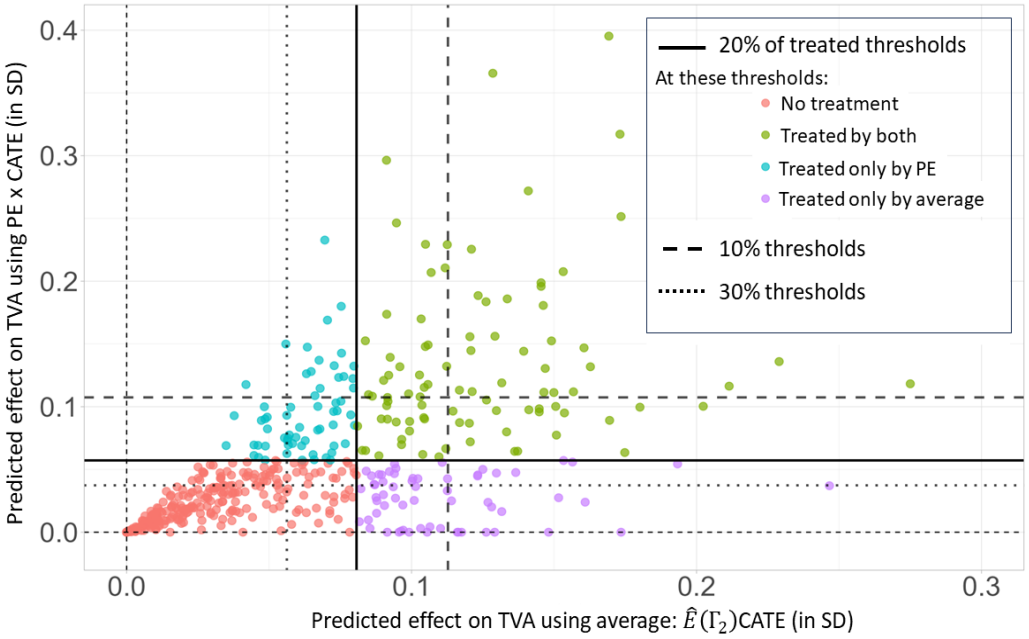
Notes: These results pool teachers from private and public schools. Figure 2(a) (resp. 2(b)) presents the estimated individual-level joint distribution of the part of the TVA that cannot be explained by teacher knowledge (resp. the estimated TVA) and the estimated effect of teacher knowledge on TVA (Γ_2). This is done using the GT estimator with varying coefficients A3 (ii), when we instrument for the teacher's mean score in the first tested year with the mean score of the second year, which reduces the sample size to 834. The dots represent the individual predictions $\widehat{\text{PE}}(X_i, Y_i)$ and the contour lines the levels of the associated fitted density. The dotted red line represents the IV estimates with an homogeneous specification (0.239). Teachers' tests scores are winsorized at a 1% level.

Figure 2: Distributions of the estimates of coefficients characterizing the TVA



Notes: These results pool teachers from private and public schools. It presents the estimated individual-level joint distribution the estimated TVA and the estimated effect of teacher knowledge on TVA (Γ_2). Estimation is performed using the GWB estimator with varying coefficients (A2 (ii)) and 95% confidence intervals (displayed in dotted black lines) are computed using subsampling. Teachers' tests scores are winsorized at a 1% level.

Figure 3: Estimated PE of knowledge on TVA for the 20% teachers' with less content knowledge



Notes: These results present the predicted effects in SD of students' tests scores based on the CATE only $\mathbb{E}(\Gamma_{2,i})\text{CATE}(X_{i,t})$ (x-axis) versus the predicted effects based on PE also $\text{PE}(X_{i,t}, Y_{i,t})$ $\text{CATE}(X_{i,t})$ (y-axis), which forms the optimal decision rule. The two plain black lines (resp. dotted and dashed) represent the threshold above which teachers would be allocated to such a program when treating 20% (resp. 10% and 30%) of the population. In this experiment, individuals represented in green (resp. in red) would be treated (resp. not treated) by both selection rules. However, the optimal policy would treat the individuals with strong predicted effect of knowledge on their performances displayed in blue, and does not treat individuals displayed in purple. Estimation is performed using the GT estimator with varying coefficients A2 (ii)).

Figure 4: Comparison of the rules based on CATE or PE $\times$ CATE

% of treated population	10%	20%	30%	50%
Estimated average welfare	0.20	0.14	0.11	0.07
Estimated gains from using PE (in SD)	0.05	0.03	0.01	0.00
In %	31.1	22.4	9.5	0.8
Average knowledge on treated with PE	1.94	2.16	2.29	2.47
Average knowledge on treated without PE	1.72	2.02	2.19	2.44

Notes: "Estimated average welfare" is the average teachers' value added $\mathbb{E}(Y_{t+1}(X_{t+1}(D)))$ under the policy with PE. "Estimated welfare gains from using PE (in SD)" are the estimated gains compared to the policy not using PE, given in percentage in "In %". "Average knowledge on treated" is the mean of $X_{i,t}$ in our population under the different policies. Estimation is performed using the GT estimator with varying coefficients A2 (ii).

Table 4: Table of estimated gains from using PE

6 Conclusion

I study the identification and inference of posterior effects in linear models. My baseline model is a stepping stone to predicting the heterogeneity of the effect of some covariates in many more complicated and empirically relevant situations, such as the analysis of the determinants of teachers' value added. A major difficulty with this model is providing estimators that allow for realistic variation in the regressors, either discrete or continuous. My approaches break new ground by providing tools to predict these effects nonparametrically in these two different contexts. My application illustrates that these are tractable estimators that provide more accurate descriptions and allow for more fine-tuning of policies by informing them of the heterogeneity of the effects. On valuable extensions in the context of this paper would be performing bias-aware inference (see Armstrong and Kolesár, 2020; Armstrong et al., 2020). Finally, my second method shows the potential of optimal transport tools to allow the estimation of posterior effects in other models than those considered here.

References

Agueh, M. and G. Carlier (2011). Barycenters in the wasserstein space. *SIAM Journal on Mathematical Analysis 43*(2), 904–924.

Agueh, M. and G. Carlier (2017). Vers un théorème de la limite centrale dans l'espace de Wasserstein? *Comptes Rendus Mathématiques 355*(7), 812–818.

Améndola, C., J.-C. Faugere, and B. Sturmfels (2015). Moment varieties of Gaussian mixtures. *arXiv preprint arXiv:1510.04654* .

Angrist, J. D., P. D. Hull, P. A. Pathak, and C. R. Walters (2017). Leveraging lotteries for school value-added: Testing and estimation. *The Quarterly Journal of Economics 132*(2), 871–919.

Arellano, M. and S. Bonhomme (2012). Identifying distributional characteristics in random coefficients panel data models. *The Review of Economic Studies 79*(3), 987–1020.

Arellano, M. and S. Bonhomme (2023). Recovering latent variables by matching. *Journal of the American Statistical Association 118*(541), 693–706.

Armstrong, T. B. and M. Kolesár (2020). Simple and honest confidence intervals in nonparametric regression. *Quantitative Economics 11*(1), 1–39.

Armstrong, T. B., M. Kolesár, and S. Kwon (2020). Bias-aware inference in regularized regression models. *arXiv preprint arXiv:2012.14823*.

Armstrong, T. B., M. Kolesár, and M. Plagborg-Møller (2022). Robust empirical Bayes confidence intervals. *Econometrica 90*(6), 2567–2602.

Athey, S. and G. Imbens (2016). Recursive partitioning for heterogeneous causal effects. *Proceedings of the National Academy of Sciences 113*(27), 7353–7360.

Athey, S., J. Tibshrani, and S. Wager (2019). Generalized random forests. *The Annals of Statistics 47*(2), 1148–1178.

Bau, N. and J. Das (2020). Teacher value added in a low-income country. *American Economic Journal: Economic Policy 12*(1), 62–96.

Beran, R., A. Feuerverger, and P. Hall (1996). On nonparametric estimation of intercept and slope distributions in random coefficient regression. *Annals of Statistics 24*, 2569–2592.

Beran, R. and P. Hall (1992). Estimating coefficient distributions in random coefficient regressions. *Annals of Statistics 20*, 1970–1984.

Beran, R. and W. Millar (1994). Minimum distance estimation in random coefficient regression models. *Annals of Statistics 22*, 1976–1992.

Bold, T., D. Filmer, G. Martin, E. Molina, B. Stacy, C. Rockmore, J. Svensson, and W. Wane (2017). Enrollment without learning: Teacher effort, knowledge, and skill in primary schools in africa. *Journal of Economic Perspectives 31*(4), 185–204.

Bonhomme, S. and M. Weidner (2022). Posterior average effects. *Journal of Business & Economic Statistics 40*(4), 1849–1862.

Borwein, P. and T. Erdélyi (1995). *Polynomials and polynomial inequalities*, Volume 161. Springer Science & Business Media.

Breunig, C. (2021). Varying random coefficient models. *Journal of Econometrics 221*(2), 381–408.

Breunig, C. and S. Hoderlein (2018). Specification testing in random coefficient models. *Quantitative Economics 9*(3), 1371–1417.

Brown, L. D. and E. Greenshtein (2009). Nonparametric empirical Bayes and compound decision approaches to estimation of a high-dimensional vector of normal means. *The Annals of Statistics*, 1685–1704.

Brunel, E., F. Comte, and C. Lacour (2010). Minimax estimation of the conditional cumulative distribution function. *Sankhya A 72*, 293–330.

Cai, T. T. (2002). On adaptive wavelet estimation of a derivative and other related linear inverse problems. *Journal of Statistical Planning and Inference 108*(1-2), 329–349.

Carlier, G., A. Delalande, and Q. Merigot (2022). Quantitative stability of barycenters in the Wasserstein space. *arXiv preprint arXiv:2209.10217*.

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

Charlier, I., D. Paindaveine, and J. Saracco (2015). Quantifquantile: an R package for performing quantile regression through optimal quantization. *The R Journal*.

Chernozhukov, V., I. Fernandez-Val, S. Hoderlein, H. Holzmann, and W. Newey (2015). Nonparametric identification in panels using quantiles. *Journal of Econometrics 188*(2), 378–392.

Chernozhukov, V., A. Galichon, M. Hallin, and M. Henry (2017). Monge-kantorovich depth, quantiles, ranks and signs. *Annals of Statistics 45*(1), 223–256.

Chetty, R., J. N. Friedman, and J. E. Rockoff (2014a). Measuring the impacts of teachers i: Evaluating bias in teacher value-added estimates. *American Economic Review 104*(9), 2593–2632.

Chetty, R., J. N. Friedman, and J. E. Rockoff (2014b). Measuring the impacts of teachers ii: Teacher value-added and student outcomes in adulthood. *American Economic Review 104*(9), 2633–2679.

Chetty, R. and N. Hendren (2018). The impacts of neighborhoods on intergenerational mobility ii: County-level estimates. *The Quarterly Journal of Economics 133*(3), 1163–1228.

Cohen, A., I. Daubechies, and P. Vial (1993). Wavelets on the interval and fast wavelet transforms. *Applied and Computational Harmonic Analysis*.

Cohen, N., E. Greenshtein, and Y. Ritov (2013). Empirical Bayes in the presence of explanatory variables. *Statistica Sinica*, 333–357.

Comte, F., V. Genon-Catalot, and A. Samson (2013). Nonparametric estimation for stochastic differential equations with random effects. *Stochastic Process. Appl. 123*(7), 2522–2551.

Comte, F. and C. Lacour (2013). Anisotropic adaptive kernel deconvolution. In *Annales de l'IHP Probabilités et statistiques*, Volume 49, pp. 569–609.

Cuturi, M. (2013). Sinkhorn distances: Lightspeed computation of optimal transport. In *Advances in Neural Information Processing Systems*, pp. 2292–2300.

Cuturi, M. and A. Doucet (2014). Fast computation of Wasserstein barycenters. In *International conference on machine learning*, pp. 685–693. PMLR.

Das, J. and T. Zajonc (2010). India shining and Bharat drowning: Comparing two Indian states to the worldwide distribution in mathematics achievement. *Journal of Development Economics 92*(2), 175–187.

Delaigle, A. and A. Meister (2008). Density estimation with heteroscedastic error. *Bernoulli 14*(2), 562–579.

Delon, J., N. Gozlan, and A. Saint-Dizier (2022). Generalized wasserstein barycenters between probability measures living on different subspaces. *Annals of Applied Probability*.

D'Haultfoeuille, X., C. Gaillac, and A. Maurel (2021). Rationalizing rational expectations: Characterizations and tests. *Quantitative Economics 12*(3), 817–842.

D'Haultfœuille, X., C. Gaillac, and A. Maurel (2022). Partially linear models under data combination. *arXiv preprint arXiv:2204.05175*.

Dion, C. (2014). New adaptive strategies for nonparametric estimation in linear mixed models. *Journal of Statistical Planning and Inference 150*, 30–48.

Dunker, F., K. Eckle, K. Proksch, and J. Schmidt-Hieber (2019). Tests for qualitative features in the random coefficients model. *Electronic Journal of Statistics 13*(2), 2257–2306.

Efron, B. (2011). Tweedie's formula and selection bias. *Journal of the American Statistical Association 106*(496), 1602–1614.

Efron, B. (2012). *Large-scale inference: empirical Bayes methods for estimation, testing, and prediction*, Volume 1. Cambridge University Press.

Efron, B. (2014). Two modeling strategies for empirical Bayes estimation. *Statistical Science 29*(2), 285.

Evdokimov, K. and H. White (2012). Some extensions of a lemma of kotlarski. *Econometric Theory 28*(4), 925–932.

Fay, R. E. I. and R. A. Herriot (1979). Estimates of income for small places: an application of James-Stein procedures to census data. *Journal of the American Statistical Association 74*(366a), 269–277.

Finkelstein, A., M. Gentzkow, and H. Williams (2021). Place-based drivers of mortality: Evidence from migration. *American Economic Review 111*(8), 2697–2735.

Flamary, R., N. Courty, A. Gramfort, M. Z. Alaya, A. Boisbunon, S. Chambon, L. Chapel, A. Corenflos, K. Fatras, N. Fournier, L. Gautheron, N. T. Gayraud, H. Janati, A. Rakotomamonjy, I. Redko, A. Rolet, A. Schutz, V. Seguy, D. J. Sutherland, R. Tavenard, A. Tong, and T. Vayer (2021). Pot: Python optimal transport. *Journal of Machine Learning Research 22*(78), 1–8.

Florens, J.-P., J. J. Heckman, C. Meghir, and E. Vytlacil (2008). Identification of treatment effects using control functions in models with continuous, endogenous treatment and heterogeneous effects. *Econometrica 76*(5), 1191–1206.

Frogner, C. and T. Poggio (2019). Fast and flexible inference of joint distributions from their marginals. In *International Conference on Machine Learning*, pp. 2002–2011.

Gaillac, C. (2021). Some problems related to random coefficients models and data combination in economics. *PhD thesis, Toulouse School of Economics*.

Gaillac, C. and E. Gautier (2021a). Estimates for the SVD of the truncated fourier transform on $L^2\left(\exp(b|\cdot|)\right)$ and stable analytic continuation. *Journal of Fourier Analysis and Applications 27*(4), 72.

Gaillac, C. and E. Gautier (2021b). Nonparametric classes for identification in random coefficients models when regressors have limited variation. *Preprint arxiv arXiv:2105.11720*.

Gaillac, C. and E. Gautier (2022). Adaptive estimation in the linear random coefficients model when regressors have limited variation. *Bernoulli 28*(1), 504–524.

Galichon, A. (2018). *Optimal transport methods in economics*. Princeton University Press.

Galichon, A. and M. Henry (2011). Set identification in models with multiple equilibria. *The Review of Economic Studies 78*(4), 1264–1298.

Gelman, A., D. K. Park, S. Ansolabehere, P. N. Price, and L. C. Minnite (2001). Models, assumptions and model checking in ecological regressions. *Journal of the Royal Statistical Society: Series A (Statistics in Society) 164*(1), 101–118.

Gilraine, M., J. Gu, and R. McMillan (2020). A new method for estimating teacher value-added. Technical report, National Bureau of Economic Research.

Giné, E. and R. Nickl (2016). *Mathematical foundations of infinite-dimensional statistical models*, Volume 40. Cambridge University Press.

Goldenshluger, A. and O. Lepski (2014). On adaptive minimax density estimation on $\mathbb{R}^d$. *Probability Theory Related Fields 159*, 479–543.

Goldfeld, Z., K. Kato, G. Rioux, and R. Sadhu (2022). Limit theorems for entropic optimal transport maps and the sinkhorn divergence. *arXiv preprint arXiv:2207.08683*.

Goodman, L. A. (1959). Some alternatives to ecological correlation. *American Journal of Sociology 64*(6), 610–625.

Graf, S. and H. Luschgy (2007). *Foundations of quantization for probability distributions*. Springer.

Gu, J. and R. Koenker (2017). Unobserved heterogeneity in income dynamics: An empirical Bayes perspective. *Journal of Business & Economic Statistics 35*(1), 1–16.

Gu, J. and R. Koenker (2022). Nonparametric maximum likelihood methods for binary response models with random coefficients. *Journal of the American Statistical Association 117*(538), 732–751.

Gu, J. and R. Koenker (2023). Invidious comparisons: Ranking and selection as compound decisions. *Econometrica 91*(1), 1–41.

Gunsilius, F. F. (2023). Distributional synthetic controls. *Econometrica 91*(3), 1105–1117.

Hahn, M. G. and E. T. Quinto (1985). Distances between measures from 1-dimensional projections as implied by continuity of the inverse Radon transform. *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete 70*(3), 361–380.

Hall, P., J. Racine, and Q. Li (2004). Cross-validation and the estimation of conditional probability densities. *Journal of the American Statistical Association 99*(468), 1015–1026.

Hanushek, E. A. et al. (2009). Teacher deselection. *Creating a new teaching profession 168*, 172–173.

Hoderlein, S., H. Holzmann, and A. Meister (2017). The triangular model with random coefficients. *Journal of Econometrics 201*, 144–169.

Hoderlein, S., J. Klemelä, and E. Mammen (2010). Analyzing the random coefficient model nonparametrically. *Econometric Theory 26*, 804–837.

Hoderlein, S. and E. Mammen (2007). Identification of marginal effects in nonseparable models without monotonicity. *Econometrica 75*(5), 1513–1518.

Hoderlein, S. and E. Mammen (2009). Identification and estimation of local average derivatives in non-separable models without monotonicity. *The Econometrics Journal 12*(1), 1–25.

Hoderlein, S. and Y. Sasaki (2013). Outcome conditioned treatment effects. Technical report, cemmap working paper.

Ignatiadis, N. and S. Wager (2019). Covariate-powered empirical Bayes estimation. *Advances in Neural Information Processing Systems 32*.

Ignatiadis, N. and S. Wager (2022). Confidence intervals for nonparametric empirical Bayes analysis. *Journal of the American Statistical Association 117*(539), 1149–1166.

Imai, K., Y. Lu, and A. Strauss (2008). Bayesian and likelihood inference for 2× 2 ecological tables: an incomplete-data approach. *Political Analysis 16*(1), 41–69.

Imai, K., Y. Lu, A. Strauss, et al. (2011). Eco: R package for ecological inference in 2x2 tables. *Journal of Statistical Software 42*(i05).

Imbens, G. W. and W. K. Newey (2009). Identification and estimation of triangular simultaneous equations models without additivity. *Econometrica 77*(5), 1481–1512.

Jacob, B. A. and L. Lefgren (2008). Can principals identify effective teachers? evidence on subjective performance evaluation in education. *Journal of Labor Economics 26*(1), 101–136.

Jakob, M., K. Büchel, D. Steffen, and A. Brunetti (2023). Participatory teaching improves learning outcomes: Evidence from a field experiment in tanzania. Technical report, Working Paper.

James, W. and C. Stein (1992). Estimation with quadratic loss. In *Breakthroughs in statistics: Foundations and basic theory*, pp. 443–460. Springer.

Jiang, W. and C.-H. Zhang (2009). General maximum likelihood empirical bayes estimation of normal means. *The Annals of Statistics 37*(4), 1647–1684.

Johnstone, I. M. and B. W. Silverman (2004). Needles and straw in haystacks: Empirical Bayes estimates of possibly sparse sequences. *The Annals of Statistics 32*(1), 1594–1649.

Kasy, M. (2022). Who wins, who loses? identification of conditional causal effects, and the welfare impact of changing wages. *Journal of Econometrics 226*(1), 155–170.

Katz, J. N. and G. King (1999). A statistical model for multiparty electoral data. *American Political Science Review 93*(1), 15–32.

King, G. (1997). *A solution to the ecological inference problem: Reconstructing individual behavior from aggregate data*. Princeton University Press.

Kitagawa, T. and A. Tetenov (2018). Who should be treated? empirical welfare maximization methods for treatment choice. *Econometrica 86*(2), 591–616.

Kline, P., E. K. Rose, and C. R. Walters (2022). Systemic discrimination among large us employers. *The Quarterly Journal of Economics 137*(4), 1963–2036.

Koenker, R. and I. Mizera (2014). Convex optimization, shape constraints, compound decisions, and empirical Bayes rules. *Journal of the American Statistical Association 109*(506), 674–685.

Kotlarski, I. (1967). On characterizing the gamma and the normal distribution. *Pacific Journal of Mathematics 20*(1), 69–76.

Lacour, C. and P. Massart (2016). Minimal penalty for Goldenshluger–Lepski method. *Stochastic Processes and their Applications 126*(12), 3774–3789.

Le Gouic, T. and J.-M. Loubes (2017). Existence and consistency of Wasserstein barycenters. *Probability Theory and Related Fields 168*, 901–917.

Lohöfer, G. (1998). Inequalities for the associated Legendre functions. *Journal of Approximation Theory 95*(2), 178–193.

Manski, C. F. (2004). Statistical treatment rules for heterogeneous populations. *Econometrica 72*(4), 1221–1246.

Masten, M., A. Poirier, and L. Zhang (2019). tesensitivity: A stata package for assessing the unconfoundedness assumption. In *2019 Stata Conference*, Number 51. Stata Users Group.

Masten, M. A. (2017). Random coefficients on endogenous variables in simultaneous equations models. *The Review of Economic Studies 85*(2), 1193–1250.

Masten, M. A. and A. Poirier (2018). Identification of treatment effects under conditional partial independence. *Econometrica 86*(1), 317–351.

Masten, M. A. and A. Torgovitsky (2016). Identification of instrumental variable correlated random coefficients models. *Review of Economics and Statistics 98*(5), 1001–1005.

Matzkin, R. L. (2007). Nonparametric identification. *Handbook of econometrics 6*, 5307–5368.

Mérigot, Q., F. Santambrogio, and C. Sarrazin (2021). Non-asymptotic convergence bounds for Wasserstein approximation using point clouds. *Advances in Neural Information Processing Systems 34*, 12810–12821.

Montiel Olea, J. L., B. O'Flaherty, and R. Sethi (2021). Empirical bayes counterfactuals in poisson regression with an application to police use of deadly force. *Available at SSRN 3857213*.

Newey, W. and S. Stouli (2020). Control variables, discrete instruments, and identification of structural functions. *Journal of Econometrics*.

Newey, W. K. and S. Stouli (2018). Heterogenous coefficients, discrete instruments, and identification of treatment effects. *Preprint arXiv arXiv:1811.09837*.

Olver, F., D. Lozier, R. Boisvert, and C. Clark (2010). *NIST handbook of mathematical functions*. Cambridge University Press.

Pagès, G. (2015). Introduction to vector quantization and its applications for numerics. *ESAIM: proceedings and surveys 48*, 29–79.

Peyré, G., M. Cuturi, et al. (2019). Computational optimal transport. *Foundations and Trends® in Machine Learning 11*(5-6), 355–607.

Poularikas, A. D. (2018). *Handbook of formulas and tables for signal processing*. CRC Press.

Robbins, H. (1956). An empirical Bayes approach to statistics. in proceedings of the third berkeley symposium of mathematical statistics and probability.

Robbins, H. (1964). The empirical Bayes approach to statistical decision problems. *The Annals of Mathematical Statistics 35*(1), 1–20.

Rockoff, J. E. (2004). The impact of individual teachers on student achievement: Evidence from panel data. *American Economic Review 94*(2), 247–252.

Rosen, O., W. Jiang, G. King, and M. A. Tanner (2001). Bayesian and frequentist inference for ecological inference: The R×C case. *Statistica Neerlandica 55*(2), 134–156.

Rothstein, J. (2010). Teacher quality in educational production: Tracking, decay, and student achievement. *The Quarterly Journal of Economics 125*(1), 175–214.

Rudin, W. (1973). *Real and complex analysis*. McGraw-Hill.

Rullgård, H. and E. T. Quinto (2010). Local Sobolev estimates of a function by means of its Radon transform. *Inverse Problems and Imaging 4*(4), 721–734.

Soloff, J. A., A. Guntuboyina, and B. Sen (2021). Multivariate, heteroscedastic empirical bayes via nonparametric maximum likelihood. *arXiv preprint arXiv:2109.03466*.

Stoye, J. (2009). Minimax regret treatment choice with finite samples. *Journal of Econometrics 151*(1), 70–81.

Tam Cho, W. K. (1998). Iff the assumption fits?: A comment on the king ecological inference solution. *Political Analysis 7*, 143–163.

Tam Cho, W. K. and B. J. Gaines (2004). The limits of ecological inference: The case of split-ticket voting. *American Journal of Political Science 48*(1), 152–171.

Tsybakov, A. (2000). On the best rate of adaptive estimation in some inverse problems. *Comptes Rendus de l'Academie des Sciences - Mathematics 330*, 835–840.

Tsybakov, A. (2008). *Introduction to nonparametric estimation*. Springer.

Ullah, A. and A. Pagan (1999). *Nonparametric econometrics*. Cambridge University Press.

Van der Vaart, A. W. (2000). *Asymptotic statistics*, Volume 3. Cambridge University Press.

Wakefield, J. (2004). Ecological inference for 2× 2 tables. *Journal of the Royal Statistical Society: Series A (Statistics in Society) 167*(3), 385–425.

Wang, X.-F. and B. Wang (2011). Deconvolution estimation in measurement error models: the R package decon. *Journal of statistical software 39*(10).

Wooldridge, J. (2010). *Econometric analysis of cross section and panel data*. MIT Press.

Appendix

Table of Contents

A Main proofs **2**

A.1 Identification 2

A.2 Inference 7

B Data-driven rule for selecting the tuning parameters **19**

C Asymptotic normality for the GT estimator **25**

D Additional material on the application **29**

E The Tweedie formula and extension **37**

F Additional Monte-Carlo simulations **38**

F.1 With conditional independence and continuous Z 38

F.2 Monte-Carlo simulations in the panel model 39

G Nonparametric Ecological inference **40**

G.1 Application to the ecological inference model 40

G.2 Results completing Section G.1 with 2 choices 41

G.3 Extension to identification in Ecological inference with more than two choices 42

G.4 Partial identification when $d_C > 2$ 44

G.5 Identification when $d_C = 3$ when restricting the dimension of the unobserved heterogeneity 45

G.6 Proofs of Appendix G.3 50

G.7 Comparison with ground truth in an election dataset: turnout by race 56

A Main proofs

Notations

I use that for all $k, l > 0$, $N \geq 1$,

$$(N+l)^k \leq ((l+1)N)^k, \tag{45}$$

$$|\{k \in \mathbb{N}_0^p : \ |k|_\infty \leq j_0\}| = (j_0+1)^p. \tag{46}$$

I endow $\mathcal{P}_2(\mathcal{S})$ with the Wasserstein distance W_2, defined for any $\rho, \mu \in \mathcal{P}_2(\mathcal{S})$ by

$$W_2(\rho,\mu) = \left(\min_{F_{X,Y}: F_X \sim \rho, F_Y \sim \mu} \mathbb{E}(\|X-Y\|^2)\right)^{1/2},$$

where the minimum is taken over the set of joint distributions satisfying the marginal constraints. For a measure μ, a vector u and a linear projection P_u onto the vector space generated by u, I denote by $P_{u,\#}\mu$ the pushforward of μ by P_u, *i.e.*, the measure on $\mathbb{R}$ such that for any Borelian $A \subset \mathbb{R}$, $(P_{u,\#}\mu)(A) = \mu(P_u^{-1}(A))$.

A.1 Identification

Proof of Proposition 1. Using Bayes' theorem for the second equality, we have for a.e. $(x,y) \in \text{Supp}(X,Y)$ and for all $k = 1, \dots, p+1$,

$$\begin{aligned}
\mathbb{E}(\Gamma_k | X=x, Y=y) &= \int_{\mathbb{R}^{p+1}} g_k d\mathbb{P}_{\Gamma|X,Y}(g|x,y) \\
&= \int_{\mathbb{R}^{p+1}} g_k \frac{\mathbb{P}_{Y|\Gamma,X}(y|g,x)}{\mathbb{P}_{Y|X}(y|x)} d\mathbb{P}_{\Gamma|X}(g|x) \\
&= \int_{g \in \mathcal{I}(x,y)} \frac{g_k}{\mathbb{P}_{Y|X}(y|x)} d\mathbb{P}_\Gamma(g) \text{ (using Assumption 2).}
\end{aligned} \tag{47}$$

This yields for all $(x,y) \in \text{Supp}(X,Y)$ and $k = 1, \dots, p+1$,

$$\mathbb{E}(\Gamma_k | X=x, Y=y)\mathbb{P}_{Y|X}(y|x) = \int_{g \in \mathcal{I}(x,y)} g_k d\mathbb{P}_\Gamma(g). \tag{48}$$

Using Theorem 1 in Gaillac and Gautier (2021b), $\mathbb{P}_\Gamma$ is identified under assumption 3-(A) and 3-(B), which yields the result for PE_k for $k = 1, \dots, p+1$.

Let us prove statement 2. We obtain equation (10) using directly Bayes theorem as in (48). Then, using statement 1, $\mathbb{P}_\Gamma$ is the unique distribution $Q \in \mathcal{P}_2(\mathbb{R}^{p+1})$ such

that $P_{(1,x),\#}Q = \mathbb{P}_{Y|X=x}$ for all $x \in \text{Supp}(X)$. Thus, it is the unique minimizer of (11). This yields the result.

We know turn to the proof of statement 3. Denote by $\varphi_{Y|X} : (t,x) \in \mathbb{R} \times \text{Supp}(X) \mapsto \mathbb{E}(e^{itY}|X = x) = \mathcal{F}[\mathbb{P}_\Gamma](t,tx)$. Using (48) and Lemma 1, the Fourier transform of $y \mapsto \int_{g \in \mathcal{I}(x,y)} g_k d\mathbb{P}_\Gamma(g)$ is well defined (see, *e.g.*, Theorem 9.13 in Rudin, 1973). Using the definition of $\mathcal{I}(x,y)$ for the second equality which yields that $g \in \mathcal{I}(x,y)$ if and only if $y = g^\top(1,x)$, and using the definition of the Fourier transform we have,

$$\begin{aligned}
\mathcal{F}\left[\int_{g\in\mathcal{I}(x,\cdot)} g_k d\mathbb{P}_\Gamma(g)\right](t) &= \int e^{ity} \int \mathbb{1}\{g \in \mathcal{I}(x,y)\} g_k d\mathbb{P}_\Gamma(g) dy \\
&= \int e^{it(g^\top(1,x))} g_k d\mathbb{P}_\Gamma(g) \\
&= \mathcal{F}[\star_k \mathbb{P}_\Gamma(\star)](t,tx). \qquad (49)
\end{aligned}$$

Then, we conclude using Theorem 9.13 d) in Rudin (1973) and taking the Fourier inverse that, for all $(x,y) \in \text{Supp}(X,Y)$ and $k = 1,\dots,p+1$,

$$E(\Gamma_k|X = x, Y = y)\mathbb{P}_{Y|X}(y|x) = \mathcal{F}^{-1}[\mathcal{F}[\star_k \mathbb{P}_\Gamma(\star)](\cdot,\cdot x)](y).$$

We denote by

$$M_k : (x,y) \mapsto \mathcal{F}^{-1}[\mathcal{F}[\star_k \mathbb{P}_\Gamma(\star)](\cdot,\cdot x)](y). \qquad (50)$$

Using Assumption 4 and the dominated convergence theorem, for all $k = 1,\dots,p+1$, the function $\varphi_{Y|X}$ admits partial derivatives with respect to t and x_k. Moreover, using that $\text{Supp}(X)$ has a nonempty interior, the latter derivatives are identified on $\text{Supp}(X)$, and we obtain, for all $t \in \mathbb{R}$ and $x \in \text{Supp}(X)$,

$$\partial_{x_k}\varphi_{Y|X}(t,x) = it\mathcal{F}[\star_{k+1}\mathbb{P}_\Gamma(\star)](t,tx), \quad k = 1,\dots,p. \qquad (51)$$

We have, using (51) for the last equality, for $k = 1,\dots,p$,

$$\begin{aligned}
\partial_y M_{k+1}(x,y) &= \partial_y \mathcal{F}^{-1}[\mathcal{F}[\star_{k+1}\mathbb{P}_\Gamma(\star)](\cdot,\cdot x)](y) \\
&= -i\mathcal{F}^{-1}[\cdot\,\mathcal{F}[\star_{k+1}\mathbb{P}_\Gamma(\star)](\cdot,\cdot x)](y) \\
&= -\mathcal{F}^{-1}\left[\partial_{x_k}\varphi_{Y|X}(\cdot,x)\right](y) \qquad (52)
\end{aligned}$$

Finally, we obtain for $k = 1,\dots,p$,

$$\begin{aligned}
\partial_y M_{k+1}(x,y) &= -\mathcal{F}^{-1}\left[\partial_{x_k}\mathcal{F}\left[f_{Y|X}(\cdot|x)\right]\right](y) \\
&= -\partial_{x_k} f_{Y|X}(y|x). \qquad (53)
\end{aligned}$$

Integrating and using that assumption 3-(B) yields $\lim_{y\to-\infty} M_k(x,y) = 0$, we obtain statement (12). Equation (13) can directly be deduced from the model's equation, taking conditional expectation with respect to (X,Y) □

Lemma 1 *Under Assumption 3-(B), then the function* $y \mapsto \mathbb{E}(\Gamma_k|X = x, Y = y)\mathbb{P}_{Y|X}(y|x)$ *belongs to* $L^1(\mathbb{R}) \cap L^2(\mathbb{R})$.

Proof of Lemma 1. Let $\epsilon \in (0,1)$ and $\lambda = (1-\epsilon)/(2R)$. We have, using that if g s.t. $g^\top(1,x) = y$ then $\|g\| \geq |y|/\|(1,x)\|$,

$$\begin{aligned}
&\int_{\mathrm{Supp}(Y)} \left| \int_{\mathbb{R}^{p+1}} \mathbb{1}\{g \in \mathcal{I}(x,y)\} g_k f_\Gamma(g) dg \right| dy \\
&\leq \int_{\mathrm{Supp}(Y)} \int_{\mathbb{R}^{p+1}} \mathbb{1}\{\|g\| \geq \frac{|y|}{\|(1,x)\|}\} |g_k| f_\Gamma(g) dg dy \\
&\leq \int_{\mathrm{Supp}(Y)} e^{-\lambda y/\|(1,x)\|} dy \int_{\mathbb{R}^{p+1}} e^{\lambda\|g\|} |g_k| f_\Gamma(g) dg
\end{aligned}$$

which is finite reasoning similarly to (9). We also have, using $\lambda = (1-\epsilon)/(4R)$ and the Cauchy-Schwarz inequality,

$$\begin{aligned}
&\int_{\mathrm{Supp}(Y)} \left| \int_{\mathbb{R}^{p+1}} \mathbb{1}\{g \in \mathcal{I}(x,y)\} g_k f_\Gamma(g) dg \right|^2 dy \\
&\leq \int_{\mathrm{Supp}(Y)} e^{-\lambda y/\|(1,x)\|} dy \int_{\mathbb{R}^{p+1}} e^{\lambda\|g\|} |g_k|^2 |f_\Gamma(g)|^2 dg
\end{aligned}$$

which is finite as $f_\Gamma \in L^2(W^{\otimes(p+1)})$. □

Proof of Proposition 2. We keep the notations of Proposition 1. Additionally, for all $k,l = 1,\dots,p$, we denote by

$$M_{k,k'} : (x,y) \mapsto \mathcal{F}^{-1}\left[\mathcal{F}\left[\star_k \star_{k'} \mathbb{P}_\Gamma(\star)\right](\cdot,\cdot x)\right](y), \tag{54}$$

a quantity which is finite using (48) and a direct adaptation of Lemma 1.
Using the integrability assumptions of the partial derivatives $\partial_{x_k}\partial_{x_l} f_{Y|X}(\cdot|x)$ and the dominated convergence theorem, for all $k,l = 1,\dots,p+1$, the function $\varphi_{Y|X}$ admits partial derivatives with respect to t and x_k, x_l. Moreover, using that $\mathrm{Supp}(X)$ has a nonempty interior, the latter derivatives are identified on $\mathrm{Supp}(X)$, and we obtain, for all $t \in \mathbb{R}$ and $x \in \mathrm{Supp}(X)$,

$$\partial_{x_k}\partial_{x_l}\varphi_{Y|X}(t,x) = -t^2\mathcal{F}\left[\star_{k+1} \star_{l+1} \mathbb{P}_\Gamma(\star)\right](t,tx), \quad k,l = 1,\dots,p. \tag{55}$$

We have, using (51) for the last equality, for $k, l = 1, \ldots, p$,

$$\begin{aligned}\partial_y^2 M_{k+1,l+1}(x,y) &= \partial_y^2 \mathcal{F}^{-1}\left[\mathcal{F}\left[\star_{k+1} \star_{l+1} \mathbb{P}_\Gamma(\star)\right](\cdot,\cdot x)\right](y) \\ &= -\mathcal{F}^{-1}\left[\cdot^2 \mathcal{F}\left[\star_{k+1} \star_{l+1} \mathbb{P}_\Gamma(\star)\right](\cdot,\cdot x)\right](y) \\ &= \mathcal{F}^{-1}\left[\partial_{x_k}\partial_{x_l}\varphi_{Y|X}(\cdot,x)\right](y) \qquad (56)\end{aligned}$$

Finally, we obtain for $k, l = 1, \ldots, p$,

$$\begin{aligned}\partial_y^2 M_{k+1,l+1}(x,y) &= -\mathcal{F}^{-1}\left[\partial_{x_k}\partial_{x_l}\mathcal{F}\left[f_{Y|X}(\cdot|x)\right]\right](y) \\ &= -\partial_{x_k}\partial_{x_l} f_{Y|X}(y|x). \qquad (57)\end{aligned}$$

Integrating and using that assumption 3-(B) yields $\lim_{y'\to-\infty} \partial_y M_{k+1,l+1}(x,y') = 0$, $\lim_{y\to-\infty} M_{k+1,l+1}(x,y) = 0$, we obtain the result. □

Proof of Theorem 1. Let us start with the proof of case 3. Using the model (15), we have, for all $(t,x) \in \mathbb{R} \times \mathrm{Supp}(X)$,

$$\varphi_{Y|X}(t,x) = \mathbb{E}(e^{itY}|X=x) = \mathcal{F}\left[\mathbb{P}_\Gamma\right](t,tx)\varphi_\varepsilon(t). \qquad (58)$$

Denote by $\widetilde{\varphi}(t,x) := \varphi_{Y|X}(t,x)/\varphi_\varepsilon(t)$. Using Bayes' theorem and Assumption 2, we first have, for all $(x,y) \in \mathrm{Supp}(X,Y)$,

$$\mathbb{E}(\Gamma_k|X=x,Y=y)\mathbb{P}_{Y|X}(y|x) = \int_{(g,u)\in\mathcal{I}(x,y)} g_k f_\varepsilon(u) d\mathbb{P}_\Gamma(b) du.$$

Using (58) and a direct adaptation of Lemma 1 when $f_\varepsilon \in L^2(W)$, the Fourier transform of $y \mapsto \int_{(g,u)\in\mathcal{I}(x,y)} g_k f_\varepsilon(u) d\mathbb{P}_\Gamma(g) du$ is well defined (see, *e.g.*, Theorem 9.13 in Rudin, 1973). Using the definition of $\mathcal{I}(x,y)$ for the second equality which yields that $(g,u) \in \mathcal{I}(x,y)$ if and only if $y = g^\top(1,x) + u$, and using the definition of the Fourier transform we have,

$$\begin{aligned}\mathcal{F}\left[\int_{(g,u)\in\mathcal{I}(x,\cdot)} g_k f_\varepsilon(u) d\mathbb{P}_\Gamma(g) du\right](t) &= \int e^{it(g'(1,x))} e^{itu} g_k f_\varepsilon(u) du d\mathbb{P}_\Gamma(g) \\ &= \varphi_\varepsilon(t)\mathcal{F}\left[\star_k \mathbb{P}_\Gamma(\star)\right](t,tx). \qquad (59)\end{aligned}$$

Then, we conclude using Theorem 9.13 d) in Rudin (1973) and taking the Fourier inverse that, for all $(x,y) \in \mathrm{Supp}(X,Y)$ and $k = 1, \ldots, p+1$,

$$\mathbb{E}(\Gamma_k|X=x,Y=y)\mathbb{P}_{Y|X}(y|x) = \mathcal{F}^{-1}\left[\varphi_\varepsilon(\cdot)\mathcal{F}\left[\star_k \mathbb{P}_\Gamma(\star)\right](\cdot,\cdot x)\right](y).$$

We denote by

$$\widetilde{M}_k : (x, y) \mapsto \mathcal{F}^{-1}\left[\varphi_\varepsilon(\cdot)\mathcal{F}\left[\star_k \mathbb{P}_\Gamma(\star)\right](\cdot, \cdot x)\right](y). \tag{60}$$

Using Assumption 4 and the dominated convergence theorem, for all $k = 1, \dots, p+1$, the function $\widetilde{\varphi}$ admits partial derivatives with respect to t and x_k. Moreover, using that $\mathrm{Supp}(X)$ has a nonempty interior, the latter derivatives are identified on $\mathrm{Supp}(X)$, and we obtain, for all $t \in \mathbb{R}$ and $x \in \mathrm{Supp}(X)$,

$$\partial_t \widetilde{\varphi}(t, x) = i(1, x)' \mathcal{F}\left[\star_{1:(p+1)} \mathbb{P}_\Gamma(\star)\right](t, tx) \tag{61}$$

$$\partial_{x_k} \widetilde{\varphi}(t, x) = it\mathcal{F}\left[\star_{k+1} \mathbb{P}_\Gamma(\star)\right](t, tx), \quad k = 1, \dots, p. \tag{62}$$

We also have, for $k = 1, \dots, p$,

$$\partial_y \widetilde{M}_{k+1}(x, y) = -\mathcal{F}^{-1}\left[\partial_{x_k} \varphi(\cdot, x)\right](y),$$

hence

$$\partial_y \widetilde{M}_{k+1}(x, y) = -\partial_{x_k} f_{Y|X}(y|x). \tag{63}$$

Finally, using that for all $(t, x) \in \mathbb{R} \times \mathrm{Supp}(X)$,

$$\partial_t \widetilde{\varphi}(t, x) = \frac{\partial_t \varphi(t, x)\varphi_\varepsilon(t) - \varphi_\varepsilon'(t)\varphi(t, x)}{\varphi_\varepsilon(t)^2},$$

we have

$$\begin{aligned}
\partial_t \varphi(t, x) &= \partial_t \widetilde{\varphi}(t, x)\varphi_\varepsilon(t) + \frac{\varphi_\varepsilon'(t)}{\varphi_\varepsilon(t)}\varphi(t, x) \\
&= i\mathcal{F}\left[(1, x)'\widetilde{M}\right] + \frac{\varphi_\varepsilon'(t)}{\varphi_\varepsilon(t)}\varphi(t, x).
\end{aligned}$$

Finally, using that $\rho(x, y) = \mathcal{F}^{-1}\left[\partial_t \varphi(t, x)\right](y)/i$, we obtain

$$\rho(x, y) = (1, x)'\widetilde{M}(x, y) + \mathcal{F}^{-1}\left[\frac{\varphi_\varepsilon'}{i\varphi_\varepsilon}\varphi(\cdot, x)\right](y),$$

hence the result.

Let us continue with the proof of case 2. Equation (18) results directly from Bayes' theorem. Then, using Assumption 5, the distribution of $\widetilde{Y}_i = (1, X_i^\top)\Gamma_i$ conditional on $X_i = x$ is given by $f(\mathbb{P}_{Y|X=x})$. $\mathbb{P}_\Gamma$ is the unique distribution $Q \in \mathcal{P}_2(\mathbb{R}^{p+1})$ such that $P_{(1,x),\#}Q = \mathbb{P}_{\widetilde{Y}|X=x}$ for all $x \in \mathrm{Supp}(X)$. Thus, it is the unique minimizer of (19). This yields the result. □

Proof of Proposition 13. This can be seen as a corollary of Theorem 1-3, or as a particular case of propositions 15 or 16, see Remark 3. □

Proof of Proposition 5 This is a direct consequence of including Assumption (8) in (47). □

A.2 Inference

A.2.1 F-modeling

Formulation of the estimator with unknown f_X and $f_{Y|X}$

Assumption 12 (On the rates of convergence of the preliminary estimators) *Assume that:*

(*Est.1*) *We have estimators $\widehat{f}_X$ based on a preliminary sample $\mathcal{P}_{n_0} = (X_i)_{i=-n_0+1}^{0}$ independent of $(X_i, Y_i)_{i=1}^{n}$ and $\widehat{f}_{Y|X}$ based on a second preliminary sample $\mathcal{P}_{n_1} = (X_i)_{i=-(n_1+n_0)+1}^{-n_0}$ independent of $(X_i, Y_i)_{i=-n_0}^{n}$;*

(*Est.2*) *$\mathcal{E}$ and $\mathcal{E}'$ are sets of densities and conditional densities on $\mathrm{Supp}(X)$ and $\mathrm{Supp}(X, Y)$ such that, for $c_X, c_{X,Y} \in (0, \infty)$, for all $f_X \in \mathcal{E}$, $\|1/f_X\|_{L^\infty(\mathrm{Supp}(X))} \leq c_X$, $\|f_X\|_{L^\infty(\mathrm{Supp}(X))} \leq C_X$, and there exists a strict subset $\mathcal{S}$ of $\mathrm{Supp}(X, Y)$ such that, for all $f_{Y|X} \in \mathcal{E}'$, $\|1/f_{Y|X}\|_{L^\infty(\mathcal{S})} \leq c_{X,Y}$; For $(v(n_0, \mathcal{E}))_{n_0 \in \mathbb{N}} \in (0,1)^{\mathbb{N}}$ and $(v(n_1, \mathcal{E}'))_{n_1 \in \mathbb{N}} \in (0,1)^{\mathbb{N}}$ which tend to 0, we have*

$$\frac{1}{v(n_0, \mathcal{E})} \sup_{f_X \in \mathcal{E}} \left\| \widehat{f}_X - f_X \right\|^2_{L^\infty(\mathrm{Supp}(X))} = O_p(1), \tag{64}$$

$$\frac{1}{v(n_1, \mathcal{E}')} \sup_{f_{Y|X} \in \mathcal{E}'} \left\| \widehat{f}_{Y|X} - f_{Y|X} \right\|^2_{L^\infty(\mathcal{S})} = O_p(1). \tag{65}$$

Giné and Nickl (2016); Tsybakov (2008) give examples of estimators for f_X and $f_{Y|X}$, $\mathcal{E}$, $\mathcal{E}'$ rates (64) and (65). Define $\widehat{f}_X^{\delta} := \widehat{f}_X \vee \sqrt{\delta(n_0)}$ and $\widehat{f}_{Y|X}^{\delta} := \widehat{f}_{Y|X} \vee \sqrt{\delta(n_1)}$, where δ is a trimming factor converging to zero. To deal with the statistical problem, I use

$$\widehat{\partial_l F_{Y|X}}^{j_0}(\star|\cdot) := \sum_{|k|_\infty \leq j_0} \widehat{d}_k(\star) \partial_l L_k(\cdot), \tag{66}$$

where, for all $y \in \mathcal{S}_Y$,

$$\widehat{d}_k(y) := \frac{1}{n}\sum_{i=1}^{n} \frac{\mathbb{1}\{Y_i \leq y\}}{\widehat{f}_X^{\delta}(X_i)} L_k(X_i) \tag{67}$$

and replace $f_{Y|X}$ by $\widehat{f}_{Y|X}^{\delta}$ in (13)-(12).

L_μ^2 risk. In this context where f_X and $f_{Y|X}$ are estimated, I use the L_μ^2 risk on $\mathcal{S}$, which is defined in Assumption (Est.3), for $k = 1, \ldots, p+1$,

$$\mathcal{R}_{n_0,n_1}\left(\widehat{\mathrm{PE}}_k, PE_k\right) := \mathbb{E}\left[\left\|\widehat{\mathrm{PE}}_k - \mathrm{PE}_k\right\|_{L^2(\mathcal{S})} \middle| \mathcal{P}_{n_0}, \mathcal{P}_{n_1}\right]$$

and we use $n_e = n \wedge \lfloor(\delta(n_0)/v(n_0,\mathcal{E}))\rfloor \wedge \lfloor(\delta(n_1)\delta(n_0)/v(n_1,\mathcal{E}'))\rfloor$ for the sample size required for an ideal estimator where f_X and $f_{Y|X}$ are known to achieve the rate of the plug-in estimator. Instead of (29), the upper bounds of Proposition 6 in this context take the form, for $k = 1, \ldots, p+1$,

$$\frac{1}{r(n_e)} \sup_{\substack{f_\Gamma \in \mathcal{H}^\sigma(l) \\ f_X \in \mathcal{E},\ f_{Y|X} \in \mathcal{E}'}} \mathcal{R}_{n_0,n_1}\left(\widehat{\mathrm{PE}}_k^{j_0}, \mathrm{PE}_k\right) = O_p(1), \tag{68}$$

and in Propostion 6, n is replaced by n_e.

Proposition 8 *There exists a constant C_0 such that for all $f \in L^2(\mathbb{R}^{p+1})$ compactly supported in $[-g_0, g_0]^{p+1}$ and with $\sigma > (p+1)/2$,*

$$\int_{\mathrm{Supp}(X)} \int_{\mathbb{R}} (1 \vee |t|)^{2\sigma+p} \left|\mathcal{F}[f](t(1,x))\right|^2 dt dx \leq C_0 \int_{\mathbb{R}^{p+1}} (1 \vee |\xi|_2)^{2\sigma} \left|\mathcal{F}[f](\xi)\right|^2 d\xi.$$

Proof of Proposition 8. I borrow arguments from the proof of Theorem 4.6 in Hahn and Quinto (1985), without using the Radon transform. On the set $\mathrm{Supp}(X) \times \mathbb{R} \setminus [-1,1]$, we use the bijective change of variable $F(t,x) = (1, tx_1, \ldots, tx_p) = \xi \in V$ with V a truncated cone in $\mathbb{R}^{p+1}$ and that for $|t| \geq 1$, $(1 \vee |t|)^p \leq 2^{p/2}|t|^p$ for the first equality

$$\begin{aligned}\int_{\mathrm{Supp}(X)} \int_{\mathbb{R}} (1 \vee |t|)^{2\sigma+p} \left|\mathcal{F}[f](t(1,x))\right|^2 dt dx &\leq 2^{p/2} \int_V (1 \vee |\xi_p|)^{2\sigma} \left|\mathcal{F}[f](\xi)\right|^2 d\xi \\ &\leq 2^{p/2} \int_{\mathbb{R}^{p+1}} (1 \vee |\xi|)^{2\sigma} \left|\mathcal{F}[f](\xi)\right|^2 d\xi.\end{aligned}$$

Then, for all $(x,t) \in \mathrm{Supp}(X) \times [-1,1]$, using the compact support of f and Parseval's identity for the second equality,

$$\begin{aligned}
&|\mathcal{F}[f](t(x,1))| \\
&= \left| \int_{\mathbb{R}^{p+1}} \mathbb{1}\{g \in \mathcal{S}_\Gamma\} e^{i(t(1,x))^\top g} f(g) dg \right| \\
&= \left| \int_{\mathbb{R}^{p+1}} \mathcal{F}\left[\mathbb{1}\{\cdot \in \mathcal{S}_\Gamma\} e^{i(t(1,x))^\top \cdot} \right](\xi) \mathcal{F}[f](\xi) d\xi \right| \\
&\leq \int_{\mathbb{R}^{p+1}} \left| \mathcal{F}\left[\mathbb{1}\{\cdot \in \mathcal{S}_\Gamma\} e^{i(t(1,x))^\top \cdot} \right](\xi) \right|^2 (1 \vee |\xi|)^{-2\sigma} d\xi \int_{\mathbb{R}^{p+1}} (1 \vee |\xi|)^{2\sigma} \left| \mathcal{F}[f](\xi) \right|^2 d\xi.
\end{aligned}$$

I conclude using that

$$\begin{aligned}
&\int_{\mathbb{R}^{p+1}} \left| \mathcal{F}\left[\mathbb{1}\{\cdot \in \mathcal{S}_\Gamma\} e^{i(t(1,x))\cdot} \right](\xi) \right|^2 (1 \vee |\xi|)^{-2\sigma} d\xi \\
&= |g_0|^{p+1} \int_{\mathbb{R}^{p+1}} \frac{\prod_{i=2}^{p+1} \left| \mathrm{sinc}\left(\xi_i + t g_0 x_i\right) \right|^2 \left| \mathrm{sinc}\left(\xi_1 + t g_0\right) \right|^2}{2^{-2(p+1)} (1 \vee |\xi|)^{2\sigma}} d\xi,
\end{aligned}$$

which is finite for $\sigma > (p+1)/2 \geq 1$ and that $\mathrm{Supp}(X) \times [-1,1]$ has finite measure. □

Use that, for all $m \in \mathbb{N}_0$, from (1) in Lohöfer (1998) and (21.4.3) in Poularikas (2018)

$$\forall x \in (-1,1), \quad |L_m(x)| \leq \frac{2}{\pi} \frac{1}{(1-x^2)^{1/4}} \text{ and } |L'_m(x)| \leq \frac{2}{\sqrt{\pi}} \frac{\sqrt{m(m+1/2)}}{1-x^2}, \tag{69}$$

and from (21.1.7) in Poularikas (2018) and Markov's inequality for polynomials (see, *e.g.*, Theorem 5.1.8 in Borwein and Erdélyi, 1995)

$$\forall x \in [-1,1], \quad |L_m(x)| \leq \sqrt{m + \frac{1}{2}} \text{ and } |L'_m(x)| \leq m^2 \sqrt{m + \frac{1}{2}}. \tag{70}$$

In the remaining, $\mathcal{E}$ and $\mathcal{E}'$ are classes of densities and conditional densities, $f_X \in \mathcal{E}$, $f_{Y|X} \in \mathcal{E}'$, and $\eta, M > 0$. Denote also by $\Delta_{f,0} := 1/\widehat{f}^{\delta}_X - 1/f_X$, $\Delta_{f,1} := 1/\widehat{f}^{\delta}_{Y|X} - 1/f_{Y|X}$,

$$Z_{n_0} := \sup_{f_X \in \mathcal{E}} \left\| \Delta_{f,0} f_X \right\|^2_{L^\infty(\mathrm{Supp}(X))}, \quad Z_{n_1} := \sup_{f_{Y|X} \in \mathcal{E}'} \left\| \Delta_{f,1} f_{Y|X} \right\|^2_{L^\infty(\mathbb{S}_{Y,X})}.$$

By Lemma A.3 in Gaillac and Gautier (2022), there exists $M_{\mathcal{E},\eta,0}$ and $M_{\mathcal{E}',\eta,1}$ such that, for all $n_0, n_1 \in \mathbb{N}$, $\mathbb{P}\left(E\left(\mathcal{P}_{n_1}, \mathcal{E}', \eta\right)\right) \geq 1 - \eta/2$ and $\mathbb{P}\left(E\left(\mathcal{P}_{n_1}, \mathcal{E}, \eta\right)\right) \geq 1 - \eta/2$ where

$$E\left(\mathcal{P}_{n_0}, \mathcal{E}, \eta\right) := \left\{ Z_{n_0} \leq \frac{M_{\mathcal{E},\eta,0} v(n_0, \mathcal{E})}{\delta(n_0)} \right\}$$

and $E(\mathcal{P}_{n_1}, \mathcal{E}', \eta) := \{Z_{n_1} \leq M_{\mathcal{E}',\eta,1} v(n_1, \mathcal{E}')/\delta(n_1)\}$. I work on $E(\mathcal{P}_{n_0}, \mathcal{P}_{n_1}, \mathcal{E}, \mathcal{E}', \eta) := E(\mathcal{P}_{n_0}, \mathcal{E}, \eta) \cap E(\mathcal{P}_{n_1}, \mathcal{E}', \eta)$, hence using independence $\mathbb{P}(E(\mathcal{P}_{n_0}, \mathcal{P}_{n_1}, \mathcal{E}, \mathcal{E}', \eta)) \geq 1 - \eta$, and use $M_{\mathcal{E},\mathcal{E}',\eta} := M_{\mathcal{E},\eta,0} \vee M_{\mathcal{E}',\eta,1}$.

All expectations are conditional on $\mathcal{P}_{n_0}$ and $\mathcal{P}_{n_1}$ when f_X and $f_{Y|X}$ are unknown and we rely on $\mathcal{P}_{n_0}$ and $\mathcal{P}_{n_1}$ to estimate it. We remove the conditioning in the notations for simplicity. Denote, for all $k \in \mathbb{N}_0^p$, by $\widetilde{d}_k$ the quantities defined as in (36) replacing $\widehat{f}_X^{\delta}$ by f_X. Denote by $\widetilde{\partial_{x_l} F}_{Y|X}^{j_0}$ the estimator $\widehat{\partial_{x_l} F}_{Y|X}^{j_0}$ where $\widehat{d}_k$ is replaced by $\widetilde{d}_k$. Denote also by $\widetilde{\mathrm{PE}}^{j_0}$ the estimator $\widehat{\mathrm{PE}}^{j_0}$ where $\widehat{f}_{Y|X}^{\delta}$ is replaced by $f_{Y|X}$.

Lemma 2 *For all $k \in \mathbb{N}_0^{p+1}$, and $y \in \mathcal{S}_Y$, we have $\mathbb{E}\left[\widetilde{d}_k(y)\right] = d_k(y)$, and*

$$\mathbb{E}\left[\left|\widetilde{d}_k(y) - d_k(y)\right|^2\right] \leq \frac{c_X}{n}.$$

Proof of Lemma 2. Let $k \in \mathbb{N}_0^{p+1}$, and $y \in \mathcal{S}_Y$. We have, using integration by part and that L_k is compactly supported,

$$\mathbb{E}\left[\widetilde{d}_k(y)\right] = \mathbb{E}\left[\frac{\mathbb{1}\{Y_i \leq y\}}{f_X(X_i)} L_k(X_i)\right] = \int_{\mathrm{Supp}(X)} \mathbb{E}\left[\mathbb{1}\{Y \leq y\} | X = x\right] L_k(x)\, dx$$

and, using that $\mathcal{B}$ is an orthonormal basis of $L^2(\mathrm{Supp}(X))$, this yields

$$\mathbb{E}\left[\left|\widetilde{d}_k(y) - d_k(y)\right|^2\right] \leq \frac{1}{n} \int_{\mathrm{Supp}(X)} \frac{1}{f_X(x)} |L_k(x)|^2\, dx \leq \frac{c_X}{n}. \qquad \square \tag{71}$$

Proof of Proposition 6. Let $(x, y) \in \mathcal{S}$, we use

$$R_{0,l}^{j_0} : (x, y) \mapsto \left(\widehat{\partial_{x_l} F}_{Y|X}^{j_0} - \widetilde{\partial_{x_l} F}_{Y|X}^{j_0}\right)(x, y) \tag{72}$$

$$R_{1,l}^{j_0} : (x, y) \mapsto \left(\widetilde{\partial_{x_l} F}_{Y|X}^{j_0} - \partial_{x_l} F_{Y|X}^{j_0}\right)(x, y) \tag{73}$$

$$R_{2,l}^{j_0} : (x, y) \mapsto \left(\partial_{x_l} F_{Y|X}^{j_0} - \partial_{x_l} F_{Y|X}\right)(x, y). \tag{74}$$

In the following I prove the result for m_1, as the cases PE_k $k \geq 1$ can be directly deduced from it. Using the triangular inequality and the convexity of $x \mapsto x^2$, we obtain

$$\begin{aligned}
&\left\|\widehat{\mathrm{PE}}_1^{j_0} - \mathrm{PE}_1\right\|_{L^2_\mu(\mathcal{S})}^2 \\
&\leq \left\|\widehat{\mathrm{PE}}_1^{j_0} - \widetilde{\mathrm{PE}}_1^{j_0}\right\|_{L^2_\mu(\mathcal{S})}^2 + \left\|\widetilde{\mathrm{PE}}_1^{j_0} - PE_1\right\|_{L^2_\mu(\mathcal{S})}^2 \\
&\leq p \sum_{l=1}^{p} \|x_l\|^2_{L^\infty(\mathrm{Supp}(X))} \left(Z_{n_1} \left\|\widehat{\partial_{x_l} F}_{Y|X}^{j_0}\right\|_{L^2_\mu(\mathcal{S})}^2 + c_{X,Y}^2 \left\|\widehat{\partial_{x_l} F}_{Y|X}^{j_0} - \partial_{x_l} F_{Y|X}\right\|_{L^2_\mu(\mathcal{S})}^2 \right).
\end{aligned}$$

Then, using the convexity of $x \mapsto x^2$, the Cauchy-Schwarz inequality and that $(\Omega_{l,k})$ is an orthonormal system of $L^2(\mathrm{Supp}(X))$ for the first inequality, that (L_k) is an orthonormal system of $L^2(\mathrm{Supp}(X))$ for the second one, and (46) for the last one, we obtain

$$\begin{aligned}\mathbb{E}\left[\left\|\widehat{\partial_{x_l} F}^{j_0}_{Y|X}\right\|^2_{L^2_\mu(\mathcal{S})}\right] &\leq \sum_{|k|_\infty \leq j_0} k_l(k_l+1)\int_{\mathcal{S}_Y} \mathbb{E}\left[\left|\widehat{d}_k(y)\right|^2\right] dy \\ &\leq \sum_{|k|_\infty \leq j_0} k_l(k_l+1)\frac{|\mathcal{S}_Y| C_X}{\delta(n_0)} \\ &\leq \frac{|\mathcal{S}_Y|(j_0+1)^{p+2} C_X}{\delta(n_0)}. \end{aligned} \tag{75}$$

Using $C_{1,l} := 3pc^2_{X,Y}\|x_l\|^2_{L^\infty(\mathrm{Supp}(X))}$ and the convexity of $x \mapsto x^2$ thus yield

$$\left\|\widehat{\mathrm{PE}}_1^{j_0} - \mathrm{PE}_1\right\|^2_{L^2_\mu(\mathcal{S})} \leq \widetilde{C}_0 \frac{Z_{n_1}(j_0+1)^{p+2}}{\delta(n_0)} + \sum_{l=1}^{p} C_{1,l} \sum_{j=0}^{2} \int_{\mathcal{S}} \left|R^{j_0}_{j,l}(x,y)\right|^2 \mu(x,y)dydx, \tag{76}$$

where $\widetilde{C}_0 := |\mathcal{S}_Y| pC_X \sum_{l=1}^p \|x_l\|^2_{L^\infty(\mathrm{Supp}(X))}$.

Term $R_{0,l}$. We obtain, using the Cauchy-Schwarz inequality and that $(\Omega_{l,k})_{k\in\mathbb{N}_0^p}$ is an orthonormal system of $L^2(\mathrm{Supp}(X))$ for the first display and (46) for the second display, for all $l = 1, \ldots, p$,

$$\begin{aligned}\mathbb{E}\left[\|R^{j_0}_{0,l}\|^2_{L^2_\mu(\mathcal{S})}\right] &\leq \sup_{y\in\mathcal{S}_Y} \sum_{|k|_\infty \leq j_0} \mathbb{E}\left[\left|\widehat{d}_k(y) - \widetilde{d}_k(y)\right|^2\right] k_l(k_l+1) \\ &\leq Z_{n_0} C_X (j_0+1)^{p+2}. \end{aligned} \tag{77}$$

Term $R_{1,l}$. We obtain, for all $l = 1, \ldots, p$, using the Cauchy-Schwarz inequality, that $(\Omega_{l,k})$ is an orthonormal system of $L^2(\mathrm{Supp}(X))$ for the first display, Lemma 2 and (46) for the third display,

$$\begin{aligned}\int_{\mathcal{S}} \mathbb{E}\left[\left|R^{j_0}_{1,l}(x,y)\right|^2\right]\mu(x)dydx &\leq \int_{\mathcal{S}_Y} \sum_{|k|_\infty \leq j_0} \mathbb{E}\left[\left|\widetilde{d}_k(y) - d_k(y)\right|^2\right] k_l(k_l+1)dy \\ &\leq \frac{|\mathcal{S}_Y| c_X (j_0+1)^{p+2}}{n} \end{aligned} \tag{78}$$

Term $R_{2,l}$. We have, using that $(\Omega_{l,k})_{k\in\mathbb{N}_0^p}$ is an orthonormal system of $L^2(\mathrm{Supp}(X))$

for the second display,

$$\int_{\mathcal{S}} \left|R_{2,l}^{j_0}(x,y)\right|^2 \mu(x)dydx \leq \int_{\mathcal{S}} \left| \sum_{j=j_0+1}^{\infty} \sum_{|k|_\infty=j} d_k(y)\sqrt{k_l(k_l+1)}\Omega_{l,k}(x) \right|^2 dxdy$$
$$\leq \int_{\mathcal{S}_Y} \sum_{j=j_0+1}^{\infty} \sum_{|k|_\infty=j} |d_k(y)|^2 k_l(k_l+1)dy.$$

Using $\mathcal{S}_Y \subseteq [\underline{y}, \infty)$, and for the second equality that under Assumption 3,

$$F_{Y|X}(y|x) = \int_{\underline{y}}^{y} \mathcal{F}^{-1}\left[\mathcal{F}\left[f_\Gamma\right](\cdot(1,x))\right](v)dv, \tag{79}$$

and using the Cauchy-Schwarz inequality for the third display, we obtain

$$\begin{aligned}
|d_k(y)| &= \left| \int_{\mathrm{Supp}(X)} F_{Y|X}(y|x)L_k(x)dx \right| \\
&= \frac{1}{2\pi} \left| \int_{\mathrm{Supp}(X)} \int_{\underline{y}}^{y} \int_{\mathbb{R}} e^{-itv} \mathcal{F}\left[f_\Gamma\right](t(1,x))L_k(x)dxdvdt \right| \\
&\leq \frac{|y-\underline{y}|}{2\pi} \int_{\mathbb{R}} \left| \mathrm{sinc}\left(\frac{t(y-\underline{y})}{2}\right) \right| \left| \int_{\mathrm{Supp}(X)} \mathcal{F}\left[f_\Gamma\right](t(1,x))L_k(x)dx \right| dt \\
&\leq \frac{\sqrt{2|y-\underline{y}|}}{\pi} \left(\int_{\mathbb{R}} |c_k(t)|^2 dt \right)^{1/2},
\end{aligned} \tag{80}$$

where $c_k(t) := \int_{\mathrm{Supp}(X)} \mathcal{F}\left[f_\Gamma\right](t(1,x))L_k(x)dx$. Thus, we have

$$\begin{aligned}
\left\|R_{2,l}^{j_0}\right\|_{L^2_\mu(\mathcal{S})}^2 &\leq 2\int_{\mathcal{S}_Y} \sum_{j\geq j_0+1} \sum_{|k|_\infty=j} |d_k(y)|^2 k_l^2 dy \\
&\leq \int_{\mathcal{S}_Y} \frac{4|y-\underline{y}|}{\pi^2} dy \sum_{j\geq j_0+1} \sum_{|k|_\infty=j} \int_{\mathbb{R}} |c_k(t)|^2 k_l^2 dt \\
&\leq \frac{8\|y^2\|_{L^\infty(\mathcal{S}_Y)} H_l}{\pi^2 (j_0+1)^{2s}}
\end{aligned} \tag{81}$$

where

$$H_l = \sum_{j\in\mathbb{N}} \sum_{|k|_\infty=j} \int_{\mathbb{R}} |c_k(t)|^2 (j+1)^{2s} k_l^2 dt.$$

Let us now prove that H_l is finite under the smoothness assumption. For all $j \in \mathbb{N}_0$, $k \in \mathbb{N}_0^p$, denote by $H_{1,l}(j,k) := \int_{|t|>j} |c_k(t)|^2 j^{2s} k_l^2 dt$ and $H_{2,l}(j,k) := \int_{|t|\leq j} |c_k(t)|^2 j^{2s} k_l^2 dt$.

Using that $(L_k)_{k\in\mathbb{N}_0^p}$ are orthonormal on $L^2(\mathrm{Supp}(X))$ for the second inequality and Assumption 10 and Proposition 8 for the third one

$$\begin{aligned}\sum_{j\in\mathbb{N}}\sum_{|k|_\infty=j} H_{1,l}(j,k)j^{2s}k_l^2 &\leq \sum_{j\in\mathbb{N}}\sum_{|k|_\infty=j}\int_{|t|>j}|c_k(t)|^2\left(|t|\vee 1\right)^{2s+2}dt\\ &\leq \int_{\mathbb{R}}\int_{\mathrm{Supp}(X)}|\mathcal{F}\left[f_\Gamma\right](t(1,x))|^2(|t|\vee 1)^{2s+2}dxdt\\ &\leq l^2C_0. \end{aligned} \tag{82}$$

Then, using for the fourth equality that $\mathcal{E}xt\left[L_k\right]$ has compact support in $\mathrm{Supp}(X)$,

$$\begin{aligned}\left|\int_{\mathrm{Supp}(X)}\mathcal{F}\left[f_\Gamma\right](t(1,x))L_k(x)dx\right| &= \left|\int_{\mathrm{Supp}(X)}\int_{\mathcal{S}_\Gamma}e^{it(1,x)^\top g}f_\Gamma(g)L_k(x)dgdx\right|\\ &= \left|\int_{\mathcal{S}_\Gamma}f_\Gamma(g)\overline{e^{-itg_1}\int_{\mathrm{Supp}(X)}e^{-itx^\top g_{-1}}L_k(x)dx}dg\right|\\ &= 2\pi\left|\int_{\mathcal{S}_\Gamma}f_\Gamma(g)\overline{e^{-itg_1}\mathcal{F}^{-1}\left[\mathcal{E}xt\left[L_k\right]\right](tg_{-1})}dg\right|\\ &\leq 2\pi\sup_{g\in[-g_0,g_0]^p}\left|\mathcal{F}^{-1}\left[\mathcal{E}xt\left[L_k\right]\right](tg)\right|. \end{aligned} \tag{83}$$

Denote by $\widetilde{L}_k$ the normalised Legendre polynomials on $L^2([-1,1])$. Using that, for all $c\neq 0$ and $g\in[-g_0,g_0]^p$,

$$\begin{aligned}|\mathcal{F}^{-1}\left[\mathcal{E}xt\left(L_k\right)\right](cg)| &= x_0^p|\mathcal{F}^{-1}\left[\mathcal{E}xt\left(\widetilde{L}_k\right)\right](cx_0g)|\\ &\leq \left(\frac{ex_0}{2}\right)^p\prod_{r=1}^p\left(\frac{ex_0|c|}{2(k_r+1/2)}\right)^{k_r}|g_r|^{k_r},\end{aligned}$$

(see p15 in Gaillac and Gautier (2021a)), we obtain, for all $|t|\leq j$, $|k|_\infty\geq j_0+1$ and $g\in[-g_0,g_0]^p$,

$$|\mathcal{F}^{-1}\left[\mathcal{E}xt\left(L_k\right)\right](tg)| \leq \left(\frac{\omega}{g_0}\right)^p\prod_{l=1}^p\left(\frac{\omega j}{k_l+1/2}\right)^{k_l}. \tag{84}$$

Then, up to re-indexing we have

$$\begin{aligned}\prod_{r=1}^p\left(\frac{\omega j}{k_r+1/2}\right)^{k_l} &\leq \omega^j\prod_{r=1}^{p-1}\left(\frac{\omega j}{k_r+1/2}\right)\\ &\leq \left(\omega e^{\omega(p-1)/e}\right)^j.\end{aligned}$$

Then, we have for all k such that $|k|_\infty = j$ and using (46),

$$\sum_{|k|_\infty=j} H_{2,l}(j,k) j^{2s} k_l^2 \leq \left(\frac{\sqrt{2}\omega}{g_0}\right)^{2p} \left(\omega e^{\omega(p-1)/e}\right)^{2j} j^{2s+2+p}. \tag{85}$$

Thus, using that $\omega e^{\omega(p-1)/e} < 1$, we obtain that $H_{2,l}$ is bounded, thus H_l is finite. Hence, using $C_8 := \sum_{l=1}^p C_{1,l}$ and

$$C_{10} := \frac{8\|y^2\|_{L^\infty(\mathcal{S}_Y)} H_l}{\pi^2},$$

we obtain,

$$\begin{aligned}
&\mathbb{E}\left[\left\|\widehat{PE}_1^{j_0} - PE_1\right\|_{L^2(\mathcal{S})}^2\right] \\
&\leq \widetilde{C}_0 \frac{Z_{n_1}(j_0+1)^{p+2}}{\delta(n_0)} + C_8\left(2Z_{n_0}C_X(j_0+1)^{p+2} + 2c_X\frac{(j_0+1)^{p+2}}{n} + \frac{C_{10}}{(j_0+1)^{2s}}\right) \\
&\leq \widetilde{C}_0 M_{\mathcal{E}',\eta,1}\frac{v(n_1,\mathcal{E}')(j_0+1)^{p+2}}{\delta(n_0)\delta(n_1)} + C_8 2Z_{n_0}C_X M_{\mathcal{E},\eta,0}\frac{v(n_0,\mathcal{E})(j_0+1)^{p+2}}{\delta(n_0)} \\
&\quad + C_8\left(2c_X\frac{(j_0+1)^{p+2}}{n} + \frac{C_{10}}{(j_0+1)^{2s}}\right).
\end{aligned}$$

Using $\widetilde{j} = n_e^{1/(2s+p+2)}$ and $j_0 \geq \widetilde{j} - 1$ yields the result. □

Proofs of Proposition 7 Let us focus on the proof for $k = 1$ are the other ones can be deduced directly from it. Denoting by $\mathbb{P}_{\Gamma,j}$ the law of $\mathbb{P}_\Gamma$, $PE_{1,j}(x,y) = \mathbb{E}[\Gamma_1 | X = x, Y = y]$, the associated functions of interest, and by $\mathbb{P}_{j,G}$ the law of an i.i.d $(X_i, Y_i)_{i=1}^n$ sample of size n, for $j = 0, \ldots, K$, $K \geq 1$, and use

$$\inf_{\widehat{\mathrm{PE}}_1} \sup_{\mathbb{P}_\Gamma \in \mathcal{H}^{s+1}(l)} \mathbb{E}\left[\left\|\widehat{\mathrm{PE}}_1 - \mathrm{PE}_1\right\|_{L^2(\mathcal{S})}\right] \geq \inf_{\widehat{\mathrm{PE}}_1} \sup_{\mathbb{P}_{\Gamma,j} \in \mathcal{H}^s(l), j=0,\ldots,K} \mathbb{E}\left[\left\|\widehat{\mathrm{PE}}_1 - \mathrm{PE}_{1,j}\right\|_{L^2(\mathcal{S})}\right]$$

and Theorem 2.6 , (2.5), and (2.9) in Tsybakov (2000) that we now recall.

Proposition 9 (Theorem 2.6 in Tsybakov (2000)) *Assume that $\mathcal{H}^{s+1}(l)$ contains $\{\mathbb{P}_{\Gamma,j}, j = 0, \ldots, K\}$, $K \geq 1$, which satisfy:*

1. *$\|PE_{1,j} - PE_{1,k}\|_{L^2(\mathcal{S})} \geq 2r(n)$, for all $0 \leq j < k \leq K$;*
2. *for all $j = 1, \ldots, K$,*

$$\frac{1}{K}\sum_{j=1}^K \chi^2\left(\mathbb{P}_{\Gamma,j}, \mathbb{P}_{\Gamma,0}\right) \leq \xi K; \tag{86}$$

Then, we have

$$\frac{1}{r(n)}\inf_{\widehat{PE}_1}\sup_{\mathbb{P}_\Gamma\in\mathcal{H}^{s+1}(l)}\mathbb{E}\left[\left\|\widehat{PE}_1-PE_1\right\|_{L^2(\mathcal{S})}\right]\geq\frac{1}{2}\left(1-\xi-\frac{1}{K}\right).$$

Before proceeding with the proof, we need to introduce some vaguelets, I refer to Gaillac (2021) for more details.

Consider here the case where $\text{Supp}(X)$ is a square $\text{Supp}(X)=\prod_{l=1}^p[\widetilde{x}_l,\widetilde{x}_l+x_0]$, where $\widetilde{x}\in\mathbb{R}^p$ and $x_0>0$. In this case, I use the boundary corrected wavelets introduced in Cohen et al. (1993) (see, *e.g.*, Section 4.3.5 in Giné and Nickl, 2016). Let $J,N\in\mathbb{N}$, $2^J\geq N$ and consider the standard 2^J-2N Daubechies wavelets $\phi_{J,k}=2^{J/2}\phi(2^J\cdot-k)$, $k\in\mathbb{Z}$ supported in the interior of $[0,1]$, the N left-edge basis functions $\phi_{J,k}^{\text{left}}$, and the right-edge basis functions $\phi_{J,k}^{\text{right}}$ introduced in Cohen et al. (1993) that are obtained from transformations (*e.g.* Gram-Schmidt orthonormalisation) of the standard wavelets. Together, they form an orthonormal system of $L^2([0,1])$ which I denote by $\left\{\phi_{J,k}^{bc},k=0,\dots,2^J-1\right\}$. Then using the construction of Section 4.3.6 in Giné and Nickl (2016), I introduce, for the purpose of this proof, for $k\in\Lambda_j:=\{k:|k|_\infty\leq 2^J-1\}$,

$$\Phi_{1,J,k}:=\frac{1}{x_0^{p/2}}\phi_{J,k_1}^{bc}\left(\frac{\cdot_1-\widetilde{x}_1}{x_0}\right)\quad\text{and}\quad\Omega_{1,J,k}=\partial_1\Phi_{1,J,k}/2^J.$$

A direct consequence of the vaguelets theory in Section 5 and condition (C) in Cai (2002), is that there exist constants $A>a>0$, which depend on $\text{Supp}(X)$, such that, for every sequence $(d_{j,k})$,

$$a\left\|(d_{j,k})\right\|_{l^2}\leq\left\|\sum_{j\geq J}\sum_{k\in\Lambda_j}d_{j,k}\Omega_{1,j,k}\right\|_{L^2(\text{Supp}(X))}\leq A\left\|(d_{j,k})\right\|_{l^2}. \tag{87}$$

I consider here the following distributions:

- $\mathbb{P}_{\Gamma,0}=\bigotimes_{l=1}^{p+1}\mathbb{P}_{\Gamma_l,0}$, $\mathbb{P}_{\Gamma_1,0}=\mathbb{P}_0$, and $\mathbb{P}_{\Gamma_2,0}=\cdots=\mathbb{P}_{\Gamma_{p+1},0}=0$. This yields for all $\mathbb{P}_Y=\mathbb{P}_0$ hence $\text{PE}_{1,0}(x,y)=y$;

- $K=2$ and $\mathbb{P}_{\Gamma,1}$ is the compactly supported function in $[0,1]^{p+1}$ such that, for all $t\in\mathbb{R}$, $x\in\text{Supp}(X)$,

$$\mathcal{F}\left[\mathbb{P}_{\Gamma,1}\right](t(1,x))=\gamma(t)\sum_{k\in\overline{\Lambda}_{j_0}}\Phi_{1,j_0,k}(x)+\mathcal{F}\left[\mathbb{P}_0\right](t), \tag{88}$$

where $\gamma(0) = 0$, $\overline{\Lambda}_{j_0} \subset \Lambda_{j_0}$ such that the support of all functions $(\Phi_{1,j_0,k})_{k\in\overline{\Lambda}_{j_0}}$ is a subset of $\mathbb{S}_{X_1}$. We have, using (13), on $\mathcal{S}$,

$$\mathrm{PE}_1(x,y) = y + \frac{x_1}{f_{Y|X}(y|x)} \int_{-\infty}^{y} \mathcal{F}^{-1}\left[\gamma(\cdot)\right](v)dv \sum_{k\in\overline{\Lambda}_{j_0}} 2^{j_0}\Omega_{1,j_0,k}\left(x\right). \tag{89}$$

From the end of page 724 in Rullgård and Quinto (2010) and arguments from Proposition 8, there exists a constant $\widetilde{C}_0$ depending only on p such that for all $f \in L^2(\mathbb{R}^{p+1})$ compactly supported in $[-1,1]^{p+1}$ and with $\sigma > (p+1)/2$,

$$\int_{\mathrm{Supp}(X)} \int_{\mathbb{R}} \left(1 \vee |t|\right)^{2s+2} \left|\mathcal{F}[f](t(1,x))\right|^2 dtdx \geq \frac{1}{\widetilde{C}_0} \int_{\mathbb{R}^{p+1}} \left(1 \vee |\xi|_2\right)^{2\sigma} \left|\mathcal{F}[f](\xi)\right|^2 d\xi.$$

Thus, using that $(\Phi_{j,k})_{j\geq J, k\in\Lambda_j}$ is an orthonormal system of $L^2(\mathrm{Supp}(X))$, $\mathcal{H}^\sigma(l)$ contains $\{\mathbb{P}_{\Gamma,j}, j = 0,1\}$, if

$$\int_{\mathbb{R}} \left(1 \vee |t|\right)^{2s+2} \gamma(t)^2 dt + \int_{\mathrm{Supp}(X)} \int_{\mathbb{R}} \left(1 \vee |t|\right)^{2s+2} \left|\mathcal{F}\left[\mathbb{P}_0\right](t)\right|^2 dt \leq \frac{l^2}{\widetilde{C}_0}. \tag{90}$$

Then, using (89) and $|\overline{\Lambda}_{j_0}| \geq c_0 2^{j_0 p}$, for $r = 1, 2$,

$$\begin{aligned}
&\left\|m_{1,1} - m_{1,0}\right\|^2_{L^2(\mathrm{Supp}(X,Y))} \qquad (91)\\
&= \left\| \frac{x_1}{f^1_{Y|X}(y|x)} \int_{-\infty}^{y} \mathcal{F}^{-1}\left[\gamma(\cdot)\right](v)dv \sum_{k\in\overline{\Lambda}_{j_0}} 2^{j_0}\Omega_{1,j_0,k}\left(x\right) \right\|^2_{L^2(\mathrm{Supp}(X,Y))}\\
&\geq ac_0 2^{j_0(p+2)} \inf_{(x,y)\in\mathrm{Supp}(X,Y)} \left|\frac{x_1}{f_{Y|X}(y|x)}\right|^2 \int_{\mathrm{Supp}(Y)} \left|\int_{-\infty}^{y} \mathcal{F}^{-1}\left[\gamma(\cdot)\right](v)dv\right|^2 dy. \qquad (92)
\end{aligned}$$

Using Step 3. in Gaillac and Gautier (2022): $\chi_2(\mathbb{P}_{k,n}, \mathbb{P}_{0,n}) \leq en\chi_2\left(\mathbb{P}_k, \mathbb{P}_0\right)$, where

$$\chi_2\left(\mathbb{P}_k, \mathbb{P}_0\right) = \int_{\mathrm{Supp}(X,Y)} \frac{f_X(x)\left(f^0_{Y|X}(y|x) - f^k_{Y|X}(y|x)\right)^2}{f^0_{Y|X}(y|x)} dxdy.$$

Using that $f^0_{Y|X}(y|x) = f^0_Y(y) \geq \inf_{y\in\mathrm{Supp}(Y)} f^0_Y(y) =: 1/c_Y > 0$ on $\mathrm{Supp}(Y)$, we have

$$\begin{aligned}
\chi_2\left(\mathbb{P}_1, \mathbb{P}_0\right) &\leq C_X c_Y \int_{\mathrm{Supp}(X,Y)} \left(f^0_{Y|X}(y|x) - f^k_{Y|X}(y|x)\right)^2 dxdy\\
&\leq C_X c_Y \int_{\mathrm{Supp}(X)} \int_{\mathbb{R}} \left|\mathcal{F}\left[\mathbb{P}_{\Gamma,k}\right](t(1,x))\right|^2 dxdt\\
&\leq C_X c_Y \int_{\mathrm{Supp}(X)} \left|\Phi_{j_0,k}\left(x\right)\right|^2 dx \int_{\mathbb{R}} \gamma(t)^2 dt\\
&= C_X c_Y \int_{\mathbb{R}} \gamma(t)^2 dt.
\end{aligned}$$

Hence, (86) is satisfied if

$$n \int_{\mathbb{R}} \gamma(t)^2 dt \leq \frac{\xi |\overline{\Lambda}_{j_0}|}{C_X c_Y e}. \tag{93}$$

Take, for all $t \in \mathbb{R}$,

$$\gamma(t) = \frac{\epsilon(1 \wedge |t/\tau|^\nu)}{(1 + (t/\tau)^{s+1})\tau^{s+p/2+3/2}(e \vee |t/\tau|)^{1/2}(\ln(e \vee t/\tau))^{1/2}},$$

with $\nu \geq 1/2$,

- $\tau = 2^{j_0}$ and j_0 such that $2^{j_0} \sim n^{1/(2s+p+2)}$
- ϵ such that

$$\epsilon \int_{\mathbb{R}} \frac{(1 \wedge |t|^{2\nu})}{(e \vee |t|)\ln(e \vee (t/\tau))} dt + \int_{\mathrm{Supp}(X)} \int_{\mathbb{R}} (1 \vee |t|)^{2s+2} \left|\mathcal{F}\left[\mathbb{P}_0\right](t)\right|^2 dt \leq \frac{l^2}{\widetilde{C}_0};$$

and $\epsilon^2 \leq \xi/(C_X c_Y e\,(1 + 1/(2s+p+2)))$ which ensures that

$$\begin{aligned} G \int_{\mathbb{R}} \gamma(t)^2 dt &\leq \int_{\mathbb{R}} \frac{n\epsilon^2}{\tau^{2s+p+3}(1 + (t/\tau)^{s+2})^2} dt \\ &\leq \left(1 + \frac{1}{2s+2}\right) n 2^{-j_0(2s+p+2)} \epsilon^2 \\ &\leq \frac{\xi \ln(n)}{C_X c_Y e} \end{aligned}$$

hence with $\ln(n) \leq K = |\overline{\Lambda}_{j_0}|$ that (93) is satisfied.

Finally, we have,

$$\left| \int_{\underline{y}}^{y} \mathcal{F}^{-1}\left[\gamma(\cdot)\right](v) dv \right| = \frac{\epsilon(y - \underline{y})}{2^{j_0(s+(p+2)/2)}} \int_{\mathbb{R}} \frac{\mathrm{sinc}\left(\tau t(y - \underline{y})/2\right)(1 \wedge |t|^\nu)}{(1 + |t|^{s+1})(e \vee |t|)^{1/2} \ln(e \vee |t|)^{1/2}} dt,$$

hence, using (92), we obtain

$$\|PE_{1,1} - PE_{1,0}\|^2_{L^2(\mathcal{S})} \geq C 2^{-2j_0 s},$$

where C is a constant independent of n, which yields the result using Proposition 9. We deduce the rate in $L^2_\mu(\mathcal{S})$ norm using that $\mu < 1$ hence $\|PE_{1,1} - PE_{1,0}\|_{L^2(\mathcal{S})} \geq \|PE_{1,1} - PE_{1,0}\|_{L^2_\mu(\mathcal{S})}$. □

A.2.2 G-modeling

I introduce, for every $F \in \mathcal{P}_2(\mathbb{R}^{p+1})$ and $(p, G) \in \Sigma_p \times \mathcal{P}_2(\mathcal{S})^\kappa$,

$$\mathcal{G}(F, p, G) = \sum_{j=1}^{\kappa} p_j W_2(F, (A^{-1/2} P_{(1,x_j)}^\top)_\# G_j).$$

For every $(x, y) \in \text{Supp}(X, Y)$ and $k \in \{1, \ldots, p+1\}$, we introduce the three functions:

$$\begin{aligned}
R_{k,x,y}: \quad & F \in \mathcal{P}(\text{Supp}(\Gamma)) \mapsto \frac{\int_{\mathcal{I}(x,y)} g_k dF(g)}{\int_{\mathcal{I}(x,y)} dF(g)} \\
\Phi: \quad & \Sigma_p \times \mathcal{P}_{a.c.}(\mathcal{S})^\kappa \to \mathcal{P}_2(\mathbb{R}^{p+1}) \\
& (p, G) \mapsto \underset{F \in \mathcal{P}_2(\mathbb{R}^{p+1})}{\text{argmin}} \, \mathcal{G}(F, p, G), \\
\underline{m}_{k,x,y}: \quad & (p, G) \in \Sigma_p \times \mathcal{P}_{a.c.}(\mathcal{S})^\kappa \mapsto R_{k,x,y}(A_\#^{-1/2} \Phi(p, G)).
\end{aligned}$$

Note that with the above restriction of Φ to $\mathcal{P}_{a.c.}(\mathcal{S})$, then it is well defined as there exists a unique solution to the Wasserstein barycenter problem if at least one the marginals is absolutely continuous (see Proposition 6 in Le Gouic and Loubes, 2017). In order to prove consistency, I rewrite the problem as a Wasserstein barycenter problem. Because A is invertible, using Proposition 3.1 in Delon et al. (2022), F_Γ^* is solution of (11) if and only if $F^* = A_\#^{1/2} F_\Gamma^*$ minimizes

$$\inf_{F \in \mathcal{P}_2(\mathbb{R}^{p+1})} \sum_{j=1}^{\kappa} \underline{p}_j W_2^2(F, G_{x_j}),$$

where $G_{x_j} = (A^{-1/2} P_{(1,x_j)}^\top)_\# F_{Y|X=x_j}$. We thus have, for every $(x, y) \in \text{Supp}(X, Y)$ and $k \in \{1, \ldots, p+1\}$,

$$\text{PE}_k(x, y) = \underline{m}_{k,x,y}(\underline{p}, F_{Y|X=x_1}, \ldots, F_{Y|X=x_\kappa}).$$

Proof of Theorem 2. Let $k = 1, \ldots, p+1$ and $(x, y) \in \text{Supp}(X, Y)$. Lemma 3 ensuring the continuity of the map $R_{x,y,k}(A_\#^{-1/2} \cdot)$ and the Proposition 6 in Le Gouic and Loubes (2017) ensuring the continuity of the unregularized Wasserstein barycenter map Φ, then the function $\underline{m}_{k,x,y}$ is continuous on $\Sigma_p \times \mathcal{P}_{a.c.}(\mathcal{S})^\kappa$. Using Glivenko-Cantelli's theorem we have that

$$(\widehat{p}, \widehat{F}_{Y|X=x_1,n_1}, \ldots, \widehat{F}_{Y|X=x_\kappa,n_\kappa})$$

converges in probability to $(\underline{p}, F_{Y|X=x_1}, \ldots, F_{Y|X=x_\kappa})$ as n goes to infinity (see, *e.g.*, Van der Vaart, 2000). Using the continuous mapping theorem yields the result. □

Lemma 3 *Let $k = 1, \ldots, p+1$ and $(x, y) \in \text{Supp}(X, Y)$, then the function $R_{k,x,y}$ is Lipschitz on $\mathcal{P}_1(\text{Supp}(\Gamma)) \cap \{F : \int_{\mathcal{I}(x,y)} dF \geq 1/c > 0\}$.*

Proof of Lemma 3. Let $k = 1, \ldots, p+1$ and $(x, y) \in \text{Supp}(X, Y)$. Consider two distributions $F_1, F_2 \in \mathcal{P}_2(\text{Supp}(\Gamma))$, then

$$
\begin{aligned}
&|R_{k,x,y}(F_1) - R_{k,x,y}(F_2)| \\
&= \left| \frac{\int_{\mathcal{I}(x,y)} g_k dF_1}{\int_{\mathcal{I}(x,y)} dF_1} - \frac{\int_{\mathcal{I}(x,y)} g_k dF_2}{\int_{\mathcal{I}(x,y)} dF_2} \right| \\
&= \left| \frac{\int_{\mathcal{I}(x,y)} g_k d(F_1 - F_2)}{\int_{\mathcal{I}(x,y)} dF_1} - \int_{\mathcal{I}(x,y)} g_k dF_2 \left(\frac{1}{\int_{\mathcal{I}(x,y)} dF_2} - \frac{1}{\int_{\mathcal{I}(x,y)} dF_1} \right) \right| \\
&\leq c \left| \int_{\mathcal{I}(x,y)} g_k d(F_1 - F_2) \right| + c^2 \mathbb{E}_{F_2}(|\Gamma|) \left| \int_{\mathcal{I}(x,y)} d(F_1 - F_2) \right| \\
&\leq c(1 + c\mathbb{E}_{F_2}(|\Gamma|)) W_1(F_1, F_2), \qquad (94)
\end{aligned}
$$

using that by duality $W_1(F_1, F_2) = \max \left\{ \int \phi d(F_1 - F_2),\ \phi \in \text{Lip}_1(\text{Supp}(\Gamma)) \right\}$, where $\text{Lip}_1(\text{Supp}(\Gamma))$ is the set of functions which are 1-Lipschitz on $\text{Supp}(\Gamma)$, which is compact. This yields the result. □

B Data-driven rule for selecting the tuning parameters

For simplicty of exposition, I first present the method when $f_{Y|X}$ and f_X are known, then turn to the general case. I use the Goldenshluger-Lepski method (see, *e.g.*, Goldenshluger and Lepski, 2014; Lacour and Massart, 2016) for the data-driven choice of j_0. Let $p_n := \theta \ln(n)$, $\theta > 6$ and, for all $j_0 \in \mathbb{N}^{\mathbb{R}}$, $j \in \mathbb{N}$, $j_{\max} = \lfloor \check{j} \rfloor$, where $\check{j}$ is solution of $(\check{j} + 1)^{p+2} = n$,

$$
\begin{aligned}
\beta_l(y, j_0) &:= \max_{j_0 + 1 \leq j' \leq j_{\max}} \left(\sum_{|k|_\infty \leq j'} k_l(k_l + 1) \left| \widetilde{d}_k(y) \right|^2 - \Sigma(j') \right)_+, \\
\Sigma(j_0) &:= \frac{24(1 + 2p_n)(j_0 + 1)^{p+2} c_X}{n}, \qquad (95)
\end{aligned}
$$

and $\widetilde{j}_0$ is defined as

$$\forall y \in \mathcal{S}_Y, \quad \widetilde{j}_{0,l}(y) \in \underset{0\leq j\leq j_{\max}}{\operatorname{argmin}} \left(\beta_l(y,j) + \Sigma(j)\right). \tag{96}$$

Proposition 10 (Data-driven convergence rates for the L^2 risk) *Let $\sigma = s + 1 - p/2$, $l > 0$, and $s > p - 1/2$. Make assumptions 2, 3-(B), 9 and 11, then we have that, for $k = 1, \dots, p+1$,*

$$\frac{1}{r(n)} \sup_{f_\Gamma \in \mathcal{H}^\sigma(l)} \mathcal{R}_n\left(\widetilde{PE}_k^{\widetilde{j}_0, GT}, PE_k\right) = O(1), \tag{97}$$

where $r(n) = (n/\ln(n))^{-s/(2s+p+2)}$.

Proposition 10 shows that choosing adaptively the parameter j_0 only yields a logarithmic penalty in the convergence rates compared to the optimal choice. Note that to establish these rates the upper bounds (95) on the variance does not need to depend on y. However, in practice keeping the term $E(\mathbb{1}\{Y_i \leq y\})$ allows to obtain better performances in practice. As standard in the literature (see, *e.g.*, Comte et al., 2013; Dion, 2014), the multiplicative constant appearing in (95) is in practice calibrated from a simulation study.

Let us now precise the statement when $f_{Y|X}$ and f_X are also estimated. Define $\widehat{j}_0$ similarly to $\widetilde{j}_0$ replacing $\widetilde{d}_k$ by $\widehat{d}_k$:

$$\beta_l(y, j_0) := \max_{j_0+1\leq j'\leq j_{\max}} \left(\sum_{|k|_\infty \leq j'} k_l(k_l+1) \left|\widehat{d}_k(y)\right|^2 - \Sigma(j') \right)_+,$$
$$\forall y \in \mathcal{S}_Y, \quad \widehat{j}_{0,l}(y) \in \underset{0\leq j\leq j_{\max}}{\operatorname{argmin}} \left(\beta_l(y,j) + \Sigma(j)\right).$$

Proposition 11 (Data-driven convergence rates for the L^2 risk, complete) *Let $\sigma = s + 1 - p/2$, $l > 0$, $N \in \mathbb{N}$. Make assumptions 2, 3-(B), 9 and 11, then we have that, for $k = 1, \dots, p$,*

$$\frac{1}{r(n_e)} \sup_{\substack{\mathbb{P}_\Gamma \in \mathcal{H}^\sigma(l) \\ f_X \in \mathcal{E},\ f_{Y|X} \in \mathcal{E}'}} \mathcal{R}^2_{n_0,n_1}\left(\widehat{PE}_k^{\widehat{j}_0}, PE_k\right) = \underset{\mathcal{O}_{n_0,n_1,n}}{O_p}(1), \tag{98}$$

where $\mathcal{O}_{n_0,n_1,n} = \{v(n_0,\mathcal{E})/\delta(n_0) \leq n^{-2}, v(n_1,\mathcal{E}')/(\delta(n_0)\delta(n_1)) \leq n^{-2}\}$, and $r(n_e) = (n_e/\ln(n_e))^{-s/(2s+p+2)}$.

Proof in the general case. Let $\mathcal{J}_n$ be the set of functions $j \in \mathbb{N}_0^{\mathbb{R}}$ such that for all $y \in \mathcal{S}_Y$, $j(y) \in \{0, \dots, j_{\max}\}$. I use, for all $k \in \mathbb{N}_0^{p+1}$, $\Delta_k := \widehat{d}_k(y) - \widetilde{d}_k(y)$, $\widetilde{\Delta}_k := \widetilde{d}_k(y) - d_k(y)$,

$$W_l^{j_0} : (y,x) \mapsto \left(\widehat{\partial_l F}_{Y|X}^{j_0} - \partial_l F_{Y|X}\right)(y,x).$$

I also use

$$\Xi_l(y, j_0) := \sum_{|k|_\infty > j_0} k_l(k_l+1)\left|d_k(y)\right|^2, \quad S_{1,l}(y, j_0) := \sum_{|k|_\infty \le j_0} k_l(k_l+1)\left|\Delta_k(y)\right|^2,$$

$$S_{2,l}(y, j_0) := \int_{\mathrm{Supp}(X)} \left|\left(\widetilde{\partial_l F}_{Y|X}^{j_0} - \partial_l F_{Y|X}^{j_0}\right)(y,x)\right|^2 \widetilde{\mu}^2(x)dx, \; L := \frac{\sqrt{2}}{42|\mathcal{S}_Y|},$$

$$\Psi_{0,n} := \exp\left(-\frac{p_n}{6}\right) + \frac{294}{n}\exp\left(-L\sqrt{np_n}\right).$$

Lemma 4 *For all $y \in \mathrm{Supp}(Y)$, $l = 1, \dots, p$, and $j_0 \in \{0, \dots, j_{\max}\}$, we have*

$$\mathbb{E}\left[S_{1,l}\left(y, \widehat{j}_0(y)\right)\right] \le Z_{n_0} C_X (j_{\max}+1)^{p+2}, \tag{99}$$

$$\mathbb{E}\left[\left(S_{2,l}(y, j_0) - \frac{\Sigma(j_0)}{6}\right)_+\right] \le 48\frac{|\mathcal{S}_Y|^2 c_X (j_0+1)^{p+2}}{n}\Psi_{0,n}. \tag{100}$$

Proof of Lemma 4. Let the parameters in the *for all* statement be given and $l \in 1, \dots, p$.

Proof of (99). Using

$$\mathbb{E}\left[\left|\Delta_k(y)\right|^2\right] \le \mathbb{E}\left[\frac{Z_{n_0}}{n^2}\left|\sum_{i=1}^n |L_k(X_i)|\right|^2\right] \le Z_{n_0} C_X$$

(46), we obtain

$$\mathbb{E}\left[S_{1,l}\left(y, \widehat{j}_0(y)\right)\right] \le \sum_{|k|_\infty \le j_{\max}} k_l(k_l+1)\mathbb{E}\left[\left|\Delta_k(y)\right|^2\right] \le Z_{n_0} C_X (j_{\max}+1)^{p+2}.$$

Proof of (100). We use

$$S_{2,l}(y, j_0) = \sup_{v \in \mathcal{U}} \left|\nu_n^y(v)\right|^2,$$

$$\nu_n^y(v) := \left\langle \left(\partial_l \widetilde{F}_{Y|X}^{j_0} - \partial_l F_{Y|X}^{j_0}\right)(y, \cdot), v(\cdot)\right\rangle_{L^2_{\widetilde{\mu}}(\mathrm{Supp}(X))}$$

$$= \frac{1}{n}\sum_{i=1}^n \left(f_v^y(X_i, Y_i) - \mathbb{E}\left[f_v^y(X_i, Y_i)\right]\right),$$

$$f_v^y(\cdot, \star) := \frac{\mathbb{1}\{\star \le y\}}{f_X(\cdot)} \int_{\mathrm{Supp}(X)} \sum_{|k|_\infty \le j_0} \sqrt{k_l(k_l+1)} L_k(\cdot)\Omega_{l,k}(x)\overline{v}(x)dx,$$

where $\mathcal{U}$ is a countable dense set of measurable functions of $\left\{v : \|v\|_{L^2_\mu(\mathrm{Supp}(X))} = 1\right\}$ and check the conditions of the Talagrand inequality given in Lemma B.15 in Gaillac and Gautier (2022) with $\eta = p_n$ and $\Lambda(p_n) = 1$. For all $u \in \mathcal{U}$, using the Cauchy-Schwarz inequality for the first display, (70) for the third, and (46) for the fourth one, we obtain

$$\begin{aligned}
\|f_v^y\|_{L^\infty(\mathcal{S})} &\leq \sqrt{c_X}|\mathcal{S}_Y| \left\| \left(\sum_{|k|_\infty \leq j_0} k_l(k_l+1) \left|L_k(\cdot)\right|^2 \int_{\mathrm{Supp}(X)} \left|\Omega_{l,k}(x)\right|^2 dx \right)^{1/2} \right\|_{L^\infty(\mathrm{Supp}(X))} \\
&\leq \sqrt{c_X}|\mathcal{S}_Y| \left\| \left(\sum_{|k|_\infty \leq j_0} k_l(k_l+1) \left|L_k(\cdot)\right|^2 \right)^{1/2} \right\|_{L^\infty(\mathrm{Supp}(X))} \\
&\leq \sqrt{c_X}|\mathcal{S}_Y| \left(\sum_{|k|_\infty \leq j_0} k_l(k_l+1)^{p+1} \right)^{1/2} \\
&\leq \sqrt{c_X}|\mathcal{S}_Y|(j_0+1)^{(p+2)/2}. \qquad (101)
\end{aligned}$$

By the Cauchy-Schwarz inequality, Lemma 2, and (46), we have

$$\begin{aligned}
\mathbb{E}\left[\sup_{v\in\mathcal{U}} |\nu_n^y(v)|\right]^2 \leq \mathbb{E}\left[\sup_{v\in\mathcal{U}} |\nu_n^y(v)|^2\right] &\leq \mathbb{E}\left[\left\| \left(\widetilde{\partial_l F}_{Y|X}^{j_0} - \partial_l F_{Y|X}^{j_0}\right)(y,\cdot)\right\|^2_{L^2_\mu(\mathrm{Supp}(X))}\right] \\
&\leq \frac{c_X}{n} \sum_{|k|_\infty \leq j_0} k_l(k_l+1) \\
&\leq \frac{c_X}{n}(j_0+1)^{p+2} = \frac{\Sigma(j_0)}{24(1+2p_n)}.
\end{aligned}$$

Finally, by the Cauchy-Schwarz inequality and (101), we have

$$\begin{aligned}
\mathrm{Var}\left(\mathfrak{R}(f_v^y(Y_i,X_i))\right) \vee \mathrm{Var}\left(\mathfrak{I}(f_v^y(Y_i,X_i))\right) &\leq \int_{\mathcal{S}_{X,Y}} \left|f_v^y(y',x)\right|^2 f_{Y,X}(y',x)dy'dx \\
&\leq c_X|\mathcal{S}_Y|^2(j_0+1)^{p+2}.
\end{aligned}$$

□

Denote by

$$\mathcal{R}^2_{n_0,n_1}\left(\widehat{\mathrm{PE}}_k^{\hat{j}_0}, \mathrm{PE}_k\right) := \mathbb{E}\left[\left\|\widehat{PE}_k^{\hat{j}_0} - PE_k\right\|^2_{L^2_\mu(\mathcal{S}_{X,Y})}\right].$$

Lemma 5 *For all $j_0 \in \mathcal{J}_n$,*

$$\mathcal{R}^2_{n_0,n_1}\left(\widehat{PE}_1^{\hat{j}_0}, PE_1\right) \leq \sum_{l=1}^{p} 21 C_{1,l} \mathbb{E}\left[\left\|W_l^{j_0}\right\|^2_{L^2_\mu(\mathcal{S}_{X,Y})}\right] + 36 C_1 p \int_{y \in \mathrm{Supp}(Y)} \Sigma(y, j_0(y)) dy + \widetilde{C}_0 \frac{Z_{n_1}(j_{\max}+1)^{p+2}}{\delta(n_0)} + (j_{\max}+1)^{p+2} 36 C_1 p \Pi(n, Z_{n_0}),$$

where $C_1 := \max_{l=1,\ldots,p} C_{1,l}$ and $\Pi(n, Z_{n_0}) := 96 |\mathcal{S}_Y|^2 c_X \Psi_{0,n}/n + 2 Z_{n_0} C_X$.

Proof of Lemma 5. Let $j_0 \in \{0, \ldots, j_{\max}\}$. We have, using (76)

$$\mathcal{R}^2_{n_0,n_1}\left(\widehat{\mathrm{PE}}_1^{\hat{j}_0}, \mathrm{PE}_1\right) \leq \widetilde{C}_0 \frac{Z_{n_1}(j_{\max}+1)^{p+2}}{\delta(n_0)} + C_1 \sum_{l=1}^{p} \mathbb{E}\left[\left\|W_l^{\hat{j}_{0,l}}\right\|^2_{L^2_\mu(\mathcal{S}_{X,Y})}\right]. \quad (102)$$

Using, for all $j_1, j_2 \in \mathbb{N}$ and $y \in \mathcal{S}_Y$,

$$\check{R}^{j_2}_{j_1,l}(y, \cdot) := \left(\widehat{\partial_l F}^{j_2 \vee j_1}_{Y|X} - \widehat{\partial_l F}^{j_1}_{Y|X}\right)(y, \cdot),$$

we have $W_l^{\hat{j}_0} = \check{R}^{j_0}_{\hat{j}_0(y),l} - \check{R}^{\hat{j}_0(y)}_{j_0,l} + L_l^{j_0}$. We obtain, using the convexity of $x \mapsto x^2$,

$$\mathbb{E}\left[\left\|W_l^{\hat{j}_0}\right\|^2_{L^2_\mu(\mathcal{S}_{X,Y})}\right] \leq 3\mathbb{E}\left[\left\|\check{R}^{j_0}_{\hat{j}_0,l}\right\|^2_{L^2_\mu(\mathcal{S}_{X,Y})}\right] + 3\mathbb{E}\left[\left\|\check{R}^{\hat{j}_0}_{j_0,l}\right\|^2_{L^2_\mu(\mathcal{S}_{X,Y})}\right] + 3\mathbb{E}\left[\left\|W_l^{j_0}\right\|^2_{L^2_\mu(\mathcal{S}_{X,Y})}\right].$$

Because

$$\beta_l(y, j_0) = \max_{j_0+1 \leq j' \leq j_{\max}} \left(\sum_{|k|_\infty \leq j'} k_l(k_l+1)\left|\widehat{d}_k(y)\right|^2 - \Sigma(j')\right)_+,$$

we have, for all $l = 1, \ldots, p$,

$$\mathbb{E}\left[\left\|\widetilde{R}^{j_2}_{j_1,l}\right\|^2_{L^2_\mu(\mathcal{S}_{X,Y})}\right] \leq \int_{\mathcal{S}_Y} \left(\mathbb{E}\left[\beta_l(y, j_1)\right] + \mathbb{E}\left[\Sigma(j_2)\right]\right) dy$$

for possibly random j_1 and j_2. Using (96) yields

$$\mathbb{E}\left[\left\|L_l^{\hat{j}_{0,l}}\right\|^2_{L^2_\mu(\mathcal{S}_{X,Y})}\right] \leq 6 \int_{\mathcal{S}_Y} \left(\mathbb{E}\left[\beta_l(y, j_0)\right] + \Sigma(j_0)\right) dy + 3\mathbb{E}\left[\left\|W_l^{j_0}\right\|^2_{L^2_\mu(\mathcal{S}_{X,Y})}\right].$$

Using the convexity of $x \mapsto x^2$ and, for all $j' \in \{0, \ldots, j_{\max}\}$,

$$\widetilde{K}^{j'}_{j_0,a}(y) := \sum_{|k|_\infty \leq j_0 \vee j'} k_l(k_l+1)\left|\widehat{d}_k(y) - d_k(y)\right|^2$$

$$\widetilde{K}^{j'}_{j_0,b}(y) := \sum_{|k|_\infty \leq j_0} k_l(k_l+1)\left|\widehat{d}_k(y) - d_k(y)\right|^2$$

$$\widetilde{K}^{j'}_{j_0,c}(y) := \sum_{j_0 \leq |k|_\infty \leq j_0 \vee j'} k_l(k_l+1)\left|d_k(y)\right|^2,$$

we have

$$\beta_l(y,j_0) \leq \max_{\substack{0\leq j'\leq j_{\max}\\ j'\in\mathbb{N}}} \left(3\sum_{m\in\{a,b,c\}} \widetilde{K}^{j'}_{j_0,m}(y) - \Sigma(j')\right)_+ .$$

We obtain, for all $y\in\mathcal{S}_Y$ and $l\in\{1,\dots,p\}$,

$$\widetilde{K}^{j'}_{j_0,c}(y) \leq \sum_{|k|_\infty\geq j_0} k_l(k_l+1)|d_k(y)|^2 \leq \left\|W_l^{j_0}(y,\cdot_1)\right\|^2_{L^2_\mu(\mathrm{Supp}(X))}$$

hence

$$\beta_l(y,j_0) \leq \max_{\substack{0\leq j'\leq j_{\max}\\ j'\in\mathbb{N}}} \left(6\sum_{j_0\leq|k|_\infty\leq j'} k_l(k_l+1)\left|\Delta_k(y)\right|^2 - \Sigma(j')\right)_+ + 3\left\|W_l^{j_0}(y,\cdot_1)\right\|^2_{L^2_\mu(\mathrm{Supp}(X))}.$$

Finally, denoting by

$$\widetilde{\beta}_l(y,j_0) := \max_{\substack{0\leq j'\leq j_{\max}\\ j'\in\mathbb{N}}} \left(\sum_{j_0\leq|k|_\infty\leq j'} k_l(k_l+1)\left|\Delta_k(y)\right|^2 - \frac{\Sigma(j')}{6}\right)_+$$

we have

$$\mathbb{E}\left[\left\|W_l^{\widehat{j}_0}\right\|^2_{L^2_\mu(\mathcal{S}_{X,Y})}\right] \leq 36\int_{\mathcal{S}_Y}\left(\widetilde{\beta}_l(y,j_0)+\Sigma\left(j_0\right)\right)dy + 21\mathbb{E}\left[\left\|W_l^{j_0}\right\|^2_{L^2_\mu(\mathcal{S}_{X,Y})}\right].$$

Using Lemma 4 for the second inequality, we obtain

$$\begin{aligned}\widetilde{\beta}_l(y,j_0) &\leq 2\mathbb{E}\left[\max_{j'\leq j_{\max}}\left(S_{2,l}(y,j') - \frac{\Sigma(j')}{6}\right)_+\right] + 2\mathbb{E}\left[\max_{j'\leq j_{\max}} S_{1,l}(y,j')\right]\\ &\leq 96\frac{|\mathcal{S}_Y|^2 c_X(j_{\max}+1)^{p+2}}{n}\Psi_{0,n} + 2Z_{n_0}C_X(j_{\max}+1)^{p+2},\end{aligned}$$

Hence the result. □

Proof of propositions 10 and 11.

Let n_0, n_1, n such that $v(n_0,\mathcal{E})/\delta(n_0) \leq n^{-2}\ln(n)^{-1}$, $v(n_1,\mathcal{E}')/(\delta(n_0)\delta(n_1)) \leq n^{-2}\ln(n)^{-1}$, and $n_e \geq 3$. Consider the case $k=1$, and take $j_0\in\mathcal{J}_n$. Start from Lemma 5 and use (76), (77) and (81), which yield

$$\begin{aligned}\mathcal{R}^2_{n_0,n_1}\left(\widehat{\mathrm{PE}}_1^{\widehat{j}_0},\mathrm{PE}_1\right) \leq& 21C_8\left(2Z_{n_0}C_X(j_0+1)^{p+2} + 2c_X\frac{(j_0+1)^{p+2}}{n} + \frac{C_{10}}{(j_0+1)^{2s}}\right)\\ &+ \frac{864C_1p|\mathcal{S}_Y|(1+2p_n)(j_0+1)^{p+2}c_X}{n} + \widetilde{C}_0\frac{Z_{n_1}(j_{\max}+1)^{p+2}}{\delta(n_0)}\\ &+ (j_{\max}+1)^{p+2}36C_1p\Pi(n,Z_{n_0}).\end{aligned}$$

Then, we have

$$\begin{aligned}
&\exp\left(-p_n/6\right)(j_{\max}+1)^{p+2} \leq 1\\
&Z_{n_0}(j_{\max}+1)^{p+2} \leq \frac{M_{\mathcal{E},\eta,0}v(n_0,\mathcal{E})n}{\delta(n_0)} \leq \frac{M_{\mathcal{E},\eta,0}}{n},\\
&\frac{Z_{n_1}(j_{\max}+1)^{p+2}}{\delta(n_0)} \leq \frac{M_{\mathcal{E}',\eta,1}nv(n_1,\mathcal{E}')}{\delta(n_1)\delta(n_0)} \leq \frac{M_{\mathcal{E}',\eta,1}}{n}\\
&\frac{(j_{\max}+1)^{p+2}\exp(-L\sqrt{np_n})}{n} \leq \exp(-L\sqrt{\theta n\ln(n)}) \leq 1.
\end{aligned}$$

We conclude using $j^* = (n_e/\ln(n_e))^{1/(2s+p+2)}$ satisfy $(j^*+1)^{p+2} \leq n_e$ hence belongs to $\mathcal{J}_n$, which yields the result as

$$\begin{aligned}
\mathcal{R}^2_{n_0,n_1}\left(\widehat{PE}_1^{\hat{j}_0}, PE_1\right) \leq &\frac{42C_8M_{\mathcal{E},\eta,0}C_X + \widetilde{C}_0M_{\mathcal{E}',\eta,1}}{n}\\
&+ 2c_X\left(21C_8 + 432C_1p|\mathcal{S}_Y|\right)\frac{(1+2p_n)(j^*+1)^{p+2}}{n} + \frac{21C_8C_{10}}{(j^*+1)^{2s}}\\
&+ \frac{36C1p}{n}\left(96|\mathcal{S}_Y|^2c_X\left(1+\frac{294}{n}\right) + \frac{2C_XM_{\mathcal{E},\eta,0}}{n}\right). \qquad \square
\end{aligned}$$

C Asymptotic normality for the GT estimator

Let us state asymptotic normality of the GT estimator $\widehat{\mathrm{PE}}^{j_0}$ defined in Section 4.1.1. For the sake of simplicity in the presentation here, I consider the case where $f_{Y|X}$ and f_X are known, but the proof in this section shows that the effect of estimating $f_{Y|X}$ and f_X under further standard assumptions 13 is negligible.

Assumption 13 *Assume (Asn.1)* $(j_0+1)^{3p+2}/n \longrightarrow 0$*; (Asn.2)* $\inf_{v\in\mathrm{Supp}(Y)} f_{Y|X}(v|x) > 0$*; (Asn.3)* $n/(j_0+1)^{2s-p} \underset{n\to\infty}{\longrightarrow} 0$.

Let $(x,y) \in \mathrm{Supp}(X,Y)$. We have, when $f_{Y|X}$ and f_X are known,

$$\widehat{\mathrm{PE}}_k^{j_0}(x,y) - y\mathbb{1}\{k=1\} = \frac{1}{n}\sum_{i=1}^{n}\zeta_{k,i}^{j_0}(x,y), \tag{103}$$

where

$$\zeta_{k,i}^{j_0}(x,y) := \sum_{l=1}^{p}\frac{x_l(\mathbb{1}\{k=1\}+\delta_{k-1,l})\mathbb{1}\{Y_i\leq y\}}{f_{Y|X}(y|x)f_X(X_i)}\sum_{|j|_\infty\leq j_0}L_j\left(X_i\right)\partial_lL_j\left(x\right).$$

Proposition 12 *(Asymptotic normality) Let* (x,y) *be in the interior of* $\text{Supp}(X,Y)$, $\sigma = s{+}1{-}p$, $s > p{+}1$, *and* $\omega < 1$. *Let* $k = 1,\dots,p{+}1$ *and* $v_l^{j_0}(x,y) := Var\left(\zeta_{l,i}^{j_0}(x,y)\right)$. *Make assumptions 2, 3-(B), 11 and 9, then we have,*

$$\sqrt{\frac{n}{v_k^{j_0}(x,y)}}\left(\widehat{PE}_k^{j_0,GT}(x,y) - PE_k(x,y)\right) \xrightarrow{\mathcal{L}} \mathcal{N}(0,1).$$

It involves undersmoothing since the optimal choice of the parameter j_0 in Proposition 6 does not satisfy (Asn.2) and should be taken lower than this optimal value to obtain Proposition 12. One can further extend Proposition 12 to (x,y) belonging to the boundary of $\text{Supp}(X,Y)$ at the cost of strengthening the Assumption C. In this case, weaker conditions could be obtained using vaguelets instead of the Legendre polynomials. However, as usual in the literature, Proposition 12 does not apply to data-driven selected parameters as in Section B, since these are random quantities.

Proof. I consider the context of Section A.2.1, where f_X and $f_{Y|X}$ are estimated. I add the following assumption.

Assumption 14 *(Asn.4)* $nv(n_1,\mathcal{E}')/(\delta(n_0)\delta(n_1)) \underset{n,n_0,n_1\to\infty}{\longrightarrow} 0$;

(Asn.5) $nv(n_0,\mathcal{E})/\delta(n_0) \underset{n,n_0\to\infty}{\longrightarrow} 0$.

Under these assumptions 14 and 12, Proposition 12 holds with f_X and $f_{Y|X}$ replaced by their respective trimmed estimators. I consider this context for the proof hereafter. I use the notation

$$\mathcal{K}_{l,j_0}(X_i,x) := \sum_{|k|_\infty \le j_0} L_k(X_i)\sqrt{k_l(k_l+1)}\Omega_{l,k}(x).$$

Proof of Proposition 12. I consider $k = 1$ as the other cases can directly be deduced from it. Using the notation (130) we have

$$\begin{aligned}\sqrt{n}\left(\widehat{PE}_1^{j_0}(x,y) - PE_1(x,y)\right) &= \sqrt{n}\sum_{l=1}^{p} x_l\left(\frac{\widehat{\partial_l F}_{Y|X}^{j_0}(y|x)}{\widehat{f}_{Y|X}^{\delta}(y|x)} - \frac{\partial_l F_{Y|X}(y|x)}{f_{Y|X}(y|x)}\right)\\ &= \sqrt{n}\sum_{j=1}^{4} R_j(x,y),\end{aligned}$$

where

$$R_1(x,y) := \sum_{l=1}^{p} x_l \left(\frac{1}{\widehat{f}^{\delta}_{Y|X}(y|x)} - \frac{1}{f_{Y|X}(y|x)} \right) \widehat{\partial_l F}^{j_0}_{Y|X}(y|x)$$

$$R_2(x,y) := \frac{1}{f_{Y|X}(y|x)} \sum_{l=1}^{p} x_l \left(\widehat{\partial_l F}^{j_0}_{Y|X}(y|x) - \widetilde{\partial_l F}^{j_0}_{Y|X}(y|x) \right)$$

$$R_3(x,y) := \frac{1}{f_{Y|X}(y|x)} \sum_{l=1}^{p} x_l \left(\widetilde{\partial_l F}^{j_0}_{Y|X}(y|x) - \partial_l F^{j_0}_{Y|X}(y|x) \right)$$

$$R_4(x,y) := \frac{1}{f_{Y|X}(y|x)} \sum_{l=1}^{p} x_l \left(\partial_l F^{j_0}_{Y|X}(y|x) - \partial_l F_{Y|X}(y|x) \right).$$

We have

$$\frac{\sqrt{n}}{f_{Y|X}(y|x)} \sum_{l=1}^{p} x_l \widetilde{\partial_l F}^{j_0}_{Y|X}(y|x) = n^{-1/2} \sum_{i=1}^{n} \zeta^{j_0}_{1,i}(x,y),$$

and $\mathbb{E}\left[\widetilde{\partial_l F}^{j_0}_{Y|X}(y|x)\right] = \partial_l F^{j_0}_{Y|X}(y|x)$. Using (103), we show below that $\zeta^{j_0}_{1,i}(x,y)$ satisfies the Lyapounov condition, for $\nu > 0$,

$$\frac{\mathbb{E}\left[\left|\zeta^{j_0}_{1,i}(x,y) - \mathbb{E}\left[\zeta^{j_0}_{1,i}(x,y)\right]\right|^{2+\nu}\right]}{n^{\nu/2}\mathrm{Var}(\zeta^{j_0}_{1,i}(x,y))^{1+\nu/2}} \longrightarrow 0.$$

Lower bound on $\mathbf{Var}(\zeta^{j_0}_{1,i}(x,y))^{1+\nu/2}$. Because $\mathbb{E}\left[\zeta^{j_0}_{1,i}(x,y)\right]$ converges to $PE_1(x,y)$, it is sufficient to get a lower bound on $\mathbb{E}\left[\left|\zeta^{j_0}_{1,i}(x,y)\right|^2\right]$. We have, using that (L_k) are orthonormal on $L^2(\mathrm{Supp}(X))$ for the last display,

$$\begin{aligned}
\mathbb{E}\left[\left|\zeta^{j_0}_{1,i}(x,y)\right|^2\right] &= \int_{\mathrm{Supp}(X,Y)} \left| \sum_{l=1}^{p} x_l \mathcal{K}_l(v,x) \right|^2 \frac{\mathbb{1}\{z \le y\} f_{X,Y}(v,z)}{f_{Y|X}(y|x)^2 f_X(v)^2} dz dv \\
&= \int_{\mathrm{Supp}(X)} \left| \sum_{l=1}^{p} x_l \mathcal{K}_l(v,x) \right|^2 \frac{F_{Y|X}(y|v)}{f_{Y|X}(y|x)^2 f_X(v)} dv \\
&\ge \frac{\widetilde{c}_{Y,X}(y)}{f_{Y|X}(y|x)^2} \int_{\mathrm{Supp}(X)} \left| \sum_{|k|_\infty \le j_0} \left(\sum_{l=1}^{p} x_l \partial_l L_k(x) \right) L_k(v) \right|^2 dv \\
&\ge \frac{\widetilde{c}_{Y,X}(y)}{f_{Y|X}(y|x)^2} \sum_{|k|_\infty \le j_0} \left| \sum_{l=1}^{p} x_l \partial_l L_k(x) \right|^2, \qquad (104)
\end{aligned}$$

where $\widetilde{c}_{Y,X}(y) := \inf_{v\in \mathrm{Supp}(X)} F_{Y|X}(y|v)/f_X(v)$.

Upper bound on the Lyapounov condition. We have,

$$\begin{aligned}
&\mathbb{E}\left[\left|\zeta_{1,i}^{j_0}(x,y)\right|^{2+\nu}\right] \\
&= \int_{\mathrm{Supp}(X,Y)} \left|\sum_{l=1}^{p} \frac{x_l \mathcal{K}_l(v,x)}{f_{Y|X}(y|x) f_X(v)}\right|^{2+\nu} f_{X,Y}(v,z)dzdv \\
&\leq \frac{c_X^{1+\nu}}{f_{Y|X}(y|x)^{2+\nu}} \int_{\mathrm{Supp}(X,Y)} \left|\sum_{|k|_\infty \leq j_0} \left(\sum_{l=1}^{p} x_l \partial_l L_k(x)\right) L_k(v)\right|^{2+\nu} f_{Y|X}(z|v)dzdv, \\
&\leq \frac{c_X^{1+\nu}}{f_{Y|X}(y|x)^{2+\nu}} \sup_{v\in\mathrm{Supp}(X)} \left|\sum_{|k|_\infty \leq j_0} \left(\sum_{l=1}^{p} x_l \partial_l L_k(x)\right) L_k(v)\right|^{\nu} B(x),
\end{aligned}$$

where, using that (L_k) are orthonormal on $L^2(\mathrm{Supp}(X))$,

$$\begin{aligned}
B(x) &:= \int_{\mathrm{Supp}(X)} \left|\sum_{|k|_\infty \leq j_0} \left(\sum_{l=1}^{p} x_l \partial_l L_k(x)\right) L_k(v)\right|^2 dv \\
&= \sum_{|k|_\infty \leq j_0} \left|\sum_{l=1}^{p} x_l \partial_l L_k(x)\right|^2.
\end{aligned}$$

This yields, using (69) for the second and third inequalities,

$$\begin{aligned}
&\frac{\mathbb{E}\left[\left|\zeta_{1,i}^{j_0}(x,y) - \mathbb{E}\left[\zeta_{1,i}^{j_0}(x,y)\right]\right|^{2+\nu}\right]}{n^{\nu/2}\mathrm{Var}(\zeta_{1,i}^{j_0}(x,y))^{1+\nu/2}} \\
&\leq \frac{c_X^{1+\nu}}{\widetilde{c}_{Y,X}(y)^{1+\nu/2} n^{\nu/2}} \sup_{v\in\mathrm{Supp}(X)} \left|\sum_{|k|_\infty \leq j_0} \left(\sum_{l=1}^{p} x_l \partial_l L_k(x)\right) L_k(v)\right|^{\nu} \\
&\leq \frac{(j_0+1)^{p\nu/2}}{n^{\nu/2}} \frac{c_X^{1+\nu}}{\widetilde{c}_{Y,X}(y)^{1+\nu/2}} \left(\sum_{l=1}^{p} |x_l| \sum_{|k|_\infty \leq j_0} |\partial_l L_k(x)|\right)^{\nu} \\
&\leq \frac{(j_0+1)^{\nu(3p/2+1)}}{n^{\nu/2}} \frac{c_X^{1+\nu}}{\widetilde{c}_{Y,X}(y)^{1+\nu/2}} \Phi(x)^{\nu},
\end{aligned}$$

where $\Phi(x) := (2/\sqrt{\pi})(2/\pi)^p \sum_{l=1}^{p} |x_l| / ((1-(x_l/x_0)^2) \prod_{k=1}^{p} (1-(x_k/x_0)^2)^{1/4})$.

Thus, under condition (Asn.1), the Lyapounov condition is satisfied and we have $\sqrt{n/v_1^{j_0}(x,y)} R_3(x,y) \xrightarrow{d} \mathcal{N}(0,1)$.

We now need to prove that the remaining terms $\sqrt{n/v_1^{j_0}(x,y)} R_j(x,y)$, $j=1,2,4$ are

$o_p(1)$. Using the lower bound (104), it suffices to show for $k \in \{1,2,4\}$

$$\sqrt{\frac{n}{(j_0+1)^{p+2}}} R_k(x,y) = o_p(1). \tag{105}$$

Term $\sqrt{n/v_1^{j_0}(x,y)}R_1(x,y)$. We have, using (69),

$$\begin{aligned}
|R_1(x,y)| &\leq \sum_{l=1}^{p} |x_l| \left| \frac{1}{\widehat{f}_{Y|X}(y|x)} - \frac{1}{f_{Y|X}(y|x)} \right| \left| \widehat{\partial_l F}^{j_0}_{Y|X}(y|x) \right| \\
&\leq \frac{(2/\pi)\Phi(x)\sqrt{Z_{n_1}}(j_0+1)^{p+1}}{\sqrt{\delta_0(n_0)}\prod_{k=1}^{p}(1-(x_k/x_0)^2)^{1/4}}.
\end{aligned}$$

Thus, using $Z_{n_1} = O_p(v(n_1,\mathcal{E}')/\delta(n_1))$ and under (Asn.4) we have (105) for $k=1$.

Term $\sqrt{n/v^{j_0}(x,y)}R_2(x,y)$. Using (69), we have,

$$\begin{aligned}
|R_2(x,y)| &\leq \frac{(2/\pi)^p c_{Y,X}\sqrt{Z_{n_0}}}{\prod_{k=1}^{p}(1-(x_k/x_0)^2)^{1/4}} \sum_{l=1}^{p} |x_l| \sum_{|k|_\infty \leq j_0} |\partial_l L_k(x)| \\
&\leq \frac{c_{Y,X}(2/\pi)^p \Phi(x)}{\prod_{k=1}^{p}(1-(x_k/x_0)^2)^{1/4}} \sqrt{Z_{n_0}}(j_0+1)^{p+1}.
\end{aligned}$$

Thus, under condition (Asn.5) we have (105) for $k=2$.

Term $\sqrt{n/v^{j_0}(x,y)}R_4(x,y)$. Using (80) for the first inequality, then (69), (82) and (85) for the second, we have

$$\begin{aligned}
|R_4(x,y)| &\leq \frac{c_{Y,X}\sqrt{4|y-\underline{y}|}}{\pi^2}\Phi(x) \sum_{j\geq j_0} \sum_{|k|_\infty = j} \frac{1}{(j+1)^s} \left(\int_{\mathbb{R}} |c_k(t)|^2 (j+1)^{2s} k_l^2 dt \right)^{1/2} \\
&\leq \frac{c_{Y,X}\sqrt{4|y-\underline{y}|H_l}}{\pi^2}\Phi(x) \sum_{j\geq j_0} \frac{1}{(j+1)^{s-p}} \\
&\leq \frac{c_{Y,X}\sqrt{4|y-\underline{y}|H_l}}{\pi^2(s-p-1)}\Phi(x) \frac{1}{(j_0+1)^{s-(p+1)}},
\end{aligned}$$

using that $\sum_{j\geq j_0}(j+1)^{p-s} \leq \int_{j_0}^{\infty}(j+1)^{p-s}dj = (j+1)^{p+1-s}/(s-p-1)$ for $s > p+1$. This yields (105) for $k=4$ using condition (Asn.3). This yields the result. □

D Additional material on the application

	OLS (1)	OLS (2)	IV (3)
Female	0.055	0.046	0.072
	(0.016)***	(0.021)*	(0.024)**
Local	-0.013	0.005	-0.007
	(0.016)	(0.022)	(0.024)
Some teacher training	-0.020	-0.032	-0.035
	(0.028)	(0.044)	(0.047)
Has bachelor's degree or better	0.021	0.013	-0.034
	(0.018)	(0.025)	(0.029)
Had > 3 years of exp. in 2007	0.003	0.030	0.014
	(0.021)	(0.040)	(0.044)
Temporary contract	0.010	0.027	0.027
	(0.021)	(0.028)	(0.031)
Public school	-0.117	-0.057	-0.053
	(0.027)***	(0.044)	(0.047)
Mean teacher knowledge	0.042	0.050	0.239
	(0.013)***	(0.017)**	(0.045)***
Fixed effects	District	District	District
Num.Obs.	1509	834	834
F-statistic			149.63
Adj. R^2	0.168	0.179	0.048

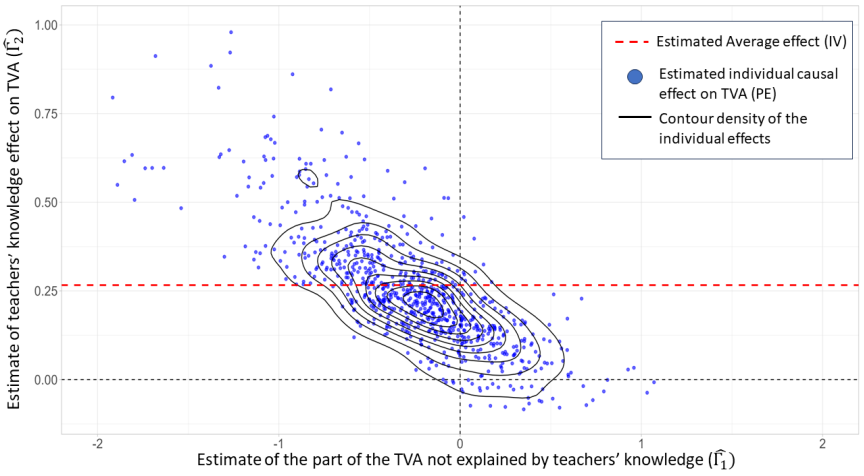
Notes: Significance levels $< 0.1\%$ ***, 1% **, 5% *. In columns (3), I instrument for the teacher's mean score in the first tested year with the mean score in the second year. The sample size is reduced from column (1) to columns (2) and (3) because not all teachers were tested in multiple years. Teachers' knwoledge measured via tests scores is winsorized at a 1% level.

Table 5: First step regression results and IV

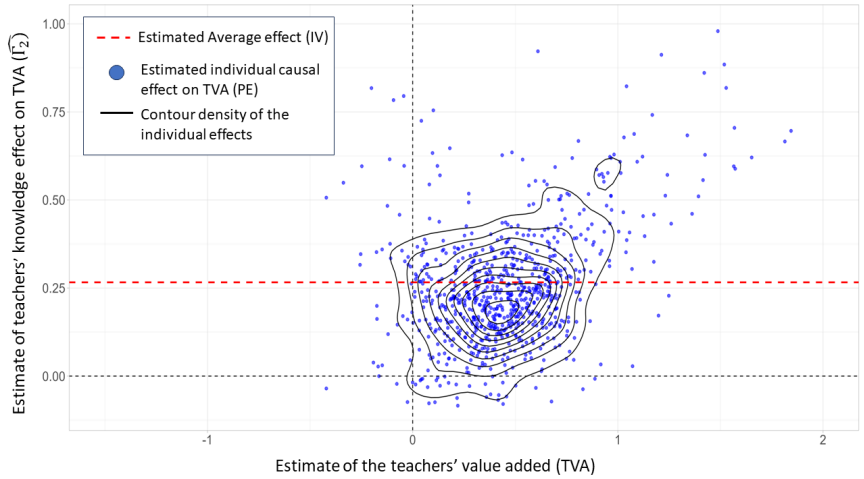
	Est. with A1	Diff. with A2 (ii)	Est. with A2 (ii)	Diff. with A2 (i)	Diff with A3
GT estimator					
First Quartile	0.191	-0.018	0.185	-0.006	-0.047
Median	0.259	-0.010	0.256	-0.002	0.000
Third Quartile	0.357	0.013	0.361	0.007	0.055
GWB estimator					
First Quartile	0.244	-0.020	0.142	-0.019	-0.095
Median	0.301	0.074	0.219	-0.002	0.018
Third Quartile	0.389	0.172	0.329	0.018	0.112

Notes: These results pool teachers from private and public schools. I instrument for the teacher's mean score in the first tested year with the mean score of the second year, which reduces the sample size to 834. Both panels present the quartiles of the distributions either of the estimate values ("Est. with A1" and "Est. with A2 (ii)") or of the differences with respect either to the estimator with A1 (2nd column) or with the estimator with A2 (ii) (5th and 6th columns).

Table 6: Sensitivity of the individual-level estimates to the independence assumption



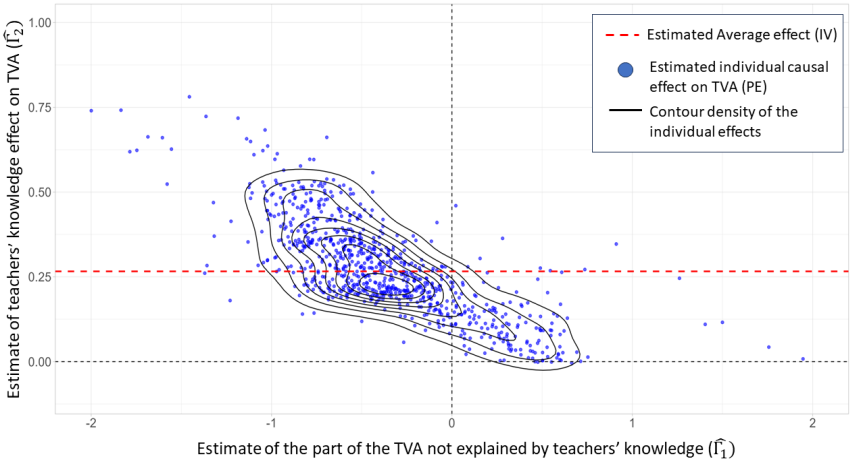
(a) Using the GT estimator



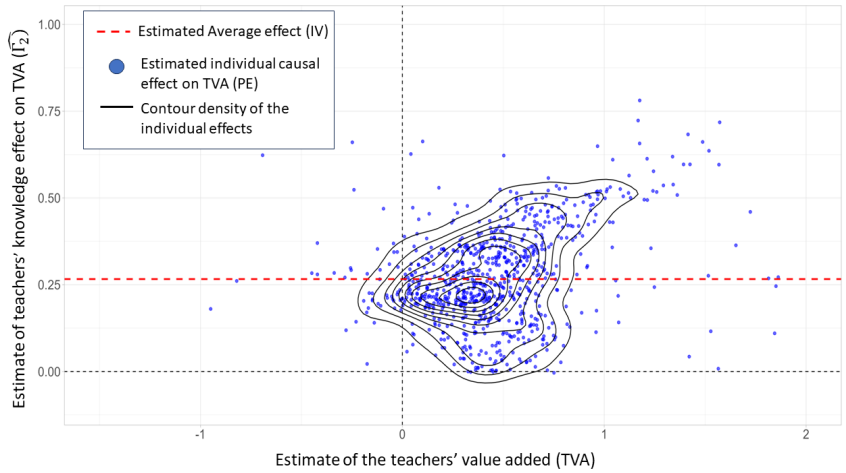
(b) Using the GT estimator and IV

Notes: These results pool teachers from private and public schools. Figure 5(a) (resp. 5(b)) present the estimated joint distribution of the part of the TVA which can not be explained by teacher's knowledge (resp. the estimated TVA) and the estimated effect of teacher knowledge on TVA (Γ_2). This is using the GT estimator with varying coefficients A3 (ii), when I instrument for the teacher's mean score in the first tested year with the mean score of the second year, which reduces the sample size to 834. The dots represent the individual predictions $\widehat{\text{PE}}(X_i, Y_i)$ and the contour lines the levels of the associated fitted density. The dotted red line represents the IV estimates with an homogeneous specification (0.239). Teachers' tests scores are winsorized at a 1% level.

Figure 5: Joint distributions of the coefficients characterizing the TVA



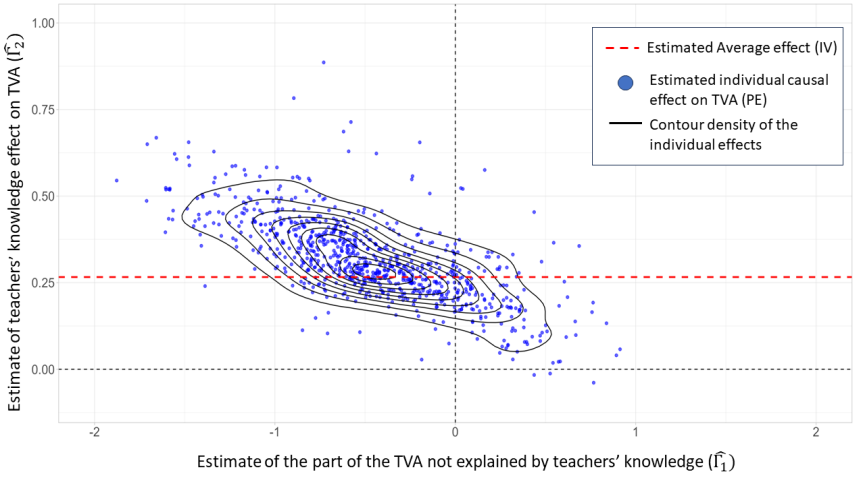
(a) Using the GT estimator



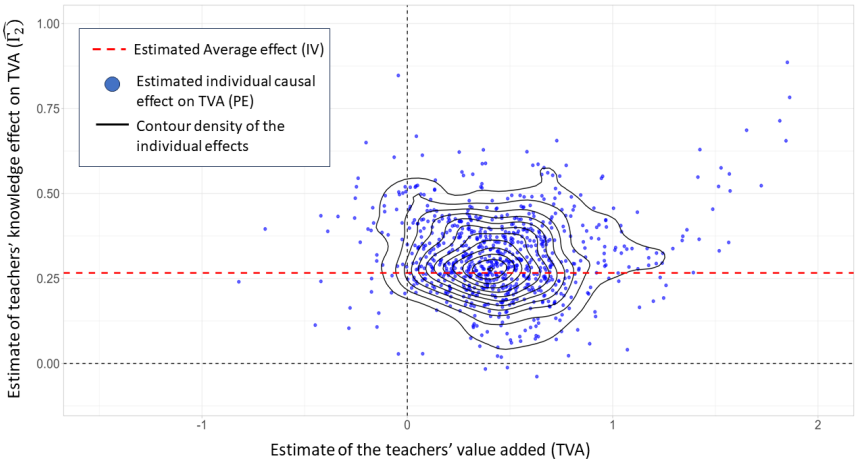
(b) Using the GT estimator and IV

Notes: These results pool teachers from private and public schools. Figure 6(a) (resp. 2(b)) present the estimated joint distribution of the part of the TVA which can not be explained by teacher's knowledge (resp. the estimated TVA) and the estimated effect of teacher knowledge on TVA (Γ_2). This is using the GT estimator with varying coefficients A3 (ii), when I instrument for the teacher's mean score in the first tested year with the mean score of the second year, which reduces the sample size to 834. The dots represent the individual predictions $\widehat{\text{PE}}(X_i, Y_i)$ and the contour lines the levels of the associated fitted density. The dotted red line represents the IV estimates with an homogeneous specification (0.239). Teachers' tests scores are winsorized at a 1% level.

Figure 6: Joint distributions of the coefficients characterizing the TVA



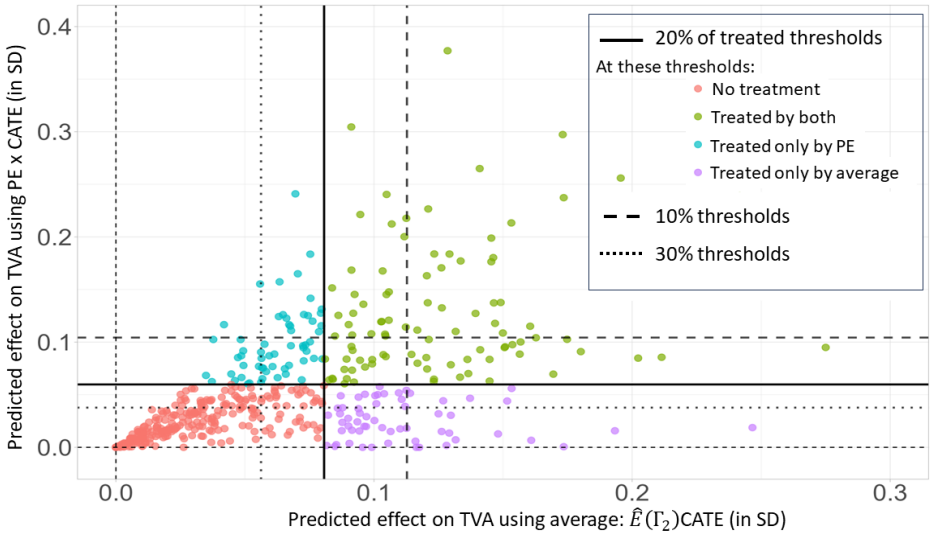
(a) Using the GT estimator



(b) Using the GT estimator and IV

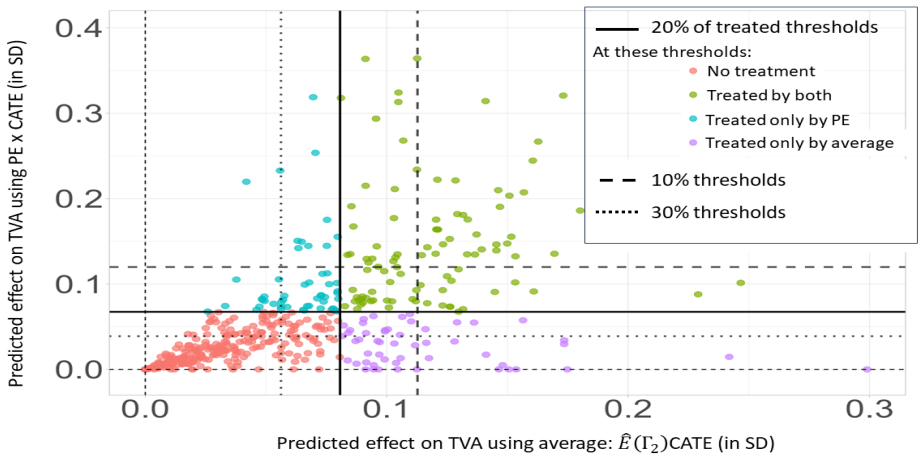
Notes: These results pool teachers from private and public schools. Figure 7(a) (resp. 7(b)) present the estimated joint distribution of the part of the TVA which can not be explained by teacher's knowledge (resp. the estimated TVA) and the estimated effect of teacher knowledge on TVA (Γ_2). This is using the GT estimator with varying coefficients A3 (ii), when I instrument for the teacher's mean score in the first tested year with the mean score of the second year, which reduces the sample size to 834. The dots represent the individual predictions $\widehat{\text{PE}}(X_i, Y_i)$ and the contour lines the levels of the associated fitted density. The dotted red line represents the IV estimates with an homogeneous specification (0.239). Teachers' tests scores are winsorized at a 1% level.

Figure 7: Joint distributions of the coefficients characterizing the TVA

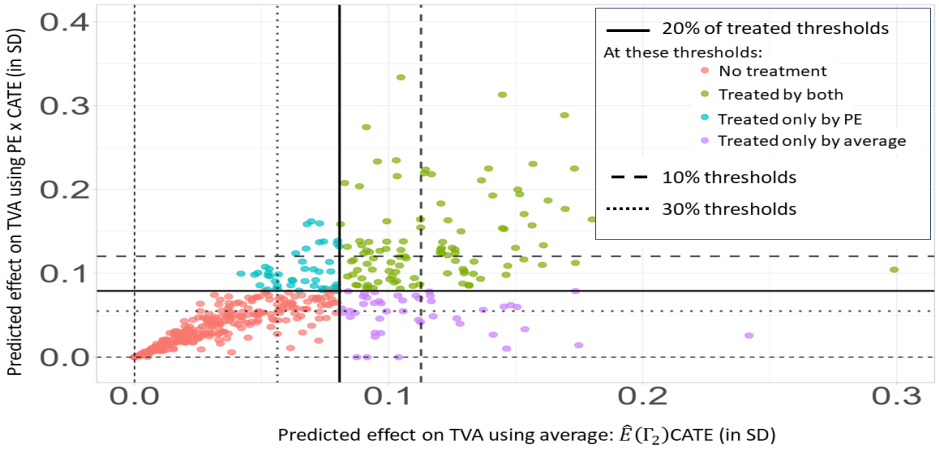


Notes: These results present the predicted effects based on the CATE only $\mathbb{E}(\Gamma_{2,i})\text{CATE}(X_{i,t})$ versus the predicted effects based on PE also $\text{PE}(X_{i,t}, Y_{i,t})$ $\text{CATE}(X_{i,t})$, which forms the optimal decision rule. The two plain black lines (resp. dotted and dashed) represent the threshold above which teachers would be allocated to such a program when treating 20% (resp. 10% and 30%) of the population. In this experiment, individuals represented in green (resp. in red) would be treated (resp. not treated) by both selection rules. However, the optimal policy would treat the individuals with strong predicted effect of knowledge on their performances displayed in blue, and does not treat individuals displayed in purple. Estimation is performed using the GT estimator under the independence assumption 1.

Figure 8: Comparison of the two decisions rules based on CATE, or the PE $\times$ CATE



(a) Using the GT estimator



(b) Using the GT estimator and IV

Notes: These results present the predicted effects based on the CATE only $\mathbb{E}(\Gamma_{2,i})\text{CATE}(X_{i,t})$ versus the predicted effects based on PE also $\text{PE}(X_{i,t}, Y_{i,t})$ $\text{CATE}(X_{i,t})$, which forms the optimal decision rule. The two plain black lines (resp. dotted and dashed) represent the threshold above which teachers would be allocated to such a program when treating 20% (resp. 10% and 30%) of the population. In this experiment, individuals represented in green (resp. in red) would be treated (resp. not treated) by both selection rules. However, the optimal policy would treat the individuals with strong predicted effect of knowledge on their performances displayed in blue, and does not treat individuals displayed in purple. Estimation is performed using the GWB estimator with varying coefficients A2 (ii) in 9(a) and independence A1 in 9(b).

Figure 9: Comparison of the two decisions rules based on CATE, or the PE $\times$ CATE

E The Tweedie formula and extension

Consider the model

$$Y_j = \alpha_j + \varepsilon_j, \tag{106}$$

α_j being independent from ε_j, $\alpha_j \sim F_\alpha$, $\varepsilon_j \sim F_\varepsilon$ being known. G can be either the true distribution of α_j in a frequentist setting where I assume "random effects", or a prior on the distribution in a Bayesian setting.

Assumption 15 *The distributions F_α and F_ε admit densities f_α and f_ε. Both f_α and $t \mapsto t f_\varepsilon(t)$ are square integrable on $\mathbb{R}$. φ_ε only vanishes on sets of null measure.*

Theorem 3 *Under assumption 15, the estimator of α_j that minimizes the Bayes risk under $\mathcal{L}_2$ loss, i.e the posterior mean of α_j conditional on Y_j takes the following form*

$$PE(y) = y + \frac{\mathcal{F}^{-1}\left[i\varphi_Y \varphi_\varepsilon' / \varphi_\varepsilon\right](y)}{f_Y(y)}. \tag{107}$$

First note that this estimator extends the so called "Tweedie formula" in the Empirical Bayes context, which for F_ε normal $\mathcal{N}(0, \sigma_\varepsilon^2)$ yields that

$$\mathrm{PE}(y) = y + \sigma_\varepsilon^2 \frac{f_Y'(y)}{f_Y(y)},$$

which we also directly gets from (107) as $\varphi_\varepsilon'/\varphi_\varepsilon = -\sigma_\varepsilon^2 t$ and $\mathcal{F}^{-1}\left[i\varphi_Y(\star)\star\right](y) = -f_Y'(y)$. If Robbins (1956) shows this results holds for F_ε belonging to an exponential family, (107) allows for more general error terms.

Proof of Theorem 3. Applying Bayes theorem yields

$$\mathbb{E}(\alpha | Y = y) = \frac{\int a p(y|a) dF_\alpha(a)}{\int p(y|a) dF_\alpha(a)},$$

where $p(y|a)$ is the conditional distribution of Y given α, hence, using $p(y|a) = f_\varepsilon(y-a)$ and $f_\alpha(y) = \int p(y|a) dF_\alpha(a)$,

$$PE(y) = y - \frac{\int (y-a) f_\varepsilon(y-a) dF_\alpha(a)}{f_Y(y)}.$$

We now rewrite $\int (y-a) f_\varepsilon(y-a) dF_\alpha(a)$ as a function of the observables. Consider now the characteristic function of the data, where using independence and Assumption 15, we have

$$\varphi_Y(t) = \varphi_\alpha(t)\varphi_\varepsilon(t). \tag{108}$$

We have under Assumption 15,

$$\begin{aligned}\int (y-a) f_\varepsilon(y-a) dF_\alpha(a) &= -\,i\left(f_\alpha \star \mathcal{F}^{-1}\left[\varphi_\varepsilon'\right]\right)(y) \\ &= -i\mathcal{F}^{-1}\left[\varphi_\alpha \varphi_\varepsilon'\right](y)\end{aligned}$$

hence with (108), assuming that φ_ε only vanishes on sets of null measure,

$$\int (y-a) f_\varepsilon(y-a) dF_\alpha(a) = \mathcal{F}^{-1}\left[\varphi_Y \frac{\varphi_\varepsilon'}{i\varphi_\varepsilon}\right](y),$$

which concludes the proof. □

F Additional Monte-Carlo simulations

F.1 With conditional independence and continuous Z

I consider the same DGP as in Section 4.3.2 but using a continuous control variable Z^* that is Beta$(2, 1.3)$ instead of a discretized version of it. The results are shown in Table 7. They reaching very similar conclusions as in the discrete case of Section 4.3.2. Important differences are that the GWB method with and without using Z performs relatively less well that in the discrete case. This is probably due to the fact that we have to discretize the variables to use it, which is more difficult in this context. Again, an important point is that although the Bayesian method actually performs better for a sample size of 1000 when Z is used, its errors remain nearly constant. On the contrary, the errors shrink for all my methods when using Z. In particular, even if Z is continuous, which yields theoretically slower rates of convergence for the nonparametric estimation, the GT method without constraint on Z ("GT") performs best at sample size 5000 and is really close to the Bayesian method for $n = 1000$ (0.047 and 0.076 (0.04 and 0.063, respectively) for the l^1 norm of Γ_1 and Γ_2).

Table 7: In-sample errors with conditional independence

	l^1 error				l^2 error			
	Γ_1		Γ_2		Γ_1		Γ_2	
	1000	5000	1000	5000	1000	5000	1000	5000
				Without Z				
Bayesian parametric	**0.076**	**0.077**	0.124	**0.126**	**0.088**	**0.09**	0.141	**0.143**
GT	0.081	0.085	0.126	0.129	0.115	0.123	0.169	0.173
GWB (disc. (X))	0.081	0.082	**0.087**	0.129	0.099	0.123	**0.132**	0.173
				With Z				
Bayesian parametric	**0.04**	0.039	**0.064**	0.063	**0.053**	0.053	**0.085**	0.083
GT varying	0.06	0.042	0.097	0.067	0.095	0.068	0.152	0.104
GT	0.047	**0.033**	0.076	**0.053**	0.068	**0.051**	0.106	**0.078**
GWB (disc. (X, Z))	0.076	0.056	0.151	0.116	0.105	0.076	0.21	0.16

Notes: in this 2 dimensional case, the in-sampled l^1 error is computed as $\sum_{i=1}^n |\widehat{\text{PE}}_k(X_i, Y_i, Z_i) - \text{PE}(X_i, Y_i, Z_i)|/n$ and the l^2 error as $(\sum_{i=1}^n (\widehat{\text{PE}}_k(X_i, Y_i, Z_i) - \text{PE}(X_i, Y_i, Z_i))^2/n)^{1/2}$, where $\widehat{\text{PE}}_k(X_i, Y_i, Z_i)$ are the different estimators. See the Appendix for non-sampled results and comparison to the true value of Γ. "Bayesian parametric" refers to King (1997) method with bivariate truncated normal prior, implemented in the R package ei. "GWB (disc. (X, Z))" refers to the GWB estimator where the distribution of (X, Z) has been discretized using the rule of Section 4.2.5. "GT varying" corresponds to the varying coefficients approach described in (23). The Monte-Carlo experiment uses 250 simulations.

F.2 Monte-Carlo simulations in the panel model

I consider simulations following the model of Section 3.2. More precisely, I consider the same DGP for X and Γ as the first one in Section 4.3.1, adding a standard normal random noise with $\sigma_\epsilon^2 = 0.01$. I compare my two estimators.

Table 8: In-sample errors in the panel model

	l^1 error				l^2 error					
	Γ_1		Γ_2		Γ_1		Γ_2		Comp. time	
	1000	5000	1000	5000	1000	5000	1000	5000	1000	5000
GT	0.084	0.062	0.137	0.105	0.109	0.083	0.184	0.139	9.47	10.23
GWB (disc. X)	**0.059**	**0.057**	**0.084**	**0.082**	**0.078**	**0.076**	**0.098**	**0.096**	30	51

Notes: in this 2×2 case, the in-sampled l^1 error is computed as $\sum_{i=1}^{n} |\widehat{\mathrm{PE}}_k(X_i, Y_i, Z_i) - \mathbb{E}[\Gamma_k|(X, Y, Z) = (X_i, Y_i, Z_i)]|/n$ and the l^2 error as $(\sum_{i=1}^{n}(\widehat{\mathrm{PE}}_k(X_i, Y_i, Z_i) - \mathbb{E}[\Gamma_k|(X, Y, Z) = (X_i, Y_i, Z_i)])^2/n)^{1/2}$, where $\widehat{\mathrm{PE}}_k(X_i, Y_i, Z_i)$ are the different estimators. "Comp. time" refers to computational time for estimation for one simulation. The Monte-Carlo experiment uses 250 simulations.

G Nonparametric Ecological inference

For a vector x of size d, denote by $\underline{x}$ the vector of size $d-1$, containing the first $d-1$ entries of x.

G.1 Application to the ecological inference model

A common and related empirical problem is to observe a sample of marginal distributions of two individual discrete variables $C_j \in \{1, 2\}$ and $R_j \in \{1, \ldots, d_R\}$ over the same groups of individuals i, while the distributions of C_j conditional on R_j for the different groups remain unknown. I consider a binary variable C_j here for simplicity, but the more general case is treated in the Appendix G.3. A simple but striking illustration is the probability of voting by race R_j for given precincts i. In this example, the precincts correspond to the groups, and the conditional probabilities are usually unobserved. Nevertheless, one can combine the margins over the precincts. Here, the margins are the turnout rates and the racial composition of each precinct, respectively.

$$Y_i := \begin{pmatrix} \mathbb{P}_i(C_j = 1) \\ \mathbb{P}_i(C_j = 2) \end{pmatrix} \in [0,1]^2 \quad \text{and} \quad X_i := \begin{pmatrix} \mathbb{P}_i(R_j = 1) \\ \vdots \\ \mathbb{P}_i(R_j = d_R) \end{pmatrix} \in [0,1]^{d_R}.$$

The former is provided by the election returns while the later is provided from the census. The conditional distributions Γ_i, or equivalently – as the margins are known – the contingency tables, are matrices with d_R rows and 2 columns whose coefficients

are the outcome probabilities conditional on the covariate for group g, namely $\Gamma_{r,c,i} := \mathbb{P}_i(C_j = c|R_j = r)$.

A common point of view used in political economy and statistics (see, *e.g.*, King, 1997; Wakefield, 2004; Imai et al., 2008), is to treat the observed sample of margins for the groups, together with the unobserved and heterogeneous conditional distributions, as random vectors and matrices drawn from a sampling distribution

$$(\Gamma_i, X_i, Y_i) \sim \mathbb{P}_{\Gamma,X,Y}.$$

As shown in King (1997), the law of total probability (8) yield that (Γ, X, Y) satisfies exactly a system of 2 linear RCs equations, which together with the constraints on the margins $X^\top 1 = Y^\top 1 = 1$, yields that we can focus on the first component $Y_{1,i}$:

$$Y_{1,i} = \sum_{r=1}^{d_R} \Gamma_{r,1,i} X_{r,i}, \ \forall r = 1, \ldots, d_R, \quad \Gamma_{r,1,i} \geq 0, \ X_{r,1,i} \geq 0, \ \sum_{r=1}^{d_R} X_{r,i} = 1. \tag{109}$$

In this context, it is first simply a matter of rewriting to obtain similar expressions for the posterior effects than the ones in Proposition 1. For the sake of completeness, this is done in Appendix G.2. I refer to Appendix G.3 for the more general case of nonbinary variable C.

Remark 1 (Identification with more than two possible outcomes) *Appendix G.3 studies partial identification with more than two possible outcomes. Proposition 15 shows that the elements of m are solutions of a system of coupled transport partial differential equations. If one limits the dimension of the unobserved heterogeneity, Appendix G.5 then provides a way to solve this system and recover point identification with 3 outcomes, which can be extended to more. In particular, I assume that some random coefficients are linearly dependent on the others. I show that here also the posterior effects can be expressed directly as function of the data.*

G.2 Results completing Section G.1 with 2 choices

I consider the context of Section G.1.

Assumption 16 *Assume that*

1. *the heterogeneous conditional probabilities are independent of the shares of the different categories across groups, namely $\Gamma \perp\!\!\!\perp X$;*

2. *The support of* $\underline{X}$ *has nonempty interior;*

3. *The conditional density* $f_{Y_1|\underline{X}}$ *exists and, for all* $l = 1, \ldots, d_R - 1$ *and* $x \in \text{Supp}(\underline{X})$*, its partial derivatives* $\partial_{x_l} f_{Y_1|\underline{X}}(\cdot|x)$ *are integrable and square integrable on* $\mathbb{R}$*.*

Proposition 13 gives the counterpart of the GT formulation in this context of ecological inference, precising explicitly how results of Section 3.1 apply here.

Proposition 13 *Let the distribution of* (Γ, X, Y) *satisfy* (109) *and make Assumption 16. Then, the prediction* m *is point identified and satisfies, for all* $r = 1, \ldots, d_R$ *and* $(x, y) \in \text{Supp}(X, Y_1)$*,*

$$PE_r(x, y) = y + \sum_{l=1}^{d_R - 1} (x_l - \mathbb{1}\{l = r\}) \frac{\partial_{x_l} F_{Y_1|\underline{X}}(y|\underline{x})}{f_{Y_1|\underline{X}}(y|\underline{x})}. \tag{110}$$

Assumption 16-1 is called the no contextual effects assumption in the ecological inference literature (NCE hereafter).

G.3 Extension to identification in Ecological inference with more than two choices

As shown in King (1997), the law of total probability (8), together with the constraints on the margins, yield that (Γ, X, Y) satisfies exactly the linear system of random coefficients equations

$$\forall c = 1, \ldots, d_C, \; Y_c = \sum_{r=1}^{d_R} \Gamma_{r,c} X_r, \quad \forall r = 1, \ldots, d_R, \quad \sum_{c=1}^{d_C} \Gamma_{r,c} = 1 \tag{111}$$

$$\forall c = 1, \ldots, d_C, \forall r = 1, \ldots, d_R, \quad \Gamma_{r,c} \geq 0, \quad X_{r,c} \geq 0, \quad \sum_{r=1}^{d_R} X_r = 1 \quad . \tag{112}$$

The system (111) is a particular type of seemingly unrelated regressions (SUR) with random coefficients which contain a common regressor, with additional constraints $X^\top \mathbf{1} = Y^\top \mathbf{1} = 1$. I now consider the following exogeneity assumption which constrains the dependence between the regressor and the random coefficients.

Assumption 17 ("No contextual effects" (NCE)) *Assume that the heterogeneous conditional probabilities are independent of the shares of the different categories across groups, namely:*

$$\Gamma \perp\!\!\!\perp X.$$

Assumption 17 is classical both in the random coefficients and in the ecological inference litteratures. This nonparametric assumption is however strong for some applications (see, *e.g.*, Tam Cho, 1998) hence the need to perform sensitivity analysis to the predictions obtained under this assumption. In assumptions 6 or 7, I consider alternative assumptions when other covariates are available. For a vector r of size d, denote by $\underline{r}$ the vector of size $d-1$, containing the first $d-1$ entries of r.

Assumption 18 *The support of $\underline{X}$ has nonempty interior.*

I maintain Assumption 18 for simplicity. Note that, because $\underline{r}$ are probabilities, this latter assumption is not restrictive in most applications. Support conditions on the regressors in this context are relaxed in Theorem 5 in Gaillac and Gautier (2022) and I could allow for discrete regressors whose support is countably infinite.

Definition of the identified set for the PE. I explicit here useful elements of nonparametric identification (see, *e.g.*, Matzkin, 2007). The distribution of the observables is $\mathbb{P}_{X,Y}$, while the distribution of the observables generated by $\mathbb{P}_{\Gamma,X}$ and the system (111)-(112) is $\mathbb{P}^{gen}(\mathbb{P}_{\Gamma,X})$. $\mathcal{R}$ is a set of restrictions defined accordingly, like satisfying the independence restriction $\mathbb{P}_{\Gamma,X} = \mathbb{P}_{\Gamma} \otimes \mathbb{P}_X$. The functional of interest is

$$\mathrm{PE}_{r,c} : (x,y) \mapsto \mathbb{E}[\Gamma_{r,c} | (X,Y) = (x,y)], \quad r = 1, \dots, d_R, \ c = 1, \dots, d_C,$$

and satisfies $m = \Lambda(\mathbb{P}_{\Gamma,X}, \mathbb{P}_{X,Y})$ for a certain deterministic function Λ.[15] The identified set for m is the set of matrix valued functionals such that there exists a unobserved associated distribution $\mathbb{P}_{\Gamma,X}$ which generates observations compatible with the distribution of the data,

$$\mathcal{J}_{X,Y}(\Lambda, \mathcal{R}) := \{\mathrm{PE} : \ \exists \, \mathbb{P}_{\Gamma,X} \in \mathcal{R}, \quad \mathbb{P}^{gen}(\mathbb{P}_{\Gamma,X}) = \mathbb{P}_{X,Y}, \quad \Gamma(\mathbb{P}_{\Gamma,X}, \mathbb{P}_{X,Y}) = \mathrm{PE}\}.$$

[15] It is detailed in the proofs, using Bayes' theorem and that the conditional distribution of Y given Γ, X is fixed by (111)-(112).

It is shown in Corollary 1 in Masten (2017) in the context of SUR that the joint distribution of Γ is necessarily not point identified (see Proposition (P14.a) below). Proposition (P14.b) below is new and shows that, with more than two choices, even the conditional expectation of the random coefficients is not identified without additional assumptions on the random matrix. When $d_C = 2$, because in my model the distribution of Γ is compactly supported $\text{Supp}(\Gamma_{\cdot,1}) \subseteq [0,1]^{d_R}$, then Proposition (P14.2) below is Proposition 2.2 in Beran and Millar (1994) and (P14.1) is a direct consequence of it.

Proposition 14 (Identification without contextual effects) *Let the distribution of (Γ, X, Y) satisfy (8) and Assumption 17. I have, for all $d_R \geq 2$, when $d_C = 2$,*

(P2.1) PE is identified under Assumption 18;

(P2.2) the distributions of Γ and of Γ conditional on (X, Y) are identified under Assumption 18;

and, when $d_C > 2$,

(P2.a) the distribution of Γ is not identified under Assumption 18;

(P2.b) PE is not identified under Assumption 18.

In Proposition 14 and under Assumption 18, I use the fact that the support of Γ is compact, hence the distribution is determined by its moments. Theorem 13 below goes further than the nonidentification result of (P2.b) with partial identification results and also showsn nonparametric constructive point identification in the case $d_C = 2$. Note that many classical parametric distributions of Γ yield that Assumption 3 holds, such as the uniform distribution, the truncated normal used by King (1997), the beta or the Dirichlet distributions with parameter strictly greater than one or the logit-normal distribution.

G.4 Partial identification when $d_C > 2$

Proposition 15 (Partial identification, $d_C > 2$) *Let the distribution of (Γ, X, Y) satisfy (8) and define the restriction $\mathcal{R}_0$ corresponding to assumptions 17, 18, and 3. Let $d_C > 2$, then $\mathcal{J}_{X,Y}(\Lambda, \mathcal{R}_0)$ is included into the set of functions of the form*

PE $= M/f_{\underline{Y}|\underline{X}}$, *where* $M_{r,c} : \text{Supp}(\underline{X}, \underline{Y}) \mapsto [0,1]$ *for* $r = 1, \ldots, d_R$ *and* $c = 1, \ldots, d_C$ *are continuous functions which admit a continuous derivative with respect to* y_c, *for* $c = 1, \ldots, d_C - 1$, $M_{r,C} = 1 - \sum_{c=1}^{d_C-1} M_{r,c}$, *and, for all* $r = 1, \ldots, d_R$, $c = 1, \ldots, d_C - 1$, *and* $(x, y) \in \text{Supp}(\underline{X}, \underline{Y})$,

$$\sum_{r=1}^{d_R-1} x_r M_{r,c}(x,y) + (1 - x'\mathbf{1}) M_{d_R,c}(x,y) = \rho_c(x,y), \tag{113}$$

$$\sum_{c=1}^{d_C-1} \partial_{y_c} M_{r,c}(x,y) = \sum_{c=1}^{d_C-1} \partial_{y_c} \rho_c(x,y) + \sum_{l=1}^{d_R-1} (x_l - \mathbb{1}\{l = r\}) \partial_{x_l} f_{\underline{Y}|\underline{X}}(y|x), \tag{114}$$

where $\rho_c(x,y) := f_{\underline{Y}|\underline{X}}(y|x) y_c$. *Moreover, for all* $c = 1, \ldots, d_C - 1$ *and* $(x,y) \in \text{Supp}(\underline{Y}, \underline{X})$, $M_{r,c}(x, y_1, \ldots, y_c = 0, \ldots, y_{d_C-1}) = 0$.

Proposition 15 shows that, when $d_C > 2$, the parameter of interest satisfies a system of coupled partial differential equations. However, the solutions are in general not unique nor explicit.

G.5 Identification when $d_C = 3$ when restricting the dimension of the unobserved heterogeneity

I consider the case where the researcher assumes that some random coefficients are linearly dependent of the others. This reduces the dimension of the unobserved heterogeneity, hence reducing the size of the identified set when we have more than two choices. I consider the case $d_C = 3$ and describe in Remark 4 the set of assumptions that one would make to handle higher dimensional cases.

Assumption 19 (Restricted heterogeneity, $d_C = 3$**)** *Let* ω *be a sequence of length* $d := d_R + d_C - 2 = d_R + 1$ *of indexes,* $\omega := \big((r,1)_{r \in \{1,\ldots,d_R\}}, (d_R, 2)\big)$. *Let the* d *coefficients* $(\Gamma_{\omega_k})_{k=1,\ldots,d}$ *be the latent unobserved heterogeneity, that I denote by* $U := (U_1, \ldots, U_d)$, *hence*

$$U_l := \Gamma_{\omega_l}, \quad l = 1, \ldots, d.$$

The $(d_R - 1)$ *other random coefficients can be expressed as*

$$\Gamma_{r,2} = \sum_{k=1}^{d} a_{r,k} U_k, \quad r = 1, \ldots, d_R - 1,$$

where $a \in M_{d_R-1,d}(\mathbb{R})$ *are fixed coefficients.*

Remark 2 (More general formulation) *A slightly more general formulation would assume instead of Assumption 19 that these are d latent sources of random unobserved heterogeneity U, and that the coefficients $\Gamma_{r,c}$ depend linearly of U. However, the simplified set up that I consider is more transparent, facilitates testing and estimation of a, and amounts to the same type of assumptions.*

Note that in the case of $d_C = 2$, Assumption 19 is not a restriction as $(d_R-1)(d_C-2) = 0$, which is in line with Theorem 13. This yields for $d_C = 3$,

$$Y_1 = \sum_{r=1}^{d_R} U_r X_r \tag{115}$$

$$Y_2 = \sum_{r=1}^{d_R-1} \sum_{k=1}^{D} a_{r,2,k} U_k X_r + U_{d_R+1} X_{d_R}. \tag{116}$$

Assumption 17 yields the system of equations

$$\mathbb{E}\left[Y_1 | X = x\right] = \sum_{r=1}^{d_R-1} \left(\mathbb{E}\left[U_r\right] - \mathbb{E}\left[U_{d_R}\right]\right) x_r + \mathbb{E}\left[U_{d_R}\right]$$

$$\mathbb{E}\left[Y_2 | X = x\right] = \sum_{r=1}^{d_R-1} \left(\sum_{k=1}^{d_R+1} a_{r,k} \mathbb{E}\left[U_k\right] - \mathbb{E}\left[U_{d_R+1}\right] \right) x_r + \mathbb{E}\left[U_{d_R+1}\right].$$

This yields using Assumption 18 with $d = d_R + 1$ that $\mathbb{E}\left[U_k\right]$ for $k = 1, \ldots, d_R + 1$ and $v_r := \sum_{k=1}^{d_R+1} a_{r,k} \mathbb{E}\left[U_k\right]$ and $r = 1, \ldots, d_R - 1$ are identified. Thus, I obtain a system of $d_R - 1$ equations and $(d_R - 1)(d_R + 1)$ unknowns coefficients $a_{r,k}$. If a is known, then Proposition 16 below shows point identification in a constructive way. Otherwise, Proposition 16 describes the identified set.

Let me introduce some notations. Under Assumption 19 and (115)-(116) I obtain, for $c = 1, 2$, ($c = 3$ being redondant with the others due to the constraint $Y^\top \mathbf{1} = 1$, I suppress it),

$$y_c = \sum_{k=1}^{d} W_{c,k}(x) \mathbb{E}\left[U_k | X = x, Y = y\right], \tag{117}$$

where $W_{1,k}(x) := x_k \mathbb{1}\{k \leq d_R\}$ for $k = 1, \ldots, d$ and

$$W_{2,k}(x) := \sum_{r=1}^{d_R-1} a_{r,k} x_r + \mathbb{1}\{k = d_R + 1\} x_{d_R}. \tag{118}$$

For convenience, I use $V_k : (x, y) \mapsto \mathbb{E}\left[U_k | (X, Y) = (x, y)\right] f_{\underline{Y}|\underline{X}}(\underline{y}|\underline{x})$, for $k = 1, \ldots, d$ and $M_{r,c} : (x, y) \mapsto \mathbb{E}\left[\Gamma_{r,c} | X = x, \underline{Y} = \underline{y}\right] f_{\underline{Y}|\underline{X}}(\underline{y}|\underline{x})$, for $r = 1, \ldots, d_R$, $c = 1, 2$.

Identification strategy. Let me explain the steps of the identification strategy:

(**Step** 1) I can express the coefficients of M in terms of V through

$$\begin{aligned} M_{r,1}(x,y) &= V_r(x,y) \\ M_{r,2}(x,y) &= \sum_{k=1}^{d_R+1} a_{r,k} V_k(x,y), \end{aligned} \tag{119}$$

for $r = 1, \ldots, d_R$. Hence the aim is to recover V.

(**Step** 2) I express V_l for $l = d_R, d_R+1$ as function of V_l for $l = 1, \ldots, d_R-1$. Denote by $\rho_c(x,y) := f_{\underline{Y}|\underline{X}}(\underline{y}|\underline{x})y_c$, for $c = 1, 2$. Under Assumption 20.1 below and using (117), the system with $d - d_R + 1 = 2$ unknowns, $V_R(x,y), V_{d_R+1}(x,y)$,

$$x_{d_R} V_{d_R}(x,y) = \rho_1(x,y) - \sum_{k=1}^{d_R-1} x_k V_k(x,y) \tag{120}$$

$$\begin{aligned} W_{2,d_R}(x) V_{d_R}(x,y) + W_{2,d_R+1}(x) V_{d_R+1}(x,y) \\ = \rho_2(x,y) - \sum_{k=1}^{d_R-1} W_{2,k}(x) V_k(x,y), \end{aligned} \tag{121}$$

has a unique solution, for $l = d_R, d_R + 1$,

$$V_l(x,y) = \sigma_{l-d_R+1}(x,y) + \sum_{k=1}^{d_R-1} Q_{l-d_R+1,k}(x) V_k(x,y), \tag{122}$$

where, for $k = 1, \ldots, d_R - 1$,

$$\sigma_1(x,y) = \frac{\rho_1(x,y)}{x_{d_R}}, \; Q_{1,k}(x) = -\frac{x_k}{x_{d_R}}, \tag{123}$$

$$\sigma_2(x,y) = \frac{x_{d_R}\rho_2(x,y) - W_{2,d_R}(x)\rho_1(x,y)}{x_{d_R} W_{2,d_R+1}(x)}, \tag{124}$$

$$Q_{2,k}(x) = \frac{x_{d_R} W_{2,k}(x) - x_k W_{2,d_R}(x)}{x_{d_R} W_{2,d_R+1}(x)}. \tag{125}$$

(**Step** 3) Then, I identify V_l for $r = 1, \ldots, d_R - 1$ as solution of a system of coupled partial transport differential equations, see the proof of Proposition 16 for details.

Denote by $\widetilde{Q} \in \mathcal{M}_{d_R-1,d_R-1}(\mathbb{R})$ with coefficients $\widetilde{Q}_{r,k}(x) := a_{r,k} + \sum_{l=d_R}^{d_R+1} a_{r,l} Q_{l-d_R+1,k}(x)$, for $r = 1, \ldots, d_R - 1$ and $k = 1, \ldots, d_R - 1$.

Assumption 20 *When $d_C = 3$, for all $x \in \text{Supp}(X)$,*

1. $x_{d_R} W_{2,d_R+1}(x) \neq 0$;
2. $\widetilde{Q}(x) \in \mathcal{M}_{d_R-1,d_R-1}(\mathbb{R})$ *is diagonalisable:* $\widetilde{Q}(x) = P^{-1}(x)\,diag\,(\Lambda(x))\,P(x)$, *where* $diag\,(\Lambda(x))$ *is a diagonal matrix and* $P(x)$ *is an orthogonal matrix.*

Proposition 16 *Consider $d_C = 3$. Let the distribution of (Γ, x, y) satisfy (8) and define the restriction $\mathcal{R}_1$ corresponding to assumptions 17, 18, 3, 19, and 20. Then $\mathcal{J}_{x,y}(\Lambda, \mathcal{R}_1)$, the identified set for PE is included into the set of functions taking the form, for all $r = 1, \ldots, d_R$, $c = 1, 2$, and $(x, y) \in \text{Supp}(X, Y)$,*

$$PE_{r,c}(x, y) = \Pi_{r,c}\left[\zeta, \sigma\right](x, y), \tag{126}$$

Π is a linear operator from $\mathcal{M}_{d\times(d_R-1)}\left(l^\infty\left(\text{Supp}(X, Y)\right)\right) \times \mathcal{M}_{2\times 1}\left(l^\infty\left(\text{Supp}(X, Y)\right)\right)$ to $\mathcal{M}_{d_R,d_C}(l^\infty\left(\text{Supp}(X, Y)\right))$,

$$\Pi_{r,1}\left[\zeta, \sigma\right] = P^{-1} Diag\left(PK\zeta\right), \quad r = 1, \ldots, d_R - 1 \tag{127}$$

$$\Pi_{d_R,1}\left[\zeta, \sigma\right] = Q_{1,.}^{\top} P^{-1} Diag\left(PK\zeta\right) + \sigma_1, \tag{128}$$

$$\Pi_{r,2}\left[\zeta, \sigma\right] = \widetilde{Q}_{r,.}^{\top} P^{-1} Diag\left(PK\zeta\right) + a_{r,d_R}\sigma_1 + a_{r,d_R+1}\sigma_2, \quad r = 1, \ldots, d_R, \tag{129}$$

where, σ is defined via (123)-(125),

$$K(x) = \begin{pmatrix} x_1 - 1 & \ldots & x_{d_R-1} & 1 & 1 \\ x_1 & x_2 - 1 & x_{d_R-1} & 1 & 1 \\ \vdots & \ddots & & 1 & 1 \\ x_1 & & x_{d_R-1} - 1 & 1 & 1 \end{pmatrix} \tag{130}$$

and where $\zeta \in \mathcal{M}_{d\times(d_R-1)}\left(l^\infty\left(\text{Supp}(X, Y)\right)\right)$ with $\zeta(x, y)/f_{\underline{Y}|\underline{X}}(\underline{y}|\underline{x})$ equals to

$$\begin{pmatrix} \int_0^{y_1} \partial_{x_1} f_{\underline{Y}|\underline{X}}(v, y_2 - \Lambda_1(x)(y_1 - v)|\underline{x})dv & \ldots & \ldots \\ \vdots & & \vdots \\ \int_0^{y_1} \partial_{x_{R-1}} f_{\underline{Y}|\underline{X}}(v, y_2 - \Lambda_1(x)(y_1 - v)|\underline{x})dv & & \vdots \\ f_{\underline{Y}|\underline{X}}(\underline{y}|\underline{x})y_1 & \ldots & f_{\underline{Y}|\underline{X}}(\underline{y}|\underline{x})y_1 \\ \int_0^{y_1} \partial_{y_2}\rho_2(x, v, y_2 - \Lambda_1(x)(y_1 - v))dv & \ldots & \int_0^{y_1} \partial_{y_2}\rho_2(x, v, y_2 - \Lambda_{R-1}(x)(y_1 - v))dv \end{pmatrix}, \tag{131}$$

where, $\rho_c(x, y) := f_{\underline{Y}|\underline{X}}(\underline{y}|\underline{x})y_c$. When $a \in \mathcal{M}_{d_R-1,d}(\mathbb{R})$ in Assumption 19 is known, then this set is reduced to one element.

Remark 3 ($d_C = 2$ as particular case) *Using Proposition 13, the case $d_C = 2$ appears as a particular case where no further assumption has to be made on the random coefficients to obtain point identification. When $d_C = 2$, (110) can be rewritten as*

$$PE_k = \Pi_{r,1}\left[\zeta\right] := (K\zeta)_r,$$

Π is a linear operator from $\mathcal{M}_{d\times(d_R-1)}\left(l^\infty\left(\text{Supp}(X,Y)\right)\right)$ to $\mathcal{M}_{d_R,d_C}(l^\infty\left(\text{Supp}(X,Y)\right))$, and K is defined like (130) with only one column of 1 and, for all $(x,y) \in \text{Supp}(X,Y)$,

$$\zeta(x,y) := \left(\frac{\partial_{x_1} F_{\underline{Y}|\underline{X}}(y_1|\underline{x})}{f_{\underline{Y}|\underline{X}}(\underline{y}|\underline{x})}, \ldots, \frac{\partial_{x_{d_R-1}} F_{\underline{Y}|\underline{X}}(y_1|\underline{x})}{f_{\underline{Y}|\underline{X}}(y_1|\underline{x})}, y_1\right)^\top \tag{132}$$

and (126) is (110), hence this set is also reduced to one element. Proposition 16 shows that cases $d_C = 2$ and $d_C = 3$ share a similar structure, where the components needing to be estimated nonparametrically are all the elements of ζ.

The proof of Proposition 16 is constructive and one can directly employ a plug-in approach using an estimator of ζ defined in (132) for $d_C = 2$, estimating (131) for $d_C = 3$. Indeed, Proposition 16 and Remark 3 underline that in cases $d_C = 2$ and $d_C = 3$, one has to nonparametrically estimate the elements of ζ, as my parameter of interest is the image by the linear operator Π (which is also bounded under Assumption 20) of ζ and σ. This assume estimation of

$$(x,y) \mapsto \int_0^{y_1} \partial_{x_l} f_{\underline{Y}|\underline{X}}(v, y_2 - \Lambda_r(x)(y_1 - v))dv$$

for $r = 1, \ldots, d_R - 1$ and $l = 1, \ldots, d_R - 1$, while the other components of ζ in Proposition 16 imply estimating also $f_{\underline{Y}|\underline{X}}$, $f_{\underline{X}}$, and $\partial_{y_c} f_{\underline{Y}|\underline{X}}$ (for $d_C = 3$ only), when these quantities exist.

Remark 4 (Cases $d_C > 3$) *Using a similar reasoning as in the proof of Proposition 16, one could handle nonparametrically the cases $d_C > 3$, assuming that the matrices $\widetilde{Q}^c(x)$ which appear in the system of coupled differential equations all commute by pairs for $c = 1, \ldots, d_C - 1$ (or equivalently that they are diagonalisable in the same basis), which puts more restrictions on the coefficients a. I left this for future research.*

G.6 Proofs of Appendix G.3

Notations and Preliminaries. For notational simplicity, we denote the multivariate Fourier transform of measures on the set of matrices $\mathfrak{M}(\mathcal{M}_{d_R,d_C}(\mathbb{R}))$ by

$$\mathcal{F}[\mu](x) = \int_{\mathcal{M}_{d_R,d_C}(\mathbb{R})} e^{i<g,x>} d\mu(g), \tag{133}$$

where $< g, x >= \text{Tr}(g^\top x) = \sum_{r=1}^{d_R}\sum_{c=1}^{d_C} g_{r,c}x_{r,c}$ is the inner product between matrices and Tr is the trace operator. This notation is simply a compact way to denote the multivariate Fourier transform, but one could also fix a way to vectorise the matrix and use the usual multivariate Fourier transform. I denote by $\underline{\Gamma}$ the submatrix of Γ keeping only the (d_C-1) first columns, hence of dimension $d_R \times (d_C-1)$.

Proof of Proposition 14. Start with the proof of (P2.2). Using $Y_1 + Y_2 = 1$, the first part of the proof is Proposition 2.2 in Beran and Millar (1994). The second part of (P2.2) can be deduced from the first one using the Bayes' theorem, (111)-(112), and Assumption 17 which yield, for all $(g, x, y) \in \mathcal{M}_{d_R\times(d_C-1)}([0,1])\times\text{Supp}(X)\times\text{Supp}(\underline{Y})$, $\mathbb{P}_{\underline{\Gamma}|X,\underline{Y}}(g|x,y) = \mathbb{1}\{y = g^\top x\}\mathbb{P}_{\underline{\Gamma}}(g)/\mathbb{P}_{\underline{Y}|X}(y|x)$.
The proof of (P2.a) is a consequence of Corollary 1 in Masten (2017) once we have used $Y^\top\mathbf{1} = 1$ to consider only equations related to $c = 1,\dots,d_C-1$ in (111)-(112). Let us now prove (P2.1) and (P2.b). Let $d_C, d_R \geq 2$. Using the constraint $Y\mathbf{1} = 1$, we consider the first d_C-1 equations in (111)-(112) because the last one can be deduced from the others. Hereafter $\underline{\Gamma}$ is thus a $d_R \times (d_C-1)$ random matrix with the d_C-1 first columns of Γ. We have, using Bayes' theorem for the second equality, for a.e. $(x,y) \in \text{Supp}(X,Y)$ and for all $r = 1,\dots,d_R$, $c = 1,\dots,d_C-1$,

$$\begin{aligned}
\mathbb{E}\left[\underline{\Gamma}_{r,c}|X = x, \underline{Y} = \underline{y}\right] &= \int_{\mathcal{M}_{d_R\times(d_C-1)}(\mathbb{R})} g_{r,c} d\mathbb{P}_{\underline{\Gamma}|X,\underline{Y}}(g|x,\underline{y}) \\
&= \int_{\mathcal{M}_{d_R\times(d_C-1)}(\mathbb{R})} g_{r,c}\frac{\mathbb{P}_{\underline{Y}|\underline{\Gamma},X}(\underline{y}|g,x)}{\mathbb{P}_{\underline{Y}|X}(\underline{y}|x)} d\mathbb{P}_{\underline{\Gamma}|X}(g|x) \\
&= \int_{g\in\mathcal{I}(x,y)} \frac{g_{r,c}}{\mathbb{P}_{\underline{Y}|X}(\underline{y}|x)} d\mathbb{P}_{\underline{\Gamma}}(g) \text{ (using Assumption 17)},
\end{aligned} \tag{134}$$

where $\mathcal{I}(x,y)$. When $d_C = 2$, under assumptions 17 and 18, using (P2.2), $\mathbb{P}_\Gamma$ is identified. Thus, we directly have from (134) that $(x,y) \mapsto \mathbb{E}[\Gamma|X = x, Y = y]$ is also identified.

Consider now the case $d_C > 2$. For simplicity, we condider the case $d_C = 3$ and $d_R = 2$, as the other cases can be adapted from it. Take $f^1_{\underline{\Gamma}}$ as, for all $g \in \mathcal{M}_{2,2}([0,1])$,

$$f^1_{\underline{\Gamma}}(g) = \frac{1}{\mathcal{Z}} \prod_{r=1}^{2} \prod_{c=1}^{2} \mathbb{1}\{g_{r,c} \in [0,1]\} g_{r,c},$$

where $\mathcal{Z}$ is a normalisation constant. Consider a second distribution, for all $g \in \mathcal{M}_{2,2}([0,1])$,

$$f^2_{\underline{\Gamma}}(g) := f^1_{\underline{\Gamma}}(g) + \delta \mathbb{1}\{g \in \mathcal{M}_{2,2}([0,1])\} \left(\partial_{11}\partial_{22} - \partial_{12}\partial_{21}\right) f^1_{\underline{\Gamma}}(g),$$

where δ is such that $f^2_{\underline{\Gamma}}(g) \geq 0$ for all $g \in \mathcal{M}_{2,2}([0,1])$. Note that we have

$$\frac{1}{\mathcal{Z}} \int_{\mathcal{M}_{d_R \times (d_C - 1)}([0,1])} \left(g_{2,1}g_{1,2} - g_{1,1}g_{2,2}\right) dg = 0$$

hence $\int_{\mathcal{M}_{d_R \times (d_C-1)}([0,1])} f^2_{\underline{\Gamma}}(g) dg = 1$. This yields, for all $\boldsymbol{z} \in \mathcal{M}_{2,2}(\mathbb{R})$,

$$\mathcal{F}\left[f^2_{\underline{\Gamma}}\right](\boldsymbol{z}) = \left(1 - \delta(\boldsymbol{z}_{11}\boldsymbol{z}_{22} - \boldsymbol{z}_{12}\boldsymbol{z}_{21})\right) \mathcal{F}\left[f^1_{\underline{\Gamma}}\right](\boldsymbol{z}),$$

hence for all $t \in \mathbb{R}^2$, $x \in \mathrm{Supp}(X)$, $\mathcal{F}\left[f^2_{\underline{\Gamma}}\right](tx^\top) = \mathcal{F}\left[f^1_{\underline{\Gamma}}\right](tx^\top)$. Using Assumption 17, we have,

$$\mathbb{E}\left[e^{it^\top \underline{Y}} | X = x\right] = \mathcal{F}\left[f_{\underline{\Gamma}}\right](tx^\top)$$

hence $f^1_{\underline{\Gamma}}$ and $f^2_{\underline{\Gamma}}$ yield the same observables, while being distinct *a.e.*, on $\mathcal{M}_{2,2}([0,1])$. Consider, for example, the coefficient $(1,1)$ of $\underline{\Gamma}$. Then, using (134), we have, for all $(x,y) \in \mathrm{Supp}(X, \underline{Y})$,

$$\begin{aligned}
&\mathbb{E}_{\mathbb{P}^1}\left[\Gamma_{1,1} | X = x, \underline{Y} = y\right] - \mathbb{E}_{\mathbb{P}^2}\left[\Gamma_{1,1} | X = x, \underline{Y} = y\right] \\
&= \int_{g \in \mathcal{I}(x,y)} \frac{g_{1,1}}{\mathbb{P}_{\underline{Y}|X}(y|x)} \left(f^1_{\underline{\Gamma}} - f^2_{\underline{\Gamma}}\right)(g) dg, \\
&= \frac{\delta}{\mathcal{Z}\mathbb{P}_{\underline{Y}|X}(y|x)} \int_{g \in \mathcal{I}(x,y)} g_{1,1}\left(g_{2,1}g_{1,2} - g_{1,1}g_{2,2}\right) dg \\
&= \frac{\delta}{\mathcal{Z}\mathbb{P}_{\underline{Y}|X}(y|x)} \left(\int_0^1 b \frac{y_1 - bx_1}{x_2} db \int_0^1 \frac{y_2 - bx_2}{x_1} db - \int_0^1 \left(\frac{y_1 - bx_2}{x_1}\right)^2 db \int_0^1 \frac{y_2 - bx_1}{x_2} db \right) \\
&= \frac{\delta}{\mathcal{Z}\mathbb{P}_{\underline{Y}|X}(y|x)(x_1x_2)^2} \left(x_1x_2 \left(\frac{y_1}{2} - \frac{x_1}{3}\right)\left(y_2 - \frac{x_2}{2}\right) - \frac{1}{3}\left(y_1^3 - (y_1 - x_2)^3\right)\left(y_2 - \frac{x_1}{2}\right) \right)
\end{aligned}$$

and using Assumption 18, there exists a subset $\mathcal{S}$ of $\mathrm{Supp}(X, \underline{Y})$ with nonempty interior such that the right-hand-side is different from zero *a.e.* $(x,y) \in \mathcal{S}$, which yields the result (P4.b). □

Lemma 6 *Let $\mathbb{P}_\Gamma$ be a measure on $\mathcal{M}_{d_R,d_C}(\mathbb{R})$ satisfying (111)-(112). Then we have, for all $(x,y) \in \text{Supp}(X,Y)$,*

$$\int_{g \in \mathcal{I}(x,y)} g d\mathbb{P}_{\underline{\Gamma}}(g) = \mathcal{F}^{-1}\left[\mathcal{F}\left[\star \mathbb{P}_{\underline{\Gamma}}(\star)\right]\left(\cdot x^\top\right)\right](\underline{y}),$$

where the Fourier transform is defined in (133).

Proof of Lemma 6. First, using (134) we have, for all $(x,y) \in \text{Supp}(X,Y)$,

$$\mathbb{E}\left[\underline{\Gamma}|X=x,\underline{Y}=\underline{y}\right]\mathbb{P}_{\underline{Y}|X}(\underline{y}|x) = \int_{g \in \mathcal{I}(x,y)} g d\mathbb{P}_{\underline{\Gamma}}(g), \tag{135}$$

and that, for all $x \in \text{Supp}(X)$, $y \in \mathbb{R}^{d_C-1} \mapsto \mathbb{E}\left[\underline{\Gamma}|X=x,\underline{Y}=y\right]\mathbb{P}_{\underline{Y}|X}(y|x)$ is compactly supported in $[0,1]^{d_C-1}$. This yields that $y \in \mathbb{R}^{d_C-1} \mapsto \int_{g \in \mathcal{I}(x,y)} g d\mathbb{P}_{\underline{\Gamma}}(g)$ belongs to $L^1(\mathbb{R}^{d_C-1}) \cap L^2(\mathbb{R}^{d_C-1})$ hence its Fourier transform is well defined (see, *e.g.*, Theorem 9.13 in Rudin, 1973). Using the definition of $\mathcal{I}(x,y)$ for the second equality which yields that $g \in \mathcal{I}(x,y)$ if and only if $\underline{y} = (x^\top g)^\top$ where $g \in \mathcal{M}_{d_R \times (d_C-1)}([0,1])$, that $t^\top (x^\top g)^\top = \sum_{c=1}^{d_C-1} t_c (x^\top g)_c = \sum_{c=1}^{d_C-1} t_c \sum_{r=1}^{d_R} x_r g_{r,c} = \sum_{c=1}^{d_C-1}\sum_{r=1}^{d_R}(t_c x_r) g_{r,c} =< tx^\top, g >$ for the third equality, and using the definition (133) of the Fourier transform, we have, for all $t \in \mathbb{R}^{d_C-1}$,

$$\begin{aligned}
\mathcal{F}\left[\int_{g \in \mathcal{I}(x,(\cdot,1-\cdot^\top \mathbf{1}))} g d\mathbb{P}_{\underline{\Gamma}}(g)\right](t) &= \int_{\mathbb{R}^{d_C-1}} e^{it^\top \underline{y}} \int_{\mathcal{M}_{d_R \times (d_C-1)}([0,1])} \mathbb{1}\{g \in \mathcal{I}(x,y)\} g d\mathbb{P}_{\underline{\Gamma}}(g) d\underline{y} \\
&= \int_{\mathcal{M}_{d_R \times (d_C-1)}([0,1])} e^{it^\top (x^\top g)^\top} g d\mathbb{P}_{\underline{\Gamma}}(g) \\
&= \int_{\mathcal{M}_{d_R \times (d_C-1)}([0,1])} e^{i<tx^\top, g>} g d\mathbb{P}_{\underline{\Gamma}}(g) \\
&= \mathcal{F}\left[\star \mathbb{P}_{\underline{\Gamma}}(\star)\right](tx^\top).
\end{aligned}$$

Then, we conclude using Theorem 9.13 d) in Rudin (1973) and taking the Fourier inverse. □

Proof of Proposition 15 and Theorem 13. Let me start with the proof of Proposition 15, then particularize the result to prove Theorem 13. Consider $\mathbb{P}_{\Gamma,x,y}$ satisfying (8) and assumptions 17 and 18. (134) and Lemma 6 brings the identification to recovering, for $r = 1,\dots,d_R$ and $c = 1,\dots,d_C-1$, the function $t \in \mathbb{R}^{d_C-1} \mapsto \mathcal{F}\left[\star_{r,c} \mathbb{P}_{\underline{\Gamma}}(\star)\right]\left(tx^\top\right)$, for all $x \in \text{Supp}(X)$. For all $x \in \text{Supp}(\underline{X})$, I use the notation $\dot{x} := \left(x^\top, 1 - x^\top \mathbf{1}\right)^\top \in \text{Supp}(X)$.

Using Assumption 17, we have, for all $x \in \text{Supp}(\underline{X})$ and $t \in \mathbb{R}^{d_C-1}$,

$$\varphi(x,t) := \mathbb{E}\left[e^{it^\top \underline{Y}} | \underline{X} = x\right] = \mathcal{F}\left[\mathbb{P}_{\underline{\Gamma}}\right](t\dot{x}^\top). \tag{136}$$

Using the dominated convergence theorem, for all $c = 1, \dots, d_C - 1$, $r = 1, \dots, d_R - 1$, the function φ admits partial derivatives with respect to t_c and x_r. Moreover, using that $\text{Supp}(\underline{X})$ has a nonempty interior, the latter derivatives are identified on $\mathbb{S}_{\underline{X}}$, and we have, for all $t \in \mathbb{R}^{d_C-1}$ and $x \in \text{Supp}(\underline{X})$,

$$\partial_{t_c}\varphi(x,t) = i\dot{x}^\top \mathcal{F}\left[\star_{1:d_R,c}\mathbb{P}_{\underline{\Gamma}}(\star)\right]\left(t\dot{x}^\top\right), \tag{137}$$

$$\partial_{x_r}\varphi(x,t) = it^\top \mathcal{F}\left[\star_{r,1:d_C-1}\mathbb{P}_{\underline{\Gamma}}(\star)\right]\left(t\dot{x}^\top\right) - it^\top \mathcal{F}\left[\star_{d_R,1:d_C-1}\mathbb{P}_{\underline{\Gamma}}(\star)\right]\left(t\dot{x}^\top\right). \tag{138}$$

This brings back identification to solving, for all $t \in \mathbb{R}^{d_C-1}$, a system of $d_R \times (d_C - 1)$ unknowns $\mathcal{F}\left[\star_{r,c}\mathbb{P}_{\underline{\Gamma}}(\star)\right]\left(t\dot{x}^\top\right)$, $r = 1 \dots, d_R$, $c = 1, \dots, d_C - 1$, and $d_R + d_C - 2$ equations. Hence, $\mathbb{E}\left[\Gamma | X = x, Y = y\right]$ is identified under Assumption 18 when $d_C = 2$. Using Assumption 18 and the dominated convergence theorem, for all $(t,x) \in \mathbb{R}^{d_C-1} \times \text{Supp}(\underline{X})$, we have

$$\partial_{t_c}\varphi(x,t) = \int_{\text{Supp}(Y_{-C})} iy_c e^{it^\top y} f_{\underline{Y}|\underline{X}}(y|x)dy = i\mathcal{F}\left[\cdot f_{\underline{Y}|\underline{X}}(\cdot|x)\right](t).$$

Thus, we obtain, for all $y \in \text{Supp}(\underline{Y})$,

$$\mathcal{F}^{-1}\left[\partial_{t_c}\varphi(x,\cdot)\right](y) = iy_c f_{\underline{Y}|\underline{X}}(y|x) = i\rho_c(x,y). \tag{139}$$

Using Assumption 3, which yields that $\partial_{x_r}\varphi(x,\cdot) \in L^2(\mathbb{R})$ and (137)-(138), we obtain

$$\begin{aligned} \rho_c(x,y) &= \dot{x}^\top \mathcal{F}^{-1}\left[\mathcal{F}\left[\star_{1:d_R,c}\mathbb{P}_{\underline{\Gamma}}(\star)\right]\left(\cdot\dot{x}^\top\right)\right](y) \\ \mathcal{F}^{-1}\left[\partial_{x_r}\varphi(x,\cdot)\right](y) &= i\mathcal{F}^{-1}\left[\cdot^\top \mathcal{F}\left[(\star_{r,1:d_C-1} - \star_{d_R,1:d_C-1})\mathbb{P}_{\underline{\Gamma}}(\star)\right]\left(\cdot\dot{x}^\top\right)\right](y). \end{aligned} \tag{140}$$

Then, using that

$$\sum_{c=1}^{d_C-1} \partial_{y_c}\mathcal{F}^{-1}\left[\mathcal{F}\left[\star_{r,c}\mathbb{P}_{\underline{\Gamma}}(\star)\right]\left(\cdot\dot{x}^\top\right)\right](y) = -i\mathcal{F}^{-1}\left[\cdot^\top \mathcal{F}\left[\star_{r,1:d_C-1}\mathbb{P}_{\underline{\Gamma}}(\star)\right]\left(\cdot\dot{x}^\top\right)\right](y) \tag{141}$$

we obtain, for all $c = 1, \dots, d_C - 1$, $r = 1, \dots, d_R - 1$,

$$\begin{aligned} -\mathcal{F}^{-1}\left[\partial_{x_r}\varphi(x,\cdot)\right](y) = &\sum_{c=1}^{d_C-1} \partial_{y_c}\mathcal{F}^{-1}\left[\mathcal{F}\left[\star_{r,c}\mathbb{P}_{\underline{\Gamma}}(\star)\right]\left(\cdot\dot{x}^\top\right)\right](y) \\ &- \sum_{c=1}^{d_C-1} \partial_{y_c}\mathcal{F}^{-1}\left[\mathcal{F}\left[\star_{d_R,c}\mathbb{P}_{\underline{\Gamma}}(\star)\right]\left(\cdot\dot{x}^\top\right)\right](y). \end{aligned} \tag{142}$$

Denote by $M_{r,c} : (x,y) \in \text{Supp}(\underline{X},\underline{Y}) \mapsto \mathcal{F}^{-1}\left[\mathcal{F}\left[\star_{r,c}\mathbb{P}_{\underline{\Gamma}}(\star)\right]\left(t\dot{x}^\top\right)\right](y)$, for $r = 1,\dots,d_R$ and $c = 1,\dots,d_C-1$, which are continuous functions which admit a continuous derivative with respect to y_c. Moreover, from (138), we have $\text{PE} = M/f_{\underline{Y}|\underline{X}}$ and the constraint, for all $(x,y) \in \text{Supp}(\underline{X},\underline{Y})$, $M_{r,c}(x,y_1,\dots,y_c=0,y_{d_C-1}) = 0$ holds. Then, using (140), we obtain, for all $(x,y) \in \text{Supp}(\underline{X},\underline{Y})$,

$$\partial_{y_c}\rho_c(x,y) = \sum_{r=1}^{d_R-1} x_r \partial_{y_c} M_{r,c}(x,y) + \partial_{y_c} M_{d_R,c}(x,y) - \sum_{r=1}^{d_R-1} x_r \partial_{y_c} M_{d_R,c}(x,y) \tag{143}$$

and summing (142) over $r = 1,\dots,d_R-1$,

$$\begin{aligned} -\sum_{r=1}^{d_R-1} x_r \mathcal{F}^{-1}\left[\partial_{x_r}\varphi(x,\cdot)\right](y) &= \sum_{c=1}^{d_C-1}\left(\sum_{r=1}^{d_R-1} x_r \partial_{y_c} M_{r,c}(x,y) - \sum_{r=1}^{d_R-1} x_r \partial_{y_c} M_{d_R,c}(x,y)\right) \\ &= \sum_{c=1}^{d_C-1}\left(\partial_{y_c}\rho_c(x,y) - \partial_{y_c} M_{d_R,c}(x,y)\right). \end{aligned}$$

This yields

$$\sum_{c=1}^{d_C-1} \partial_{y_c} M_{d_R,c}(x,y) = \sum_{c=1}^{d_C-1} \partial_{y_c}\rho_c(x,y) + \sum_{r=1}^{d_R-1} x_r \mathcal{F}^{-1}\left[\partial_{x_r}\varphi(x,\cdot)\right](y).$$

Then, using Assumption 3 and the dominated convergence theorem for the first equality, then Theorem 9.13 d) in Rudin (1973) for the second, we have

$$\begin{aligned} \mathcal{F}^{-1}\left[\partial_{x_r}\varphi(x,\cdot)\right](y) &= \mathcal{F}^{-1}\left[\partial_{x_r}\mathcal{F}\left[f_{\underline{Y}|\underline{X}}(\cdot|x)\right]\right](y) \\ &= \partial_{x_r} f_{\underline{Y}|\underline{X}}(y|x). \end{aligned} \tag{144}$$

Using (142), we obtain (114). This yields that m takes the form described in the statement of Proposition 15.

When $d_C = 2$, integrating (114), using $M_{r,c}(x,0) = 0$, and $\rho_1(x,0) = 0$ for the first equality, and Assumption 3 and the dominated convergence theorem for the second one, we obtain, for all $r = 1,\dots,d_R$ and $(x,y) \in \text{Supp}(\underline{X},\underline{Y})$,

$$\begin{aligned} M_{r,1}(x,y) &= \rho_1(x,y) + \sum_{l=1}^{d_R-1}(x_l - \mathbb{1}\{l=r\})\int_0^y \partial_{x_l} f_{\underline{Y}|\underline{X}}(v|x)dv \\ &= \rho_1(x,y) + \sum_{l=1}^{d_R-1}(x_l - \mathbb{1}\{l=r\})\partial_{x_l} F_{\underline{Y}|\underline{X}}(y|x). \end{aligned}$$

Using $\rho_1(x,y) = y f_{\underline{Y}|\underline{X}}(y|x)$ yields the result of Theorem 13. □

Proof of Proposition 16. Denote the right hand side of (114) by, for $(x,y) \in \text{Supp}(X,Y)$,

$$\Theta_r(x,y) := \sum_{c=1}^{2} \partial_{y_c} \rho_c(x,y) + \sum_{l=1}^{d_R-1} (x_l - \mathbb{1}\{l=r\}) \partial_{\underline{x}_l} f_{\underline{Y}|\underline{X}}(\underline{y}|\underline{x}). \tag{145}$$

Then, (114) can we rewritten as, for $r = 1, \ldots, d_R - 1$,

$$\partial_{y_1} V_r(x,y) + \sum_{c=2}^{d_C-1} \partial_{y_c} M_{r,c}(x,y) = \Theta_r(x,y). \tag{146}$$

Using (119), we have, for $r = 1, \ldots, d_R - 1$

$$M_{r,2}(x,y) = \sum_{k=1}^{d_R-1} \widetilde{Q}_{r,k}(x) V_k(x,y) + \sum_{k=d_R}^{d_R+1} a_{r,k} \sigma_{k-d_R+1}(x,y) \tag{147}$$

which yields the system of coupled partial differential equations, for $r = 1, \ldots, d_R - 1$:

$$\partial_{y_1} V_r(x,y) + \sum_{k=1}^{d_R-1} \widetilde{Q}_{r,k}(x) \partial_{y_2} V_k(x,y) = \Theta_r(x,y), \tag{148}$$

with boundary constraints given by $V_r(0,y_2,x) = 0$ for $r = 1, \ldots, d_R - 1$. (148) is a system of coupled $(d_R-1) \times (d_R-1)$ transport partial differential equations that can be put into matrix form

$$\partial_{y_1} V(x,y) + \widetilde{Q}(x) \partial_{y_2} V(x,y) = \Theta(x,y). \tag{149}$$

When $d_C = 3$, using assumption (20.2) yields in (149),

$$\partial_{y_1} \widetilde{V}(x,y) + \text{diag}\left(\Lambda(x)\right) \partial_{y_2} \widetilde{V}(x,y) = P\Theta(x,y),$$

where $\widetilde{V} := PV$. Hence we can solve separately these $d_R - 1$ transport differential equations, for $r = 1, \ldots, d_R - 1$,

$$\widetilde{V}_r(x,y) = \sum_{k=1}^{d_R-1} P_{r,k}(x) \int_0^{y_1} \Theta_k(x, v, y_2 - \Lambda_r(x)(y_1 - v)) dv, \tag{150}$$

using that $\underline{V}_r(0, y_2, x) = 0$ for $r = 1, \ldots, d_R - 1$. Thus, using (145), we obtain

$$
\begin{aligned}
&\underline{V}_r(x, y) \\
&= \sum_{k=1}^{d_R-1} P_{r,k}(x) \int_0^{y_1} \partial_{y_1} \rho_1(x, v, y_2 - \Lambda_r(x)(y_1 - v)) dv \\
&+ \sum_{k=1}^{d_R-1} P_{r,k}(x) \int_0^{y_1} \partial_{y_2} \rho_2(x, v, y_2 - \Lambda_r(x)(y_1 - v)) dv \\
&+ \sum_{k=1}^{d_R-1} P_{r,k}(x) \sum_{l=1}^{d_R-1} (x_l - \mathbb{1}\{l = k\}) \int_0^{y_1} \mathcal{F}^{-1} \left[\partial_{x_l} \varphi(\underline{x}, \cdot)\right](v, y_2 - \Lambda_r(x)(y_1 - v)) dv \\
&= \mathrm{Diag}(PK\zeta)_r(x, y) \quad \text{(using (144))}.
\end{aligned}
$$

This yields the result using (122). □

G.7 Comparison with ground truth in an election dataset: turnout by race

I provide several validations of my methods in finite samples. Monte-Carlo simulations with the baseline independence assumption (Section 4.3.1) and in the panel data model (Section F.2) are given in appendix. Here, I consider an application to ecological inference where the true value of the parameters is known using specific register data.

This empirical illustration thus concerns the estimation of turnout by race and electoral precinct. This has important political implications, since racially homogeneous voting patterns are precluded by law. This falls into the context of ecological inference described in section G.1. More specifically, I focus on the case of studying the binary decision to vote according to $d_R = 3$ racial categories: White, Black, and Other, in the 2008 United States presidential election. I perform the analysis at the precinct level (8,843 observations), where the turnout by race is observed, allowing us to assess the performance of my method.

Several estimators emerge in this context, in particular that of Rosen et al. (2001), which relies on a Bayesian framework to make predictions. The independence assumption (2) has been the focus of some literature (see, *e.g.*, Tam Cho, 1998). Note, however, that since I am considering a national election, aggregation bias due to local

stakes is less likely.[16] Nevertheless, I consider three types of assumptions:

1. Assumption 16-1 (assuming NCE);
2. Assumption 6 conditioning the share of individuals registered as Democrat at the district level (Z_1);
3. Assumption 6 conditioning on the share of individuals registered as Other (Z_2).

In the case of 2, I thus control for local activism that could create aggregation effects at the precinct level. Despite the fact that the regressors here can be considered continuous as they represent the minority shares of the precinct population, I want to use my GWB estimator as a benchmark. To do so, I consider my discretization rule for each covariate (see section 4.2.4) before computing the estimator.

Table 9 shows the results, and Figure 10 shows contour plots, as well as a sample of the predictions and the actual realizations. First, note that in this example the computational time is actually reduced by using my nonparametric GT and GWB estimators rather than the Bayesian method based on simulations. Second, there does not seem to be much difference in this case between estimators using the additional variable Z or not, so there is some robustness of the independence assumption here 16-1. Third, my nonparametric GT estimator seems to have the best predictive performance, especially when using Z_2. It is closely followed by my GWB, which also performs well without controls, although it is not perfectly adapted to this context since the regressors are continuous. The GWB estimator, however, suffers an important loss when using the controls Z, mainly due to the fact that the sample is split according to the values of Z to be discretized (here in 3).

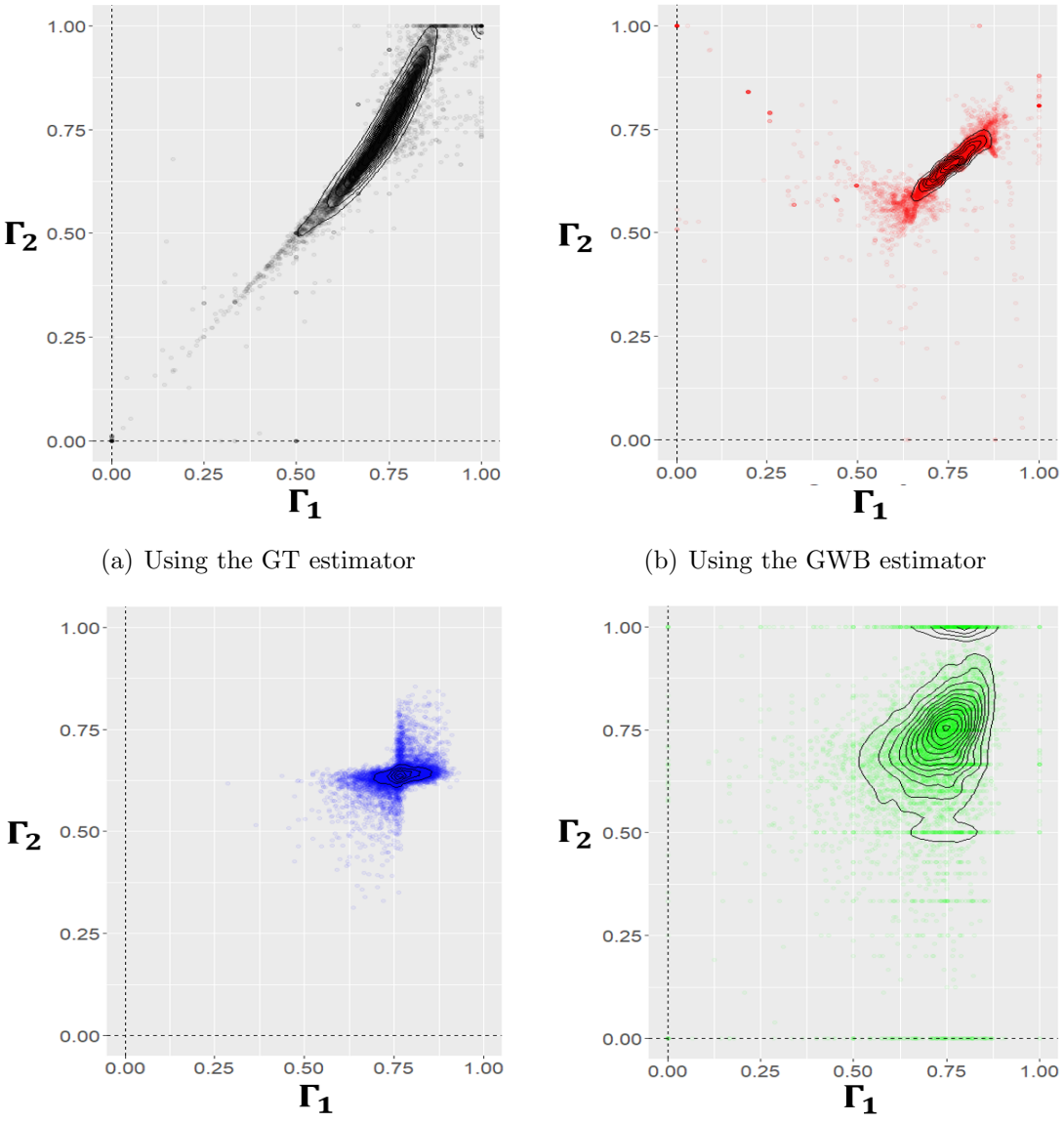
Finally and more importantly, the Bayesian estimator seems to miss the positive correlation that is observed in the true data, simply meaning that in some precinct people vote more independently of race. This important feature of the problem is well captured by my two estimators.

[16] However, due to the joint vote in the House and for the presidential election, individuals could decide to participate based on unobservable district-level stakes for the House election.

Table 9: In-sample errors in turnout by race in Florida

	MAE			RMSE			Time (s.)
	$\Gamma_{1,1}$	$\Gamma_{2,1}$	$\Gamma_{3,1}$	$\Gamma_{1,1}$	$\Gamma_{2,1}$	$\Gamma_{3,1}$	
Rosen et al. (2001), without controls	0.048	0.132	0.097	0.099	0.174	0.133	>3600
Rosen et al. (2001), with Z_1	0.044	0.196	0.088	0.102	0.230	0.130	>3600
Rosen et al. (2001), with Z_2	0.057	0.193	0.090	0.122	0.231	0.136	>3600
GT, without controls	0.024	**0.100**	0.085	0.152	**0.156**	0.206	12.0
GT, with Z_1	0.026	0.104	0.076	0.070	0.159	0.119	13.3
GT, with Z_2	**0.022**	0.102	**0.072**	**0.056**	0.162	**0.114**	41.6
GWB, without controls	0.029	0.117	**0.072**	0.069	0.163	**0.114**	150
GWB, with Z_1	0.054	0.246	0.101	0.092	0.328	0.139	155
GWB, with Z_2	0.043	0.166	0.087	0.082	0.220	0.132	161

Notes: in this 3×2 case, the in-sampled MAE is computed as $\sum_{i=1}^{n} |\widehat{\text{PE}}_{r,1}(X_i, Y_i, Z_i) - \Gamma_{r,1,i}|/n$ and the $RMSE$ as $(\sum_{i=1}^{n}(\widehat{\text{PE}}_{r,1}(X_i, Y_i, Z_i) - \Gamma_{r,1,i})^2/n)^{1/2}$, where $\widehat{\text{PE}}_{r,1}(X_i, Y_i, Z_i)$ are the different estimators. $\boldsymbol{B}_{1,1}$ is probability to vote conditional on being White, $\Gamma_{2,1}$ is probability to vote conditional on being Black, and $\Gamma_{3,1}$ is probability to vote conditional on being neither White nor Black. I use as Z_1 the share of individuals registered as democrats in the precinct and Z_2 the share of individuals whose party is neither Democrat nor Republican. "Comp. time" refers to computational time for estimation for one simulation. I use the implementation of Rosen et al. (2001) provided in the R package eiPack.



(a) Using the GT estimator

(b) Using the GWB estimator

(c) Rosen et al. (2001) param. Bayesian

(d) True values

Notes: These results represent the joint predictions for all Florida's 2008 electoral precincts of the probabilities to vote conditionally on being White (Γ_1) or Black (Γ_2), conditional on the observed values of the aggregate turnout rates and the racial composition of each precinct. The dots represent the individual predictions $\widehat{\text{PE}}(X_j, Y_j)$ and the contour lines the levels of the associated fitted density.

Figure 10: Joint distributions of the probabilities to vote conditionally on being White (Γ_1) or Black (Γ_2) for all Florida's 2008 electoral precincts