Electronic Journal of Statistics
Vol. 17 (2023) 2289–2316
ISSN: 1935-7524
https://doi.org/10.1214/23-EJS2144

Asymptotic normality of a change plane estimator in fixed dimension with near-optimal rate

Debarghya Mukherjee

Department of Statistics, University of Michigan
e-mail: mdeb@umich.edu

Moulinath Banerjee

Department of Statistics, University of Michigan
e-mail: mdeb@umich.edu

Debasri Mukherjee

Department of Economics, Western Michigan University
e-mail: debasri.mukherjee@wmich.edu

Ya'acov Ritov

Department of Statistics, University of Michigan
e-mail: mdeb@umich.edu

Abstract: Linear thresholding models postulate that the conditional distribution of a response variable in terms of covariates differs on the two sides of a (typically unknown) hyperplane in the covariate space. A key goal in such models is to learn about this separating hyperplane. Exact likelihood or least squares methods to estimate the thresholding parameter involve an indicator function which make them difficult to optimize and are, therefore, often tackled by using a surrogate loss that uses a smooth approximation to the indicator. In this paper, we demonstrate that the resulting estimator is asymptotically normal with a near optimal rate of convergence: n^{-1} up to a log factor, in both classification and regression thresholding models. This is substantially faster than the currently established convergence rates of smoothed estimators for similar models in the statistics and econometrics literatures. We also present a real-data application of our approach to an environmental data set where CO_2 emission is explained in terms of a separating hyperplane defined through per-capita GDP and urban agglomeration.

Keywords and phrases: Change plane estimator, CO_2 emission data, near-optimal rate.

Received March 2022.

Contents

1. Introduction

The simple linear regression model assumes a uniform linear relationship between the covariate and the response, in the sense that the regression parameter β is the same over the entire covariate domain. In practice, the situation can be more complicated: for instance, the regression parameter may differ from sub-population to sub-population within a large (super-) population. Some common techniques to account for such heterogeneity include mixed linear models, introducing an interaction effect, or fitting different models among each sub-population which corresponds to a supervised classification setting where the true groups (sub-populations) are *a priori known*.

A more difficult scenario arises when the sub-populations are unknown, in which case regression and classification must happen simultaneously. Consider the scenario where the conditional mean of Y_i given X_i is different for different unknown sub-groups. A well-studied treatment of this problem – the so-called change point problem – considers a simple thresholding model where membership in a sub-group is determined by whether a real-valued observable X falls to the left or right of an unknown parameter γ. More recently, there has been work for multi-dimensional covariates, namely when the membership is determined by which side a random vector X falls with respect to a hyperplane with unknown normal vector θ_0. A concrete example appears in [30] who extend the linear thresholding model due to [14] to general dimensions:

$$Y_t = \mu_1 + \mu_2 \cdot 1_{Q_t^\top \psi_0 > 0} + \varepsilon_t, \tag{1.1}$$

and studied computational algorithms and consistency of the same. Here Q_t can be thought as a subset of X_t (or may include some other covariates other than those in X_t) that determines the change hyperplane, ψ_0 is the (fixed) change-plane parameter, and t can be viewed as a time index. A natural extension of this change plane model incorporates the information from background covariates by replacing μ_1, μ_2 with non-trivial (typically linear) functions of the covariates. An example is the following threshold regression model:

$$Y_t = \beta^\top X_t + \delta^\top \tilde{X}_t \mathbb{1}_{Q_t^\top \psi_0 > 0} + \varepsilon_t. \tag{1.2}$$

For more details, see [10, 11] and the references therein. This model and others with similar structure, called *change plane models*, are useful in various fields of

research, e.g. modeling treatment effect heterogeneity in drug treatment ([13]), modeling sociological data on voting and employment ([13]), or cross country growth regressions in econometrics ([25]).

Other aspects of this model have also been investigated. [8] examined the change plane model from the statistical testing point of view, with the null hypothesis being the absence of a separating hyperplane. They proposed a test statistic, studied its asymptotic distribution and provided sample size recommendations for achieving target values of power. [18] extended the threshold regression model of (1.2) to incorporate multi-change plane indices, depending on the value of $Q_t^\top \psi_0$.

The key difficultly with change plane type models is the inherent discontinuity in the optimization criteria involved where the parameter of interest appears as an argument to some indicator function, rendering the optimization extremely hard. To alleviate this, one option is to kernel smooth the indicator function, an approach that was adopted by Seo and Linton [25] to analyze the model presented in (1.2), motivated by earlier results of Horowitz [12] that dealt with a smoothed version of the maximum score estimator. More precisely, one of the estimators analyzed in [25] is as follows:

$$(\hat{\beta}^S, \hat{\delta}^S, \hat{\psi}^S) = \arg\min_{\beta,\delta,\psi} \left\{ \frac{1}{n} \sum_{i=1}^{n} \left[\left(Y_i - X_i^\top \beta\right)^2 + \left\{ \left(Y_i - X_i^\top \beta - \tilde{X}_i^\top \delta\right)^2 - \left(Y_i - X_i^\top \beta\right)^2 \right\} K\left(\frac{Q_i^\top \psi}{\sigma_n} \right) \right] \right\} \quad (1.3)$$

for some smooth kernel K and some bandwidth parameter σ_n. The value of σ_n determines the approximation of the indicator by the kernel, the smaller is σ_n, the better is the approximation. Under a set of assumptions on the model (Assumptions 1 and 2 of their paper), the authors of [25] established asymptotic normality of $\hat{\psi}$ obtained in (1.3) with the rate of convergence being $\sqrt{n/\sigma_n}$. As noted in Remark 3 of [25], under the special case of i.i.d. observations, their requirement that $\log n/(n\sigma_n^2) \to 0$ translates to a maximal convergence rate of $n^{3/4}$ up to a logarithmic factor. The work of [18] who considered multiple parallel change planes (determined by a fixed dimensional normal vector) and high dimensional linear models in the regions between consecutive hyperplanes also builds partly upon the methods of [25] and obtains the same (almost) $n^{3/4}$ rate for the normal vector (as can be seen by putting Condition 6 in their paper in conjunction with the conclusion of Theorem 3).

We note that kernel smoothing is no panacea: as the resulting loss function is non-convex, it is possible to hit a local minima using first order methods like gradient descent. To escape from local minima, we may start gradient descent from a consistent estimator, but for that we need a computable consistent estimator to begin with. This being said, the kernel smoothed estimator is interesting in its own right, as it provides an *easier* way to obtain an estimate with tractable asymptotic distribution, while the distribution of the non-smoothed least squares estimator is extremely complicated. Recently [16]

developed a method based on mixed integer programing to calculate the least squares estimator. It would be interesting to compare these two procedures but that is beyond the scope of this manuscript.

While it is established that the condition $n\sigma_n^2 \to \infty$ is sufficient (upto a log factor) for achieving asymptotic normality of the smoothed estimator, there is no result in the existing literature to ascertain whether its necessity. Intuitively speaking, the necessary condition for asymptotic normality ought to be $n\sigma_n \to \infty$, as this will ensure a growing number of observations in a σ_n neighborhood around the true hyperplane, allowing the central limit theorem to kick in. In this paper we *bridge this gap* by proving that asymptotic normality of the smoothed change point estimator is, in fact, achievable with $n\sigma_n \to \infty$. This implies that the best possible rate of convergence of the smoothed estimator can be arbitrarily close to n^{-1}, the minimax optimal rate of estimation for this problem. In this paper, we prove this intuitive claim for a (slightly) modified version of (1.2):

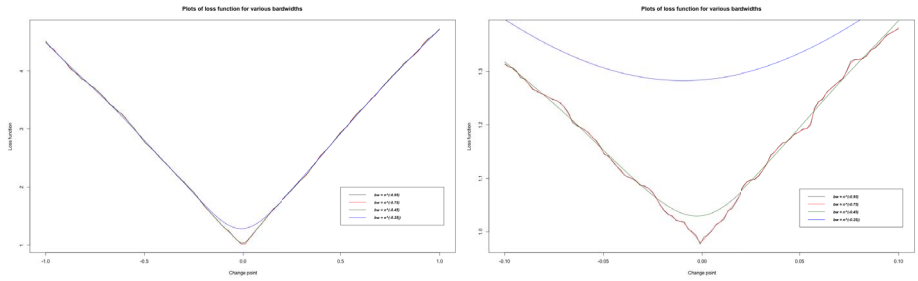
$$Y_i = \beta_0^\top X_i + \delta_0^\top X_i \mathbb{1}_{Q_i^\top \psi_0 > 0} + \epsilon_i, \tag{1.4}$$

for i.i.d. observations $\{(X_i, Y_i, Q_i\}_{i=1}^n$, where the zero-mean transitory shocks $\epsilon_i \perp\!\!\!\perp (X_i, Q_i)$. Our calculation can be easily extended to the case when the covariates on the either side of the change hyperplane are different (i.e. $X \neq \tilde{X}$ in (1.2)) and $\mathbb{E}[\epsilon \mid X, Q] = 0$ with more tedious bookkeeping. As this generalization adds little of interest, conceptually, to our proof, we posit the simpler model for ease of understanding. As the parameter ψ_0 is only identifiable upto its norm, we assume that the first co-ordinate is 1 (along the lines of [25]) which removes one degree of freedom and makes the parameter identifiable. We show that $\sqrt{n/\sigma_n}(\hat{\psi}^S - \psi_0)$ converges to zero-mean normal distribution as long as $n\sigma_n \to \infty$, where $\hat{\psi}^S$ is the smoothed estimator, obtained by solving a version of (1.3), modified appropriately for (1.4). Therefore, the rate of convergence of the smoothed estimator $\hat{\psi}^S$ can be arbitrarily close to the minimax optimal rate n^{-1}.

Before going into further details, let us present a brief visual demonstration of the effect of the bandwidth on the smoothness and curvature of the loss function. The non-smoothed squared error loss function – see equation (2.1) – has a sharp (linear) curvature near the true parameter (like the curvature of $|x|$ at the origin) which yields the faster n^{-1} rate. However, as soon as we replace the indicator in the squared error loss by a smooth kernel, the curvature becomes quadratic (like x^2 near origin) and the rate of convergence slows down. This trade-off is governed by the choice of the bandwidth parameter σ_n: when σ_n is small, the effect of the kernel is diminished and the curvature is close to linear, whereas increasing the value of $sigma_n$ smoothens out the loss function, making the curvature close to quadratic. To illustrate this effect, we simulate from the following setup:

$$Y_i = X_i^\top \beta_0 + X_i^\top \delta_0 \mathbb{1}_{X_{i,1} > 0} + \epsilon_i$$

where $X_i \sim \mathcal{N}(0, I_2)$, $\epsilon_i \sim \mathcal{N}(0, 1)$ and the true parameter $\psi_0 = (1, 0)$. In Fig. 1a, we plot the smoothed loss function (equation (1.3)) with the stan-



(a) Entire region (b) Zoomed in near origin

FIG 1. *Effect of bandwidth on smoothing.*

dard Gaussian kernel on the region $[-1, 1]$ for four different bandwidths: $\sigma_n \in \{n^{-0.95}, n^{-0.75}, n^{-0.45}, n^{-0.25}\}$. It is evident from the picture all the loss functions are minimized close to 0, but the change of smoothness with respect to the bandwidth is not immediately clear. Therefore, we provide another figure (Fig. 1b), which magnifies Fig. 1a in the vicinity of 0 (on $[-0.1, 0.1]$), which clarifies the effect of the bandwidth: as σ_n becomes larger, the loss function becomes smoother and the curvature approaches quadratic.

The key difference between our approach and that of [25] is Lemma 2.6, which posits that the curvature of the loss function basically interpolates between a linear and a quadratic curvature in terms of the bandwidth parameter. As we have elaborated in the previous paragraph, linear curvature, roughly speaking corresponds to the n^{-1} rate, whereas quadratic curvature corresponds to the $n^{-1/2}$ rate. The precise interpolation between these regimes in terms of the bandwidth σ_n yields a rate with optimal dependence on σ_n.

We further extend our analysis for the binary response model, which can be briefly described as follows: As before, the covariate $Q \sim P$ where P is the distribution on $\mathbb{R}^d$ and the conditional distribution of Y given Q is modeled as follows:

$$P(Y = 1|Q) = \alpha_0 \mathbb{1}(Q^\top \psi_0 \le 0) + \beta_0 \mathbb{1}(Q^\top \psi_0 > 0) \tag{1.5}$$

for some parameters $\alpha_0, \beta_0 \in (0, 1)$ and $\psi_0 \in \mathbb{R}^d$ (with first co-ordinate being one for identifiability issue as for the continuous response model), the latter being of primary interest for estimation. This model is identifiable up to a permutation of (α_0, β_0), so we further assume $\alpha_0 < \beta_0$. As in the case of threshold regression model, here we also show that $\sqrt{n/\sigma_n}(\hat{\psi} - \psi_0)$ converges to the zero-mean normal distribution as long as $n\sigma_n \to \infty$, but the calculations for the binary model are relegated to the supplement in the interests of a more focused narrative for the core part of the paper (see Section 2 and Section 3 of the supplementary document).

Organization of the paper The rest of the paper is organized as follows: In Section 2 we present the methodology, the statement of the asymptotic dis-

tributions and a sketch of the proof for the continuous response model (1.4). In Section 3 we present a brief description of the inferential methods based on the smoothed estimator. In Section 4 we present some simulation results, both for the binary and the continuous response models to study the effect of the bandwidth on the quality of the normal approximation in finite samples. In Section 5, we present a real data analysis where we analyze the effect of income and urbanization on the CO_2 emission in different countries.

Notations Before delving into the technical details, we first setup some notations here. We assume from now on, $X \in \mathbb{R}^p$ and $Q \in \mathbb{R}^d$. For any vector v we define by $\tilde{v}$ as the vector with all the co-ordinates expect the first one. We denote by K the kernel function used to smooth the indicator function. For any matrix A, we denote by $\|A\|_2$ (or $\|A\|_F$) as its Frobenius norm and $\|A\|_{op}$ as its operator norm. For any vector, $\|\cdot\|_2$ denotes its ℓ_2 norm.

2. Methodology and theory for continuous response model

In this section we present our analysis for the continuous response model. Without smoothing, the original estimating equation is:

$$f_{\beta,\delta,\psi}(Y,X,Q) = \left(Y - X^\top\beta - X^\top\delta \mathbb{1}_{Q^\top\psi>0}\right)^2$$

and we estimate the parameters as:

$$\begin{aligned}\left(\hat{\beta}^{LS},\hat{\delta}^{LS},\hat{\psi}^{LS}\right) &= \arg\min_{(\beta,\delta,\psi)\in\Theta}\mathbb{P}_n f_{\beta,\delta,\psi} \\ &:= \arg\min_{(\beta,\delta,\psi)\in\Theta}\mathbb{M}_n(\beta,\delta,\psi). \end{aligned} \tag{2.1}$$

where $\mathbb{P}_n$ is empirical measure based on i.i.d. observations $\{(X_i,Y_i,Q_i)\}_{i=1}^n$ and Θ is the parameter space. Henceforth, we assume Θ is a compact subset of dimension $\mathbb{R}^{2p+d}$. We also define $\theta = (\beta,\delta,\psi)$, i.e. all the parameters together as a vector and by θ_0 is used to denote the true parameter vector $(\beta_0,\delta_0,\psi_0)$. Some modification of equation (2.1) leads to the following:

$$\begin{aligned}(\hat{\beta}^{LS},\hat{\delta}^{LS},\hat{\psi}^{LS}) &= \arg\min_{\beta,\delta,\psi}\sum_{i=1}^n\left(Y_i - X_i^\top\beta - X_i^\top\delta\mathbb{1}_{Q_i^\top\psi>0}\right)^2 \\ &= \arg\min_{\beta,\delta,\psi}\sum_{i=1}^n\Big[\left(Y_i - X_i^\top\beta\right)^2\mathbb{1}_{Q_i^\top\psi_0\le 0} \\ &\qquad + \left(Y_i - X_i^\top\beta - X_i^\top\delta\right)^2\mathbb{1}_{Q_i^\top\psi>0}\Big] \\ &= \arg\min_{\beta,\delta,\psi}\sum_{i=1}^n\Big[\left(Y_i - X_i^\top\beta\right)^2 + \Big\{\left(Y_i - X_i^\top\beta - X_i^\top\delta\right)^2 \\ &\qquad - \left(Y_i - X_i^\top\beta\right)^2\Big\}\mathbb{1}_{Q_i^\top\psi>0}\Big]\end{aligned}$$

Typical empirical process calculations yield under mild conditions:

$$\|\hat{\beta}^{LS} - \beta_0\|^2 + \|\hat{\delta}^{LS} - \delta_0\|^2 + \|\hat{\psi}^{LS} - \psi_0\|_2 = O_p(n^{-1})$$

but inference is difficult as the limit distribution is unknown, and in any case, would be a highly non-standard distribution. Recall that even in the one dimensional change point model with fixed jump size, the least squares change point estimator converges at rate n to the truth with a non-standard limit distribution, namely a minimizer of a two-sided compound Poisson process (see [15] for more details). To obtain a computable estimator with tractable limiting distribution, we resort to a smooth approximation of the indicator function in (2.1) using a distribution kernel with suitable bandwidth, i.e we replace $\mathbb{1}_{Q_i^\top \psi > 0}$ by $K(Q_i^\top \psi/\sigma_n)$ for some appropriate distribution function K and bandwidth σ_n, i.e.

$$\begin{aligned}(\hat{\beta}^S, \hat{\delta}^S, \hat{\psi}^S) &= \arg\min_{\beta,\delta,\psi} \left\{ \frac{1}{n} \sum_{i=1}^{n} \left[\left(Y_i - X_i^\top \beta\right)^2 + \left\{ \left(Y_i - X_i^\top \beta - X_i^\top \delta\right)^2 \right.\right.\right. \\ &\qquad \left.\left.\left. - \left(Y_i - X_i^\top \beta\right)^2 \right\} K\left(\frac{Q_i^\top \psi}{\sigma_n}\right) \right] \right\} \\ &= \arg\min_{(\beta,\delta,\psi)\in\Theta} \mathbb{P}_n f^s_{(\beta,\delta,\psi)}(X, Y, Q) \\ &:= \arg\min_{\theta\in\Theta} \mathbb{M}^s_n(\theta).\end{aligned}$$

Define $\mathbb{M}$ (resp. $\mathbb{M}^s$) to be the population counterpart of $\mathbb{M}_n$ and $\mathbb{M}^s_n$ respectively which are defined as:

$$\begin{aligned}\mathbb{M}(\theta) &= \mathbb{E}\left(Y - X^\top \beta\right)^2 + \mathbb{E}\left(\left[-2\left(Y_i - X^\top \beta\right) X^\top \delta + (X^\top \delta)^2\right] \mathbb{1}_{Q^\top \psi > 0}\right), \\ \mathbb{M}^s(\theta) &= \mathbb{E}\left[(Y - X^\top \beta)^2 + \left\{-2(Y - X^\top \beta)(X^\top \delta) + (X^\top \delta)^2\right\} K\left(\frac{Q^\top \psi}{\sigma_n}\right)\right].\end{aligned}$$

As noted in the proof of Seo and Linton, the assumption $\log n/n\sigma_n^2 \to 0$ was only used to show:

$$\frac{\|\hat{\psi}^s - \psi_0\|}{\sigma_n} = o_p(1).$$

In this paper, we show that one can achieve the same conclusion as long as $n\sigma_n \to \infty$. The rest of the proof for the normality is similar to that of [25], we will present it briefly for the ease the readers. The proof is quite long and technical, therefore we break the proof into several lemmas. We, first, list our assumptions:

Assumption 2.1. *1. Define $f_\psi(\cdot \mid \tilde{Q})$ to be the conditional distribution of $Q^\top \psi$ given $\tilde{Q}$. (In particular we will denote by $f_0(\cdot \mid \tilde{q})$ to be conditional distribution of $Q^\top \psi_0$ given $\tilde{Q}$ and $f_s(\cdot \mid \tilde{q})$ to be the conditional distribution of $Q^\top \psi_0^s$ given $\tilde{Q}$. Assume that there exists F_+ such that $\sup_t f_0(t|\tilde{Q}) \le F_+$ almost surely on $\tilde{Q}$ and for all ψ in a neighborhood of ψ_0 (in particular for ψ_0^s). Further assume that f_ψ is differentiable and the derivative is bounded by F_+ for all ψ in a neighborhood of ψ_0 (again in particular for ψ_0^s).*

2. *Define* $g(Q) = \mathsf{var}(X \mid Q)$. *There exists* c_- *and* c_+ *such that* $c_- \leq \lambda_{\min}(g(Q)) \leq \lambda_{\max}(g(Q)) \leq c_+$ *almost surely. Also assume that* g *is a Lipschitz with constant* G_+ *with respect to* Q.
3. *There exists* $p_+ < \infty$ *and* $p_- > 0, r > 0$ *such that:*

$$p_- \|\psi - \psi_0\| \leq \mathbb{P}\left(\mathsf{sign}\left(Q^\top \psi\right) \neq \mathsf{sign}\left(Q^\top \psi_0\right)\right) \leq p_+ \|\psi - \psi_0\|,$$

for all ψ *such that* $\|\psi - \psi_0\| \leq r$.
4. *For all* ψ *in the parameter space* $0 < \mathbb{P}\left(Q^\top \psi > 0\right) < 1$.
5. *Define* $m_2(Q) = \mathbb{E}\left[\|X\|^2 \mid Q\right]$ *and* $m_4(Q) = \mathbb{E}\left[\|X\|^4 \mid Q\right]$. *Assume* m_2, m_4 *are bounded Lipschitz function of* Q.

2.1. Sufficient conditions for above assumptions

We now demonstrate some sufficient conditions for the above assumptions to hold. The first condition is essentially a condition on the conditional density of the first co-ordinate of Q given all other co-ordinates. If this conditional density is bounded and has bounded derivative, then first assumption is satisfied. This condition is satisfied in fair generality. The second assumption implies that the conditional distribution of X given Q has variance in all the direction over all Q. This is also very weak condition, as is satisfied for example if X and Q and independent (with X has non-degenerate covariance matrix) or (X, Q) are jointly normally distributed to name a few. This condition can further be weaken by assuming that the maximum and minimum eigenvalues of $\mathbb{E}[g(Q)]$ are bounded away from ∞ and 0 respectively but it requires more tedious book-keeping. The third assumption is satisfied as long as as $Q^\top \psi$ has non-zero density near origin, while the fourth assumption merely states that the support of Q is not confined to one side of the hyperplane for any hyperplane and a simple sufficient condition for this is Q has continuous density with non-zero value at the origin. The last assumption is analogous to the second assumption for the conditional fourth moment which is also satisfied in fair generality.

Remark 2.2. *In this paper, we assume* r *is fixed, however one can certainly let* r *to go to* 0 *as* $n \uparrow \infty$, *at the cost of a slower rate of convergence of the change plane estimator. The intuitive explanation is that,* r *quantifies the curvature of the risk around the truth. In the extreme case of* $r = 0$, *we don't have any curvature of the risk function and there will be identifiability issue as the excess risk* $\mathbb{P}(\mathsf{sign}(Q^\top \psi) \neq \mathsf{sign}(Q^\top \psi_0)) = 0$ *will no longer imply that* $\psi = \psi_0$. *Consequently, the risk consistency of our estimator won't imply the* ℓ_2 *consistency of the estimator itself. This is why it is crucial to assume* $r > 0$. *This intuition implies that if we let* r *to go to* 0 *with* n, *then we are losing the curvature and consequently we expect that the rate of the convergence of the estimator will be slower. To quantify the precise effect of* r *on the rate of convergence, we need to track it carefully throughout the proof, which requires more book-keeping. This* r *will start appearing whenever we lower bound* $\mathbb{M}(\theta) - \mathbb{M}(\theta_0)$ *or* $\mathbb{M}^s(\theta) - \mathbb{M}^s(\theta_0^s)$, *for example in the proof of Lemma 2.5.*

Kernel function and bandwidth We take $K(x) = \Phi(x)$ (distribution of standard normal random variable) for our analysis. For the bandwidth we assume $n\sigma_n^2 \to 0$ and $n\sigma_n \to \infty$ as the other case, (i.e. $n\sigma_n^2 \to \infty$) is already established in [25].

Based on Assumption 2.1 and our choice of kernel and bandwidth we establish the following theorem:

Theorem 2.3. *Under Assumption 2.1 and the above choice of kernel and bandwidth we have:*

$$\sqrt{n}\left(\begin{pmatrix}\hat{\beta}^s\\ \hat{\delta}^s\end{pmatrix} - \begin{pmatrix}\beta_0\\ \delta_0\end{pmatrix}\right) \overset{\mathscr{L}}{\Longrightarrow} \mathcal{N}(0, \Sigma_{\beta,\delta})$$

and

$$\sqrt{n/\sigma_n}\left(\hat{\psi}^s - \psi_0\right) \overset{\mathscr{L}}{\Longrightarrow} \mathcal{N}(0, \Sigma_\psi),$$

for matrices $\Sigma_{\beta,\delta}$ and Σ_ψ mentioned explicitly in the proof. Moreover they are asymptotically independent.

The proof of the theorem is relatively long, so we break it into several lemmas. We provide a roadmap of the proof in this section while the elaborate technical derivations of the supporting lemmas can be found in Appendix. Let $\nabla\mathbb{M}_n^s(\theta)$ and $\nabla^2\mathbb{M}_n^s(\theta)$ be the gradient and Hessian of $\mathbb{M}_n^s(\theta)$ with respect to θ. As $\hat{\theta}^s$ minimizes $\mathbb{M}_n^s(\theta)$, we have from the first order condition, $\nabla\mathbb{M}_n^s(\hat{\theta}^s) = 0$. Using one step Taylor expansion we have:

$$\nabla\mathbb{M}_n^s(\hat{\theta}^s) = \nabla\mathbb{M}_n^s(\theta_0) + \nabla^2\mathbb{M}_n^s(\theta^*)\left(\hat{\theta}^s - \theta_0\right) = 0$$

i.e.

$$\left(\hat{\theta}^s - \theta_0\right) = -\left(\nabla^2\mathbb{M}_n^s(\theta^*)\right)^{-1}\nabla\mathbb{M}_n^s(\theta_0) \tag{2.2}$$

for some intermediate point θ^* between $\hat{\theta}^s$ and θ_0. Following the notation of [25], define a diagonal matrix D_n of dimension $2p + d$ with first $2p$ elements being 1 and the last d elements being $\sqrt{\sigma_n}$. we can write:

$$\begin{aligned}
&\sqrt{n}D_n^{-1}(\hat{\theta}^s - \theta_0)\\
&= -\sqrt{n}D_n^{-1}\nabla^2\mathbb{M}_n^s(\theta^*)^{-1}\nabla\mathbb{M}_n^s(\theta_0)\\
&= \begin{pmatrix}\nabla^2\mathbb{M}_n^{s,\gamma}(\theta^*) & \sqrt{\sigma_n}\nabla^2\mathbb{M}_n^{s,\gamma\psi}(\theta^*)\\ \sqrt{\sigma_n}\nabla^2\mathbb{M}_n^{s,\gamma\psi}(\theta^*) & \sigma_n\nabla^2\mathbb{M}_n^{s,\psi}(\theta^*)\end{pmatrix}^{-1}\begin{pmatrix}\sqrt{n}\nabla\mathbb{M}_n^{s,\gamma}(\theta_0)\\ \sqrt{n\sigma_n}\nabla\mathbb{M}_n^{s,\psi}(\theta_0)\end{pmatrix}
\end{aligned} \tag{2.3}$$

where $\gamma = (\beta, \delta) \in \mathbb{R}^{2p}$. The following lemma establishes the asymptotic properties of $\nabla\mathbb{M}_n^s(\theta_0)$:

Lemma 2.4 (Asymptotic Normality of $\nabla\mathbb{M}_n^s(\theta_0)$)**.** *Under Assumption 2.1 we have:*

$$\begin{aligned}
\sqrt{n}\nabla\mathbb{M}_n^{s,\gamma}(\theta_0) &\Longrightarrow \mathcal{N}\left(0, 4V^\gamma\right),\\
\sqrt{n\sigma_n}\nabla\mathbb{M}_n^{s,\psi}(\theta_0) &\Longrightarrow \mathcal{N}\left(0, V^\psi\right).
\end{aligned}$$

for some n.n.d. matrices V^γ and V^ψ which is mentioned explicitly in the proof. Further more $\sqrt{n}\nabla\mathbb{M}_n^{s,\gamma}(\theta_0)$ and $\sqrt{n\sigma_n}\nabla\mathbb{M}_n^{s,\psi}(\theta_0)$ are asymptotically independent.

Next, we analyze the convergence of $\nabla^2\mathbb{M}_n^s(\theta^*)$ which is stated in the following lemma:

Lemma 2.5 (Convergence in Probability of $\nabla^s\mathbb{M}_n^s(\theta^*)$)**.** *Under Assumption (2.1), for any random sequence $\breve{\theta} = \left(\breve{\beta}, \breve{\delta}, \breve{\psi}\right)$ such that $\breve{\beta} \xrightarrow{p} \beta_0, \breve{\delta} \xrightarrow{p} \delta_0, \|\breve{\psi} - \psi_0\|/\sigma_n \xrightarrow{P} 0$, we have:*

$$\begin{aligned}\nabla^2_\gamma \mathbb{M}_n^s(\breve{\theta}) &\xrightarrow{p} 2Q^\gamma,\\ \sqrt{\sigma_n}\nabla^2_{\psi\gamma}\mathbb{M}_n^s(\breve{\theta}) &\xrightarrow{p} 0,\\ \sigma_n\nabla^2_{\psi}\mathbb{M}_n^s(\breve{\theta}) &\xrightarrow{p} Q^\psi.\end{aligned}$$

for some matrices Q^γ, Q^ψ mentioned explicitly in the proof. This, along with equation (2.3), establishes:

$$\begin{aligned}\sqrt{n}\left(\hat{\gamma}^s - \gamma_0\right) &\overset{\mathscr{L}}{\Longrightarrow} \mathcal{N}\left(0, {Q^\gamma}^{-1} V^\gamma {Q^\gamma}^{-1}\right),\\ \sqrt{n/\sigma_n}\left(\hat{\psi}^s - \psi_0\right) &\overset{\mathscr{L}}{\Longrightarrow} \mathcal{N}\left(0, {Q^\psi}^{-1} V^\psi {Q^\psi}^{-1}\right),\end{aligned}$$

where as before $\hat{\gamma}^s = (\hat{\beta}^s, \hat{\delta}^s)$.

It will be shown later that the condition $\|\breve{\psi}_n - \psi_0\|/\sigma_n \xrightarrow{P} 0$ needed in Lemma 2.5 holds for the (random) sequence ψ^*, the intermediate point in the Taylor expansion. Then, combining Lemma 2.4 and Lemma 2.5 we conclude the proof of Theorem 2.3. Observe that, to show $\|\psi^* - \psi_0\| = o_P(\sigma_n)$, it suffices to to prove that $\|\hat{\psi}^s - \psi_0\| = o_P(\sigma_n)$. Towards that direction, we have following lemma:

Lemma 2.6 (Rate of convergence)**.** *Under Assumption 2.1 and our choice of kernel and bandwidth,*

$$n^{2/3}\sigma_n^{-1/3} d_*^2\left(\hat{\theta}^s, \theta_0^s\right) = O_P(1),$$

where

$$\begin{aligned}d_*^2(\theta, \theta_0^s) = \|\beta - \beta_0^s\|^2 + \|\delta - \delta_0^s\|^2 \\ + \frac{\|\psi - \psi_0^s\|^2}{\sigma_n}\mathbb{1}_{\|\psi-\psi_0^s\|\le\mathcal{K}\sigma_n} + \|\psi - \psi_0^s\|\mathbb{1}_{\|\psi-\psi_0^s\|>\mathcal{K}\sigma_n},\end{aligned}$$

for some specific constant $\mathcal{K}$. (This constant will be mentioned precisely in the proof.) Hence as $n\sigma_n \to \infty$, we have $n^{2/3}\sigma_n^{-1/3} \gg \sigma_n^{-1}$ which implies $\|\hat{\psi}^s - \psi_0^s\|/\sigma_n \xrightarrow{P} 0$.

The above lemma establishes $\|\hat{\psi}^s - \psi_0^s\|/\sigma_n = o_p(1)$ but our goal is to show that $\|\hat{\psi}^s - \psi_0\|/\sigma_n = o_p(1)$. Therefore, we further need $\|\psi_0^s - \psi_0\|/\sigma_n \to 0$ which is demonstrated in the following lemma:

Lemma 2.7 (Convergence of population minimizer)**.** *Under Assumption 2.1 and our choice of kernel and bandwidth, we have:* $\|\psi_0^s - \psi_0\|/\sigma_n \to 0$.

Hence the final roadmap is the following: Using Lemma 2.7 and Lemma 2.6 we establish that $\|\hat{\psi}^s - \psi_0\|/\sigma_n = o_p(1)$ if $n\sigma_n \to 0$. This, in turn, enables us to prove Lemma 2.5, i.e. $\sigma_n \nabla^2 \mathbb{M}_n^s(\theta^*) \xrightarrow{P} Q$,which, along with Lemma 2.4, establishes the main theorem.

Remark 2.8. *As the entire proof is quite long and involved, we present the proof of the key lemma i.e. Lemma 2.6 in the Appendix of this main draft, whereas the proofs of the other Lemmas can be found in Section 1 of Supplementary document.*

3. Inferential methods

We draw inferences on $(\beta_0, \delta_0, \psi_0)$ by resorting to similar techniques as in [25]. For the continuous response model, we need consistent estimators of $V^\gamma, Q^\gamma, V^\psi$ and Q^ψ (see Lemma 2.5 for the definitions) for hypothesis testing. By virtue of the aforementioned Lemma, we can estimate Q^γ and Q^ψ as follows:

$$\hat{Q}^\gamma = \nabla^2_\gamma \mathbb{M}_n^s(\hat{\theta}),$$
$$\hat{Q}^\psi = \sigma_n \nabla^2_\psi \mathbb{M}_n^s(\hat{\theta}).$$

The consistency of the above estimators is established in the proof of Lemma 2.5. For the other two parameters V^γ, V^ψ we use the following estimators:

$$\hat{V}^\psi = \frac{1}{n\sigma_n^2} \sum_{i=1}^n \left(\left(Y_i - X_i^\top (\hat{\beta} + \hat{\delta}) \right)^2 - \left(Y_i - X_i^\top \hat{\beta} \right)^2 \right)^2 \tilde{Q}_i \tilde{Q}_i^\top \left(K' \left(\frac{Q_i^\top \hat{\psi}}{\sigma_n} \right) \right)^2$$

$$\hat{V}^\gamma = \hat{\sigma}_\epsilon^2 \begin{pmatrix} \frac{1}{n} X_i X_i^\top & \frac{1}{n} X_i X_i^\top \mathbb{1}_{Q_i^\top \hat{\psi} > 0} \\ \frac{1}{n} X_i X_i^\top \mathbb{1}_{Q_i^\top \hat{\psi} > 0} & \frac{1}{n} X_i X_i^\top \mathbb{1}_{Q_i^\top \hat{\psi} > 0} \end{pmatrix}$$

where $\hat{\sigma}_\epsilon^2$ can be obtained as $(1/n)(Y_i - X_i^\top \hat{\beta} - X_i^\top \hat{\delta} \mathbb{1}(Q_i^\top \hat{\psi} > 0))^2$, i.e. the residual sum of squares. The explicit value of V_γ (as derived in equation (1.23) in the proof Lemma 2.4) is:

$$V^\gamma = \sigma_\epsilon^2 \begin{pmatrix} \mathbb{E}\left[XX^\top \right] & \mathbb{E}\left[XX^\top \mathbb{1}_{Q^\top \psi_0 > 0} \right] \\ \mathbb{E}\left[XX^\top \mathbb{1}_{Q^\top \psi_0 > 0} \right] & \mathbb{E}\left[XX^\top \mathbb{1}_{Q^\top \psi_0 > 0} \right] \end{pmatrix}$$

Therefore, the consistency of $\hat{V}_\gamma$ is immediate from the law of large numbers. The consistency of $\hat{V}^\psi$ follows via arguments similar to those employed in proving Lemma 2.5 but under somewhat more stringent moment conditions: in particular, we need $\mathbb{E}[\|X\|^8] < \infty$ and $\mathbb{E}[(X^\top \delta_0)^k \mid Q]$ to be Lipschitz functions over Q for $1 \le k \le 8$. The inferential techniques for the classification model are similar and hence skipped, to avoid repetition.

4. Simulation studies

In this section, we present some simulation results to analyse the effect of the choice of σ_n on the finite sample approximation of asymptotic normality, i.e. Berry-Essen type bounds. If we choose a smaller sigma, the rate of convergence is accelerated but the normal approximation error at smaller sample sizes will be higher, as we don't have enough observations in the vicinity of the change hyperplane for the CLT to kick in. This problem is alleviated by choosing σ_n larger, but this, on the other hand, compromises the convergence rate. Ideally, a Berry-Essen type of bound will quantify this, but this will require a different set of techniques and is left as an open problem. In our simulations, we generate data from following setup:

1. Set $N = 50000, p = 3, \alpha_0 = 0.25, \beta = 0.75$ and some $\theta_0 \in \mathbb{R}^p$ with first co-ordinate $= 1$.
2. Generate $X_1, \ldots, X_n \sim \mathcal{N}(0, I_p)$.
3. Generate $Y_i \sim \textbf{Bernoulli}\left(\alpha_0 \mathbb{1}_{X_i^\top \theta_0 \leq 0} + \beta_0 \mathbb{1}_{X_i^\top \theta_0 > 0}\right)$.
4. Estimate $\hat{\theta}$ by minimizing $\mathbb{M}_n(\theta)$ (replacing γ by $\bar{Y}$) based on $\{(X_i, Y_i)\}_{i=1}^n$ for different choices of σ_n.

We repeat Step 2–Step 4 a hundred times to obtain $\hat{\theta}_1, \ldots, \hat{\theta}_{100}$. Define s_n to be the standard deviation of $\{\hat{\theta}_i\}_{i=1}^{100}$. Figures 2 and 3 show the qqplots of $\tilde{\theta}_i = (\hat{\theta}_i - \theta_0)/s_n$ against the standard normal for four different choices of $\sigma_n = n^{-0.6}, n^{-0.7}, n^{-0.8}, n^{-0.9}$. It is evident that smaller value of σ_n yield a poor normal approximation. Although our theory shows that asymptotic normality holds as long as $n\sigma_n \to \infty$, in practice we recommend choosing σ_n such that $n\sigma_n \geq 30$ for the central limit of theorem to take effect.

5. Real data analysis

We illustrate our method using cross-country data on pollution (carbon-dioxide), income and urbanization obtained from the World Development Indicators, World Bank. The Environmental Kuznets Curve hypothesis (EKC henceforth), a popular and ongoing area of research in environmental economics, posits that at an initial stage of economic development pollution increases with economic growth, and then diminishes when society's priorities change, leading to an inverted U-shaped relation between income (measured via real GDP per capita) and pollution. The hypothesis has led to numerous empirical papers (i) testing the hypothesis (whether the relation is inverted U-shaped for countries/regions of interest in the sample), (ii) exploring the threshold level of income at which pollution starts falling, as well as (iii) examining the countries/regions which belong to the upward rising part versus the downward sloping part of the inverted U-shape, if at all. The studies have been performed using US state level data or cross-country data (e.g. [26, 22, 1, 17, 4, 21, 9, 3, 2, 28] to name a few). While some of these papers have found evidence in favor of the EKC hypothesis (inverted U-shaped income-pollution relation), others have found evidence against

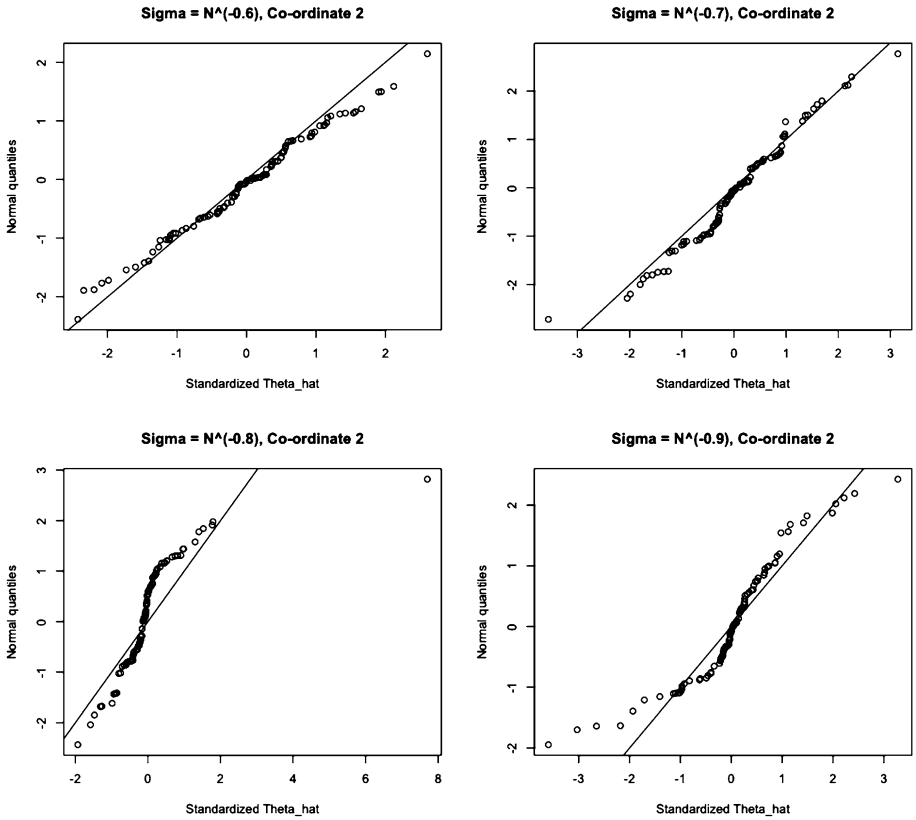


FIG 2. *In this figure, we present qqplot for estimating second co-ordinate of* θ_0 *with different choices of* σ_n *mentioned at the top of each plots.*

it (monotonically increasing or other shapes for the relation). The results often depend on countries/regions in the sample, period of analysis, as well as the pollutant studied.

While income-pollution remains the focal point of most EKC studies, several of them have also included urban agglomeration (UA) or some other measures of urbanization as an important control variable especially while investigating carbon emissions.[1] (See for example, [26, 4] and [19].) The theory of ecological economics posits potentially varying effects of increased urbanization on pollution – (i) urbanization leading to more pollution (due to its close links with sanitations, dense transportations, and proximities to polluting manufacturing

[1]Although income growth is connected to urbanization, countries are heterogenous and follow different growth paths due to their varying geographical structures, population densities, infrastructures, ownerships of resources making a case for using urbanization as another control covariate in the income-pollution study. The income growth paths of oil rich UAE, manufacturing based China, serviced based Singapore, low population density Canada (with vast land) are all different.

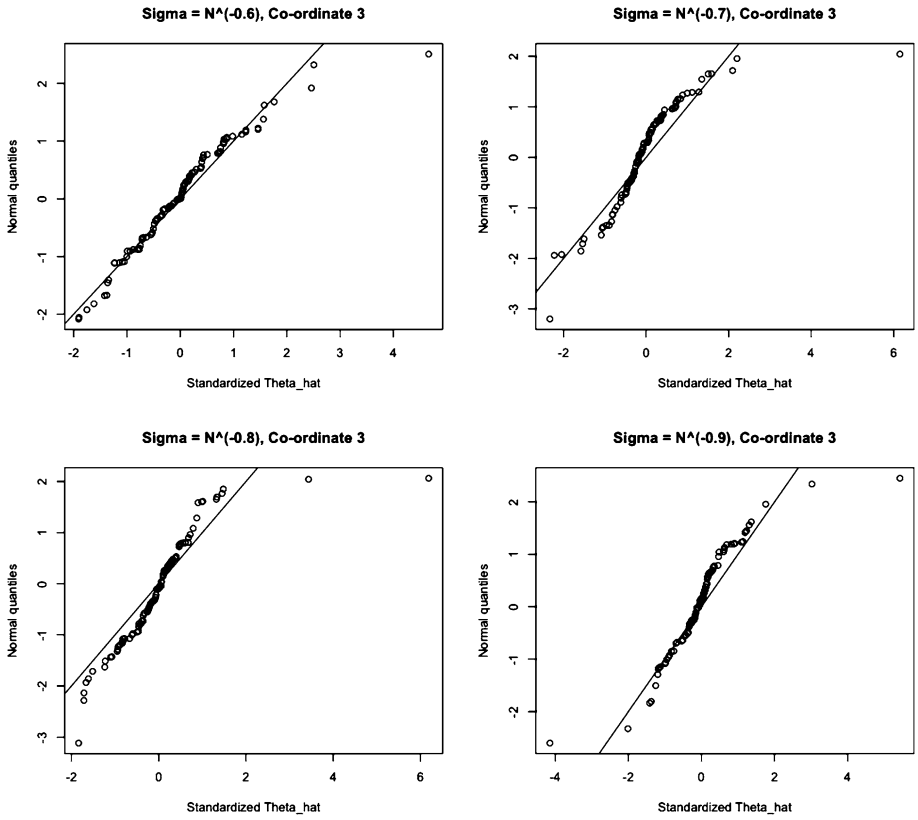


FIG 3. *In this figure, we present qqplot for estimating third co-ordinate of* θ_0 *with different choices of* σ_n *mentioned at the top of each plots.*

industries), (ii) urbanization potentially leading to less pollution based on 'compact city theory' (see [5, 6, 24]) that explains the potential benefits of increased urbanization in terms of economies of scale (for example, replacing dependence on automobiles with large scale subway systems, using multi-storied buildings instead of single unit houses, keeping more open green space). [20], using 17 developed countries, find a positive and significant effect of urbanization on pollution. On the contrary, using a set of 69 countries [27] find a negative and significant effect of urbanization on pollution while [7] find an insignificant effect of urbanization on carbon emission. Using various empirical strategies [24] conclude that the positive and negative effects of urbanization on carbon pollution may cancel out depending on the countries involved often leaving insignificant effects on pollution. They also note that many countries are yet to achieve a sizeable level of urbanization which presumably explains why many empirical works using less developed countries find insignificant effect of urbanization. In summary, based on the existing literature, both the relationship between

urbanization and pollution as well as the relationship between income and pollution appear to depend largely on the set of countries considered in the sample. This motivates us to use UA along with income in our change plane model for analyzing carbon-dioxide emission to plausibly separate the countries into two regimes.

Following the broad literature we use pollution emission per capita (carbon-dioxide measured in metric tons per capita) as the dependent variable and real GDP per capita (measured in 2010 US dollars), its square (as is done commonly in the EKC literature) and a popular measure of urbanization, namely urban agglomeration (UA)[2] as covariates (in our notation X) in our regression. In light of the preceding discussions we fit a change plane model comprising real GDP per capita and UA (in our notation Q). To summarize the setup, we use the continuous response model as described in equation (1.4), i.e

$$\begin{aligned} Y_i &= X_i^\top \beta_0 + X_i^\top \delta_0 \mathbb{1}_{Q_i^\top \psi_0 > 0} + \epsilon_i \\ &= X_i^\top \beta_0 \mathbb{1}_{Q_i^\top \psi_0 \le 0} + X_i^\top (\beta_0 + \delta_0) \mathbb{1}_{Q_i^\top \psi_0 > 0} + \epsilon_i \end{aligned}$$

with the per capita CO_2 emission in metric ton as Y, per capita GDP, square of per capita GDP and UA as X (hence $X \in \mathbb{R}^3$) and finally, per capita GDP and UA as Q (hence $Q \in \mathbb{R}^2$). Observe that β_0 represents the regression coefficients corresponding to the countries with $Q_i^\top \psi_0 \le 0$ (henceforth denoted by Group 1) and $(\beta_0 + \delta_0)$ represents the regression coefficients corresponding to the countries with $Q_i^\top \psi_0 \ge 0$ (henceforth denoted by Group 2). As per our convention, in the interests of identifiability we assume $\psi_{0,1} = 1$, where $\psi_{0,1}$ is the change plane parameter corresponding to per capita GDP. Therefore the only change plane coefficient to be estimated is $\psi_{0,2}$, the change plane coefficient for UA. For numerical stability, we divide per capita GDP by 10^{-4} (consequently square of per capital GDP is scaled by 10^{-8}).[3] After some pre-processing (i.e. removing rows consisting of NA and countries with 100% UA) we estimate the coefficients $(\beta_0, \delta_0, \psi_0)$ of our model based on data from 115 countries with $\sigma_n = 0.05$ and test the significance of the various coefficients using the methodologies described in Section 3. We present our findings in Table 1.

TABLE 1

Table of the estimated regression and change plane coefficients along with their p-values.

Coefficients	Estimated values	p-values
$\beta_{0,1}$ (RGDPPC for Group 1)	6.98555060	4.961452e-10
$\beta_{0,2}$ (squared RGDPPC for Group 1)	−0.43425991	7.136484e-02
$\beta_{0,3}$ (UA for Group 1)	−0.02613813	1.066065e-01
$\beta_{0,1} + \delta_{0,1}$ (RGDPPC for Group 2)	2.0563337	0.000000e+00
$\beta_{0,2} + \delta_{0,2}$ (squared RGDPPC for Group 2)	−0.1866490	4.912843e-04
$\beta_{0,3} + \delta_{0,3}$ (UA for Group 2)	0.1403171	1.329788e-05
$\psi_{0,2}$ (Change plane coeff for UA)	−0.07061785	0.000000e+00

[2]The exact definition can be found in the World Development Indicators database from the World Bank website.

[3]This scaling helps in the numerical stability of the gradient descent algorithm used to optimize the least squares criterion.

From the above analysis, we find that GDP has significantly positive effect on pollution for both groups of countries. The effect of its squared term is negative for both groups; but the effect is significant for Group-2 consisting of mostly high income countries whereas its effect is insignificant (at the 5% level) for the Group-1 countries (consisting of mostly low or middle income and few high income countries). Thus, not surprisingly, we find evidence in favor of EKC for the developed countries, but not for the mixed group. Notably, Group-1 consists of a mixed set of countries like Angola, Sudan, Senegal, India, China, Israel, UAE etc., whereas Group-1 consists of rich and developed countries like Canada, USA, UK, France, Germany etc. The urban variable, on the other hand, is seen to have insignificant effect on Group-1 which is in keeping with [7, 24]. Many of them are yet to achieve substantial urbanization and this is more true for our sample period.[4] In contrast, UA has a positive and significant effect on Group-2 (developed) countries which is consistent with the findings of [20], for example. Note that UA plays a crucial role in dividing the countries into different regimes, as the estimated value of $\psi_{0,2}$ is significant. Thus, we are able to partition countries into two regimes: a mostly rich and a mixed group.

Note that many underdeveloped countries and poorer regions of emerging countries are still swamped with greenhouse gas emissions from burning coal, cow dung etc., and usage of poor exhaust systems in houses and for transport. This is more true for rural and semi-urban areas of developing countries. So even while being less urbanized compared to developed nations, their overall pollution load is high (due to inefficient energy usage and higher dependence on fossil fuels as pointed out above) and rising with income and they are yet to reach the descending part of the inverted U-shape for the income-pollution relation. On the contrary, for countries in Group-2, the adoption of more efficient energy and exhaust systems are common in households and transportations in general, leading to eventually decreasing pollution with increasing income (supporting EKC). Both the results are in line with the existing EKC literature. Additionally we find that the countries in Group 2 are yet to achieve 'compact city' and green urbanization. This is a stylized fact that is confirmed by the positive and significant effect of UA on pollution in our analysis.

There are many future potential applications of our method in economics. Similar analyses can be performed for other pollutants (such as sulfur emission, electrical waste/e-waste, nitrogen pollution etc.). While income/GDP remains a common, indeed the most crucial variable in pollution studies, other covariates (including change plane defining variables) may vary, depending on the pollutant of interest. Another potential application can be that of identifying the determinants of family health expenses in household survey data. Families are often asked about their health expenses incurred in the past one year. An interesting case in point may be household surveys collected in India where one

[4] We use 6 years average from 2010–2015 for GDP and pollution measures. Such averaging is in accordance with the cross-sectional empirical literature using cross-country/regional data and helps avoid business cycle fluctuations in GDP. It also minimizes the impacts of outlier events such as the financial crisis or great recession period. The years that we have chosen are ones for which we could find data for the largest number of countries.

finds numerous (large) joint families with several children and old people residing in the same household and most families are uninsured. It is often seen that health expenditure increases with income with a major factor being the costs associated with regularly performed preventative medical examinations which are affordable only once a certain income level is reached. The important covariates here are per capita family income, family wealth, 'dependency ratio' (number of children and old to the total number of people in the family) and the binary indicator of any history of major illness/hospitalizations in the family in the past year. Family income per capita and history of major illness are natural candidate covariates for defining the change plane.

6. Conclusion and future research

In this paper we have established that under some mild assumptions the kernel-smoothed change plane estimator is asymptotically normal with near optimal rate n^{-1}. To the best of our knowledge, the state of the art result in this genre of problems is due to [25], where they demonstrate a best possible rate about $n^{-3/4}$ for i.i.d. data. The main difference between their approach and ours is mainly the proof of Lemma 2.7. Our techniques are based upon modern empirical process theory which allow us to consider much smaller bandwidths σ_n compared to those in [25], who appear to require larger values to achieve the result, possibly owing to their reliance on the techniques developed in [12]. Although we have established it is possible to have asymptotic normality with really small bandwidths, we believe that the finite sample approximation (e.g. Berry-Essen bound) to normality could be poor, which is also evident from our simulation.

There are some natural directions in which the model proposed in this paper can be extended. We describe some of these below for the classification model:

(i) Assume Y is Bernoulli and the success probability can take k different values, i.e.

$$\mathbb{P}(Y = 1 \mid Q) = \sum_{k=1}^{K} \alpha_k \mathbb{1}_{\gamma_{k-1} \le Q^\top \psi \le \gamma_k},$$

for some sequence $\{\gamma_k\}_{0 \le k \le K}$ with $\gamma_0 = -\infty$ and $\gamma_k = \infty$.

(ii) Assume Y takes C different values with the conditional multinomial distribution changing across a hyperplane: for $1 \le j \le C$,

$$\mathbb{P}\left(Y = j \mid Q\right) = \alpha_j \mathbb{1}_{Q^\top \psi \le 0} + \beta_j \mathbb{1}_{Q^\top \psi > 0}$$

where $\sum_j \alpha_j = \sum_j \beta_j = 1$. Further, one can combine (i) and (ii) where Y is a multinomial and the probability has K structural changes depending on the magnitude of $Q^\top \psi$. We believe that similar regularity conditions to those used in the paper will yield asymptotic normality for the change hyperplanes in this multinomial problem.

Another interesting research direction is to analyze the kernel smoothed estimator in the growing dimensional regime, i.e. when the dimension of the change

plane grows with n (or may be larger than n). In this case, the proof will be more tricky, especially if one wants to quantify the minimax effect of dimension. Recently, the authors of this paper ([23]) have established the minimax optimal rate of change plane estimation in the growing dimension regime and showed that the least squares estimate (in the presence of light-tailed error) and the Huber estimate (in the presence of heavy tailed error) are minimax optimal (up to log factors). Whether the kernel-smoothed estimate (almost) attains that minimax optimal rate under an appropriate bandwidth choice remains an open question.

Appendix A

In this appendix, we present the proof of Lemma 2.6, which lies at the heart of our refined analysis of the smoothed change plane estimator. Proofs of the other lemmas and our results for the binary response model are available in the supplementary document.

A.1. Proof of Lemma 2.6

Proof. The proof of Lemma 2.6 is quite long, hence we further break it into few more lemmas.

Lemma A.1. *Under Assumption (2.1), there exists $u_-, u_+ > 0$ such that:*

$$u_- d^2(\theta, \theta_0) \leq \mathbb{M}(\theta) - \mathbb{M}(\theta_0) \leq u_+ d^2(\theta, \theta_0),$$

for θ in a (non-srinking) neighborhood of θ_0, where:

$$d(\theta, \theta_0) := \sqrt{\|\beta - \beta_0\|^2 + \|\delta - \delta_0\|^2 + \|\psi - \psi_0\|}.$$

Lemma A.2. *Under Assumption 2.1 the smoothed loss function $\mathbb{M}^s(\theta)$ is uniformly close to the non-smoothed loss function $\mathbb{M}(\theta)$:*

$$\sup_{\theta \in \Theta} |\mathbb{M}^s(\theta) - \mathbb{M}(\theta)| \leq K_1 \sigma_n,$$

for some constant K_1.

Lemma A.3. *Under certain assumptions:*

$$\begin{aligned}\mathbb{M}^s(\theta) - \mathbb{M}^s(\theta_0^s) \gtrsim \|\beta - \beta_0^s\|^2 + \|\delta - \delta_0^s\|^2 \\ + \frac{\|\psi - \psi_0^s\|^2}{\sigma_n} \mathbb{1}_{\|\psi - \psi_0^s\| \leq \mathcal{K}\sigma_n} + \|\psi - \psi_0^s\| \mathbb{1}_{\|\psi - \psi_0^s\| > \mathcal{K}\sigma_n} \\ := d_*^2(\theta, \theta_0^s),\end{aligned}$$

for some constant $\mathcal{K}$ and for all θ in a neighborhood of θ_0, which does not change with n.

The proofs of the three lemmas above can be found in Section 1 the supplementary document. In Lemma A.3 we have established the curvature of the smooth loss function $\mathbb{M}^s(\theta)$ around θ_0^s. To determine the rate of convergence of $\hat{\theta}^s$ to θ_0^s, we further need an upper bound on the modulus of continuity of our loss function. Towards that end, first recall that our loss function is:

$$f_\theta(Y,X,Q) = \left(Y - X^\top\beta\right)^2 + \left[-2\left(Y - X^\top\beta\right)X^\top\delta + (X^\top\delta)^2\right] K\left(\frac{Q^\top\psi}{\sigma_n}\right)$$

The centered loss function can be written as:

$$\begin{aligned}
&f_\theta(Y,X,Q) - f_{\theta_0^s}(Y,X,Q)\\
&= \left(Y - X^\top\beta\right)^2 + \left[-2\left(Y - X^\top\beta\right)X^\top\delta + (X^\top\delta)^2\right] K\left(\frac{Q^\top\psi}{\sigma_n}\right)\\
&- \left(Y - X^\top\beta_0^s\right)^2 - \left[-2\left(Y - X^\top\beta_0^s\right)X^\top\delta_0^s + (X^\top\delta_0^s)^2\right] K\left(\frac{Q^\top\psi_0^s}{\sigma_n}\right)\\
&= \left(Y - X^\top\beta\right)^2 + \left[-2\left(Y - X^\top\beta\right)X^\top\delta + (X^\top\delta)^2\right] K\left(\frac{Q^\top\psi}{\sigma_n}\right)\\
&- \left(Y - X^\top\beta_0^s\right)^2 - \left[-2\left(Y - X^\top\beta_0^s\right)X^\top\delta_0^s + (X^\top\delta_0^s)^2\right] K\left(\frac{Q^\top\psi}{\sigma_n}\right)\\
&- \left[-2\left(Y - X^\top\beta_0^s\right)X^\top\delta_0^s + (X^\top\delta_0^s)^2\right]\left\{K\left(\frac{Q^\top\psi_0^s}{\sigma_n}\right) - K\left(\frac{Q^\top\psi}{\sigma_n}\right)\right\}\\
&= \underbrace{\left(Y - X^\top\beta\right)^2 - \left(Y - X^\top\beta_0^s\right)^2}_{M_1}\\
&+ \underbrace{\left\{\left[-2\left(Y - X^\top\beta\right)X^\top\delta + (X^\top\delta)^2\right] - \left[-2\left(Y - X^\top\beta_0^s\right)X^\top\delta_0^s + (X^\top\delta_0^s)^2\right]\right\} \times K\left(\frac{Q^\top\psi}{\sigma_n}\right)}_{M_2}\\
&- \underbrace{\left[-2\left(Y - X^\top\beta_0^s\right)X^\top\delta_0^s + (X^\top\delta_0^s)^2\right]\left\{K\left(\frac{Q^\top\psi_0^s}{\sigma_n}\right) - K\left(\frac{Q^\top\psi}{\sigma_n}\right)\right\}}_{M_3}\\
&:= M_1 + M_2 + M_3
\end{aligned} \tag{A.1}$$

For the rest of the analysis, fix $\zeta > 0$ and consider the collection of functions $\mathcal{F}_\zeta$ which is defined as:

$$\mathcal{F}_\zeta = \{f_\theta - f_{\theta^s} : d_*(\theta,\theta^s) \le \zeta\}.$$

First note that $\mathcal{F}_\zeta$ has bounded uniform entropy integral (henceforth BUEI) over ζ. To establish this, it is enough to argue that the collection $\mathcal{F} = \{f_\theta : \theta \in \Theta\}$ is BUEI. Note that the functions $X \mapsto X^\top\beta$ has VC dimension p and so is the map $X \mapsto X^\top(\beta+\delta)$. Therefore the functions $(X,Y) \mapsto (Y -$

$X^\top(\beta+\delta))^2-(Y-X^\top\beta)^2$ is also BUEI, as composition with monotone function (here x^2) and taking difference keeps this property. Further by the hyperplane $Q \mapsto Q^\top\psi$ also has finite dimension (only depends on the dimension of Q) and the VC dimension does not change by scaling it with σ_n. Therefore the functions $Q \mapsto Q^\top\psi/\sigma_n$ has same VC dimension as $Q \mapsto Q^\top\psi$ which is independent of n. Again, as composition of monotone function keeps BUEI property, the functions $Q \mapsto K(Q^\top\psi/\sigma_n)$ is also BUEI. As the product of two BUEI class is BUEI, we conclude that $\mathcal{F}$ (and hence $\mathcal{F}_\zeta$) is BUEI.

Now to bound the modulus of continuity we use Lemma 2.14.1 of [29]:

$$\sqrt{n}\mathbb{E}\left[\sup_{\theta:d_*(\theta,\theta_0^s)\le\zeta}\left|(\mathbb{P}_n-P)\left(f_\theta-f_{\theta_0^s}\right)\right|\right] \lesssim \mathcal{J}(1,\mathcal{F}_\zeta)\sqrt{\mathbb{E}\left[F_\zeta^2(X,Y,Q)\right]}$$

where F_ζ is some envelope function of $\mathcal{F}_\zeta$. As the function class $\mathcal{F}_\zeta$ has bounded entropy integral, $\mathcal{J}(1,\mathcal{F}_\zeta)$ can be bounded above by some constant independent of n. We next calculate the order of the envelope function F_ζ. Recall that, by definition of envelope function is:

$$F_\zeta(X,Y,Q) \ge \sup_{\theta:d_*(\theta,\theta_0^s)\le\zeta}\left|f_\theta-f_{\theta_0}\right|,$$

and we can write $f_\theta - f_{\theta_0^s} = M_1+M_2+M_3$ which follows from equation (A.1). Therefore, to find the order of the envelope function, it is enough to find the order of bounds of M_1, M_2, M_3 over the set $d_*(\theta,\theta_0^s)\le\zeta$. We start with M_1:

$$\begin{aligned}
&\sup_{d_*(\theta,\theta_0^s)\le\zeta}|M_1|\\
&=\sup_{d_*(\theta,\theta_0^s)\le\delta}\left|\left(Y-X^\top\beta\right)^2-\left(Y-X^\top\beta_0^s\right)^2\right|\\
&=\sup_{d_*(\theta,\theta_0^s)\le\zeta}\left|2YX^\top(\beta_0^s-\beta)+(X^\top\beta)^2-(X^\top\beta_0^S)^2\right|\\
&\le\sup_{d_*(\theta,\theta_0^s)\le\zeta}\|\beta-\beta_0^s\|\left[2|Y|\|X\|+(\|\beta_0^s\|+\zeta)\|X\|^2\right]\\
&\le\zeta\left[2|Y|\|X\|+(\|\beta_0^s\|+\zeta)\|X\|^2\right]:=F_{1,\zeta}(X,Y,Q)\quad\text{[Envelope function of } M_1\text{]}
\end{aligned}\tag{A.2}$$

and the second term:

$$\begin{aligned}
&\sup_{d_*(\theta,\theta_0^s)\le\zeta}|M_2|\\
&=\sup_{d_*(\theta,\theta_0^s)\le\zeta}\Big|\left\{\left[-2\left(Y-X^\top\beta\right)X^\top\delta+(X^\top\delta)^2\right]\right.\\
&\qquad\qquad\left.-\left[-2\left(Y-X^\top\beta_0^s\right)X^\top\delta_0^s+(X^\top\delta_0^s)^2\right]\right\}\Big|\,K\left(\frac{Q^\top\psi}{\sigma_n}\right)\\
&\le\sup_{d_*(\theta,\theta_0^s)\le\zeta}\Big|\left\{\left[-2\left(Y-X^\top\beta\right)X^\top\delta+(X^\top\delta)^2\right]\right.\\
&\qquad\qquad\left.-\left[-2\left(Y-X^\top\beta_0^s\right)X^\top\delta_0^s+(X^\top\delta_0^s)^2\right]\right\}\Big|
\end{aligned}$$

$$
\begin{aligned}
&= \sup_{d_*(\theta,\theta_0^s)\le\zeta} \left|\left\{\left[2Y(X^\top\delta_0^s - X^\top\delta) + 2[(X^\top\beta)(X^\top\delta)\right.\right.\right. \\
&\qquad\qquad \left.\left.\left. -(X^\top\beta_0^s)(X^\top\delta_0^s)] + (X^\top\delta)^2 - (X^\top\delta_0^s)^2\right]\right\}\right| \\
&\le \sup_{d_*(\theta,\theta_0^s)\le\zeta} \{\|\delta-\delta_0^s\|2|Y|\|X\| + 2\|\beta-\beta_0\|\|X\|\|\delta\| \\
&\qquad\qquad +2\|\delta-\delta_0^s\|\|X\|\|\beta_0^s\| + 2\|X\|\|\delta+\delta_0^s\|\|\delta-\delta_0^s\|\} \\
&\le \zeta\,[2|Y|\|X\| + 2\|X\|(\|\delta_0^s\| + \|\zeta\|) + 2\|X\|\|\beta_0^s\| + 2\|X\|(\|\delta_0^s\| + \zeta)] \\
&= \zeta \times 2\|X\|\,[2|Y| + 2(\|\delta_0^s\| + \|\zeta\|) + \|\beta_0^s\|] := F_{2,\zeta}(X,Y,Q) \qquad \text{(A.3)}
\end{aligned}
$$

For the third term, note that:

$$
\begin{aligned}
&\sup_{d_*(\theta,\theta_0^s)\le\zeta} |M_3| \\
&\le \left|\left[-2\left(Y - X^\top\beta_0^s\right)X^\top\delta_0^s + (X^\top\delta_0^s)^2\right]\right| \\
&\qquad \times \sup_{d_*(\theta,\theta_0^s)\le\zeta} \left|\left\{K\left(\frac{Q^\top\psi_0^s}{\sigma_n}\right) - K\left(\frac{Q^\top\psi}{\sigma_n}\right)\right\}\right| \\
&:= F_{3,\zeta}(X,Y,Q)
\end{aligned}
$$

Henceforth, we define the envelope function to be $F_\zeta = F_{\zeta,1} + F_{\zeta,2} + F_{\zeta,3}$. Hence we have by triangle inequality:

$$
\sqrt{\mathbb{E}\left[F_\zeta^2(X,Y,Q)\right]} \le \sum_{i=1}^{3} \sqrt{\mathbb{E}\left[F_{i,\zeta}^2(X,Y,Q)\right]}
$$

From equation (A.2) and (A.3) we have:

$$
\sqrt{\mathbb{E}\left[F_{1,\zeta}^2(X,Y,Q)\right]} + \sqrt{\mathbb{E}\left[F_{2,\zeta}^2(X,Y,Q)\right]} \lesssim \zeta. \qquad \text{(A.4)}
$$

For $F_{3,\zeta}$, first note that:

$$
\begin{aligned}
&\mathbb{E}\left[\left|\left[-2\left(Y - X^\top\beta_0^s\right)X^\top\delta_0^s + (X^\top\delta_0^s)^2\right]\right|^2 \mid Q\right] \\
&\le 8\mathbb{E}\left[\left(Y - X^\top\beta_0^s\right)^2 (X^\top\delta_0)^2 \mid Q\right] + 2\mathbb{E}[(X^\top\delta_0^s)^4 \mid Q] \\
&\le \left\{8\|\beta-\beta_0^s\|^2\|\delta_0\|^2 + 8\|\delta_0\|^4 + 2\|\delta_0^s\|^4\right\} m_4(Q),
\end{aligned}
$$

where $m_4(Q)$ is defined in Assumption 2.1. In this part, we have to tackle the dichotomous behavior of ψ around ψ_0^s carefully. Henceforth define $d_*^2(\psi,\psi_0^s)$ as:

$$
d_*^2(\psi,\psi_0^s) = \frac{\|\psi-\psi_0^s\|^2}{\sigma_n}\mathbb{1}_{\|\psi-\psi_0^s\|\le\mathcal{K}\sigma_n} + \|\psi-\psi_0^s\|\mathbb{1}_{\|\psi-\psi_0^s\|>\mathcal{K}\sigma_n}
$$

This is a slight abuse of notation, but the reader should think of it as the part of ψ in $d_*^2(\theta,\theta_0^s)$. Define $B_\zeta(\psi_0^s)$ to be set of all ψ's such that $d_*^2(\psi,\psi_0^s) \le \zeta^2$. We can decompose $B_\zeta(\psi_0^s)$ as a disjoint union of two sets:

$$
B_{\zeta,1}(\psi_0^s) = \left\{\psi : d_*^2(\psi,\psi_0^s) \le \zeta^2, \|\psi-\psi_0^s\| \le \mathcal{K}\sigma_n\right\}
$$

$$
\begin{aligned}
&= \left\{\psi : \frac{\|\psi - \psi_0^s\|^2}{\sigma_n} \leq \zeta^2, \|\psi - \psi_0^s\| \leq \mathcal{K}\sigma_n\right\} \\
&= \{\psi : \|\psi - \psi_0^s\| \leq \zeta\sqrt{\sigma_n}, \|\psi - \psi_0^s\| \leq \mathcal{K}\sigma_n\} \\
B_{\zeta,2}(\psi_0^s) &= \left\{\psi : d_*^2(\psi, \psi_0^s) \leq \zeta^2, \|\psi - \psi_0^s\| > \mathcal{K}\sigma_n\right\} \\
&= \left\{\psi : \|\psi - \psi_0^s\| \leq \zeta^2, \|\psi - \psi_0^s\| > \mathcal{K}\sigma_n\right\}
\end{aligned}
$$

Assume $\mathcal{K} > 1$. The case where $\mathcal{K} < 1$ follows from similar calculations and hence skipped for brevity. Consider the following two cases:

Case 1: Suppose $\zeta \leq \sqrt{\mathcal{K}\sigma_n}$. Then $B_{\zeta,2} = \phi$. Also as $\mathcal{K} > 1$, we have: $\zeta\sqrt{\sigma_n} \leq \mathcal{K}\sigma_n$. Hence we have:

$$
\sup_{d_*^2(\psi,\psi_0^s)\leq\zeta^2} \|\psi - \psi_0^s\| = \sup_{B_{\zeta,1}} \|\psi - \psi_0^s\| = \zeta\sqrt{\sigma_n}.
$$

This implies:

$$
\begin{aligned}
&\sup_{d_*(\theta,\theta_0^s)\leq\zeta} \left|\left\{K\left(\frac{Q^\top\psi_0^s}{\sigma_n}\right) - K\left(\frac{Q^\top\psi}{\sigma_n}\right)\right\}\right|^2 \\
&\leq \max\left\{\left|\left\{K\left(\frac{Q^\top\psi_0^s}{\sigma_n}\right) - K\left(\frac{Q^\top\psi_0^s}{\sigma_n} + \|\tilde{Q}\|\frac{\zeta}{\sqrt{\sigma_n}}\right)\right\}\right|^2,\right. \\
&\qquad\qquad \left.\left|\left\{K\left(\frac{Q^\top\psi_0^s}{\sigma_n}\right) - K\left(\frac{Q^\top\psi_0^s}{\sigma_n} - \|\tilde{Q}\|\frac{\zeta}{\sqrt{\sigma_n}}\right)\right\}\right|^2\right\} \\
&:= \max\{T_1, T_2\}.
\end{aligned}
$$

Therefore we have:

$$
\mathbb{E}\left[F_{3,\zeta}^2(X, Y, Q)\right] \leq \mathbb{E}[m_4(Q)T_1] + \mathbb{E}[m_4(Q)T_2].
$$

Now:

$$
\begin{aligned}
&\mathbb{E}[m_4(Q)T_1] \\
&= \mathbb{E}\left[m_4(Q)\left|\left\{K\left(\frac{Q^\top\psi_0^s}{\sigma_n}\right) - K\left(\frac{Q^\top\psi_0^s}{\sigma_n} + \|\tilde{Q}\|\frac{\zeta}{\sqrt{\sigma_n}}\right)\right\}\right|^2\right] \\
&= \sigma_n \int_{\mathbb{R}^{p-1}} \int_{-\infty}^{\infty} m_4(\sigma_n t - \tilde{q}^\top\tilde{\psi}_0^s, \tilde{q})\left|K(t) - K\left(t + \|\tilde{q}\|\frac{\zeta}{\sqrt{\sigma_n}}\right)\right|^2 \\
&\qquad \times f_s(\sigma_n t \mid \tilde{q})\, dt\ f(\tilde{q})\, d\tilde{q} \\
&\leq \sigma_n \int_{\mathbb{R}^{p-1}} \int_{-\infty}^{\infty} m_4(\sigma_n t - \tilde{q}^\top\tilde{\psi}_0^s, \tilde{q})\left|K(t) - K\left(t + \|\tilde{q}\|\frac{\zeta}{\sqrt{\sigma_n}}\right)\right| \\
&\qquad \times f_s(\sigma_n t \mid \tilde{q})\, dt\ f(\tilde{q})\, d\tilde{q} \\
&= \sigma_n \int_{\mathbb{R}^{p-1}} \int_{-\infty}^{\infty} m_4(\sigma_n t - \tilde{q}^\top\tilde{\psi}_0^s, \tilde{q}) \int_t^{t+\|\tilde{q}\|\frac{\zeta}{\sqrt{\sigma_n}}} K'(s)\, ds
\end{aligned}
$$

$$
\begin{aligned}
&\quad \times f_s(\sigma_n t \mid \tilde{q})\, dt\; f(\tilde{q})\, d\tilde{q} \\
&= \sigma_n \int_{\mathbb{R}^{p-1}} \int_{-\infty}^{\infty} K'(s) \int_{s-\|\tilde{q}\| \frac{\zeta}{\sqrt{\sigma_n}}}^{s} m_4(\sigma_n t - \tilde{q}^\top \tilde{\psi}_0^s, \tilde{q}) \\
&\quad \times f_s(\sigma_n t \mid \tilde{q})\, dt\, ds\; f(\tilde{q})\, d\tilde{q} \\
&= \zeta \sqrt{\sigma_n} \mathbb{E}[\|\tilde{Q}\| m_4(-\tilde{Q}^\top \psi_0^s, \tilde{Q}) f_s(0 \mid \tilde{Q})] + R
\end{aligned}
$$

where as before we split R into three parts $R = R_1 + R_2 + R_3$.

$$
\begin{aligned}
&|R_1| \qquad\qquad (A.5)\\
&= \left| \sigma_n \int_{\mathbb{R}^{p-1}} \int_{-\infty}^{\infty} K'(s) \int_{s-\|\tilde{q}\| \frac{\zeta}{\sqrt{\sigma_n}}}^{s} m_4(-\tilde{q}^\top \tilde{\psi}_0^s, \tilde{q}) (f_s(\sigma_n t \mid \tilde{q}) - f_s(0 \mid \tilde{q}))\, dt\, ds \right. \\
&\qquad \left. f(\tilde{q})\, d\tilde{q} \right| \\
&\le \sigma_n^2 \int_{\mathbb{R}^{p-1}} m_4(-\tilde{q}^\top \tilde{\psi}_0^s, \tilde{q}) \dot{f}_s(\tilde{q}) \int_{-\infty}^{\infty} K'(s) \int_{s-\|\tilde{q}\| \frac{\zeta}{\sqrt{\sigma_n}}}^{s} |t|\, dt\, ds\; f(\tilde{q})\, d\tilde{q} \qquad (A.6)
\end{aligned}
$$

We next calculate the inner integral (involving (s,t)) of equation (A.6):

$$
\begin{aligned}
&\int_{-\infty}^{\infty} K'(s) \int_{s-\|\tilde{q}\| \frac{\zeta}{\sqrt{\sigma_n}}}^{s} |t|\, dt\, ds \\
&= \left(\int_{-\infty}^{0} + \int_{0}^{\|\tilde{q}\| \frac{\zeta}{\sqrt{\sigma_n}}} + \int_{\|\tilde{q}\| \frac{\zeta}{\sqrt{\sigma_n}}}^{\infty} \right) K'(s) \int_{s-\|\tilde{q}\| \frac{\zeta}{\sqrt{\sigma_n}}}^{s} |t|\, dt\, ds \\
&= \frac{1}{2} \int_{-\infty}^{0} K'(s) \left[\left(s - \|\tilde{q}\| \frac{\zeta}{\sqrt{\sigma_n}} \right)^2 - s^2 \right] ds \\
&\quad + \frac{1}{2} \int_{0}^{\|\tilde{q}\| \frac{\zeta}{\sqrt{\sigma_n}}} K'(s) \left[\left(s - \|\tilde{q}\| \frac{\zeta}{\sqrt{\sigma_n}} \right)^2 + s^2 \right] ds \\
&\qquad + \frac{1}{2} \int_{\|\tilde{q}\| \frac{\zeta}{\sqrt{\sigma_n}}}^{\infty} K'(s) \left[s^2 - \left(s - \|\tilde{q}\| \frac{\zeta}{\sqrt{\sigma_n}} \right)^2 \right] ds \\
&= -\|\tilde{q}\| \frac{\zeta}{\sqrt{\sigma_n}} \int_{-\infty}^{0} K'(s) s\, ds + \|\tilde{q}\|^2 \frac{\zeta^2}{2\sigma_n} \int_{-\infty}^{0} K'(s)\, ds + \int_{0}^{\|\tilde{q}\| \frac{\zeta}{\sqrt{\sigma_n}}} s^2 K'(s)\, ds \\
&\quad - \|\tilde{q}\| \frac{\zeta}{\sqrt{\sigma_n}} \int_{0}^{\|\tilde{q}\| \frac{\zeta}{\sqrt{\sigma_n}}} s K'(s)\, ds + \|\tilde{q}\|^2 \frac{\zeta^2}{2\sigma_n} \int_{0}^{\|\tilde{q}\| \frac{\zeta}{\sqrt{\sigma_n}}} K'(s)\, ds \\
&\qquad + \|\tilde{q}\| \frac{\zeta}{\sqrt{\sigma_n}} \int_{\|\tilde{q}\| \frac{\zeta}{\sqrt{\sigma_n}}}^{\infty} s K'(s)\, ds - \|\tilde{q}\|^2 \frac{\zeta^2}{2\sigma_n} \int_{\|\tilde{q}\| \frac{\zeta}{\sqrt{\sigma_n}}}^{\infty} K'(s)\, ds \\
&= \|\tilde{q}\|^2 \frac{\zeta^2}{2\sigma_n} \left[2K\left(\|\tilde{q}\| \frac{\zeta}{\sqrt{\sigma_n}} \right) - 1 \right] + \|\tilde{q}\| \frac{\zeta}{\sqrt{\sigma_n}} \left[- \int_{-\infty}^{0} K'(s) s\, ds \right.
\end{aligned}
$$

$$
\begin{aligned}
&\quad - \int_0^{\|\tilde{q}\|\frac{\zeta}{\sqrt{\sigma_n}}} K'(s)s\,ds + \int_{\|\tilde{q}\|\frac{\zeta}{\sqrt{\sigma_n}}}^{\infty} sK'(s)\,ds \Bigg] + \int_0^{\|\tilde{q}\|\frac{\zeta}{\sqrt{\sigma_n}}} s^2K'(s)\,ds \\
&= \|\tilde{q}\|^2\frac{\zeta^2}{\sigma_n}\left[K\left(\|\tilde{q}\|\frac{\zeta}{\sqrt{\sigma_n}}\right) - K(0)\right] \\
&\quad + \|\tilde{q}\|\frac{\zeta}{\sqrt{\sigma_n}}\left[-\int_{-\infty}^{-\|\tilde{q}\|\frac{\zeta}{\sqrt{\sigma_n}}} K'(s)s\,ds + \int_{\|\tilde{q}\|\frac{\zeta}{\sqrt{\sigma_n}}}^{\infty} sK'(s)\,ds\right] \\
&\quad + \int_0^{\|\tilde{q}\|\frac{\zeta}{\sqrt{\sigma_n}}} s^2K'(s)\,ds \\
&= \|\tilde{q}\|^2\frac{\zeta^2}{\sigma_n}\left[K\left(\|\tilde{q}\|\frac{\zeta}{\sqrt{\sigma_n}}\right) - K(0)\right] \\
&\quad + \|\tilde{q}\|\frac{\zeta}{\sqrt{\sigma_n}}\int_{-\infty}^{\infty} K'(s)|s|\mathbb{1}_{|s|\geq\|\tilde{q}\|\frac{\zeta}{\sqrt{\sigma_n}}}\,ds + \int_0^{\|\tilde{q}\|\frac{\zeta}{\sqrt{\sigma_n}}} s^2K'(s)\,ds \\
&\leq \dot{K}_+\|\tilde{q}\|^3\frac{\zeta^3}{\sigma_n^{3/2}} + \|\tilde{q}\|\frac{\zeta}{\sqrt{\sigma_n}}\int_{-\infty}^{\infty} K'(s)|s|\,ds + \|\tilde{q}\|^2\frac{\zeta^2}{\sigma_n}\left(K\left(\|\tilde{q}\|\frac{\zeta}{\sqrt{\sigma_n}}\right) - K(0)\right) \\
&\lesssim \|\tilde{q}\|^3\frac{\zeta^3}{\sigma_n^{3/2}} + \|\tilde{q}\|\frac{\zeta}{\sqrt{\sigma_n}}
\end{aligned}
$$

Putting this bound in equation (A.6) we obtain:

$$
\begin{aligned}
&|R_1| \\
&\leq \frac{\sigma_n^2}{2}\int_{\mathbb{R}^{p-1}} m_4(-\tilde{q}^\top\tilde{\psi}_0^s,\tilde{q})\dot{f}_s(\tilde{q})\left(\|\tilde{q}\|^3\frac{\zeta^3}{\sigma_n^{3/2}} + \|\tilde{q}\|\frac{\zeta}{\sqrt{\sigma_n}}\right) f(\tilde{q})\,d\tilde{q} \\
&\leq \frac{\zeta^3}{2\sqrt{\sigma_n}}\mathbb{E}\left[m_4(-\tilde{Q}^\top\tilde{\psi}_0^s,\tilde{Q})\dot{f}_s(\tilde{Q})\|\tilde{Q}\|^3\right] + \frac{\zeta\sqrt{\sigma_n}}{2}\mathbb{E}\left[m_4(-\tilde{Q}^\top\tilde{\psi}_0^s,\tilde{Q})\dot{f}_s(\tilde{Q})\|\tilde{Q}\|\right]
\end{aligned}
$$

and

$$
\begin{aligned}
&|R_2| \\
&= \Bigg|\sigma_n\int_{\mathbb{R}^{p-1}}\int_{-\infty}^{\infty} K'(s)\int_{s-\|\tilde{q}\|\frac{\zeta}{\sqrt{\sigma_n}}}^{s}\left(m_4(\sigma_n t - \tilde{q}^\top\tilde{\psi}_0^s,\tilde{q}) - m_4(-\tilde{q}^\top\tilde{\psi}_0^s,\tilde{q})\right) \\
&\qquad\qquad \times f_s(0\mid\tilde{q})\,dt\,ds\;f(\tilde{q})\,d\tilde{q}\Bigg| \\
&\leq \sigma_n^2\int_{\mathbb{R}^{p-1}}\dot{m}_4(\tilde{q})f_s(0\mid\tilde{q})\int_{-\infty}^{\infty} K'(s)\int_{s-\|\tilde{q}\|\frac{\zeta}{\sqrt{\sigma_n}}}^{s}|t|\,dt\,ds\;f(\tilde{q})\,d\tilde{q} \\
&\leq \sigma_n^2\int_{\mathbb{R}^{p-1}}\dot{m}_4(\tilde{q})f_s(0\mid\tilde{q})\left(\|\tilde{q}\|^3\frac{\zeta^3}{\sigma_n^{3/2}} + \|\tilde{q}\|\frac{\zeta}{\sqrt{\sigma_n}}\right) f(\tilde{q})\,d\tilde{q} \\
&= \zeta\sigma_n^{3/2}\mathbb{E}\left[\dot{m}_4(\tilde{Q})f_s(0\mid\tilde{Q})\|\tilde{Q}\|\right] + \zeta^3\sqrt{\sigma_n}\mathbb{E}\left[\dot{m}_4(\tilde{Q})f_s(0\mid\tilde{Q})\|\tilde{Q}\|^3\right]
\end{aligned}
$$

The third residual R_3 is even higher order term and hence skipped. It is immediate that the order of the remainders are equal to or smaller than $\zeta\sqrt{\sigma_n}$ which

implies:

$$\mathbb{E}[m_4(Q)T_1] \lesssim \zeta\sqrt{\sigma_n}.$$

The calculation for T_2 is similar and hence skipped for brevity. Combining conclusions for T_1 and T_2 we conclude when $\zeta \leq \sqrt{\mathcal{K}\sigma_n}$:

$$\begin{aligned}
&\mathbb{E}\left[F^2_{3,\zeta}(X,Y,Q)\right] \\
&= \mathbb{E}\Bigg[\left|\left[-2\left(Y - X^\top\beta_0^s\right)X^\top\delta_0^s + (X^\top\delta_0^s)^2\right]\right|^2 \\
&\qquad \times \sup_{d_*(\theta,\theta_0^s)\leq\zeta}\left|\left\{K\left(\frac{Q^\top\psi_0^s}{\sigma_n}\right) - K\left(\frac{Q^\top\psi}{\sigma_n}\right)\right\}\right|^2\Bigg] \\
&\lesssim \mathbb{E}\left[m_4(Q)\sup_{d_*(\theta,\theta_0^s)\leq\zeta}\left|\left\{K\left(\frac{Q^\top\psi_0^s}{\sigma_n}\right) - K\left(\frac{Q^\top\psi}{\sigma_n}\right)\right\}\right|^2\right] \\
&\lesssim \zeta\sqrt{\sigma_n}. \qquad \text{(A.7)}
\end{aligned}$$

Case 2: Now consider $\zeta > \sqrt{\mathcal{K}\sigma_n}$. Then it is immediate that:

$$\sup_{d_*^2(\psi,\psi_0^s)\leq\zeta^2}\|\psi - \psi_0^s\| = \zeta^2.$$

Using this we have:

$$\begin{aligned}
&\mathbb{E}[m_4(Q)T_1] \\
&= \mathbb{E}\left[m_4(Q)\left|\left\{K\left(\frac{Q^\top\psi_0^s}{\sigma_n}\right) - K\left(\frac{Q^\top\psi_0^s}{\sigma_n} + \|\tilde{Q}\|\frac{\zeta^2}{\sqrt{\sigma_n}}\right)\right\}\right|^2\right] \\
&= \sigma_n\int_{\mathbb{R}^{p-1}}\int_{-\infty}^{\infty} m_4(\sigma_n t - \tilde{q}^\top\tilde{\psi}_0^s,\tilde{q})\left|K(t) - K\left(t + \|\tilde{q}\|\frac{\zeta^2}{\sigma_n}\right)\right|^2 \\
&\qquad \times f_s(\sigma_n t \mid \tilde{q})\,dt\; f(\tilde{q})\,d\tilde{q} \\
&\leq \sigma_n\int_{\mathbb{R}^{p-1}}\int_{-\infty}^{\infty} m_4(\sigma_n t - \tilde{q}^\top\tilde{\psi}_0^s,\tilde{q})\left|K(t) - K\left(t + \|\tilde{q}\|\frac{\zeta^2}{\sigma_n}\right)\right| \\
&\qquad \times f_s(\sigma_n t \mid \tilde{q})\,dt\; f(\tilde{q})\,d\tilde{q} \\
&\leq \sigma_n\int_{\mathbb{R}^{p-1}}\int_{-\infty}^{\infty} m_4(\sigma_n t - \tilde{q}^\top\tilde{\psi}_0^s,\tilde{q})\|\tilde{q}\|\frac{\zeta^2}{\sigma_n} \\
&\qquad \times f_s(\sigma_n t \mid \tilde{q})\,dt\; f(\tilde{q})\,d\tilde{q} \\
&= \zeta^2\int_{\mathbb{R}^{p-1}} m_4(-\tilde{q}^\top\tilde{\psi}_0^s,\tilde{q})f_s(0 \mid \tilde{q})\|\tilde{q}\|\; f(\tilde{q})\,d\tilde{q} + R \\
&\leq \zeta^2\mathbb{E}\left[\|\tilde{Q}\|m_4\left(-\tilde{Q}^\top\tilde{\psi}_0^s,\tilde{Q}\right)f_s(0 \mid \tilde{Q})\right] + R
\end{aligned}$$

The analysis of the remainder term is similar and if is of higher order. This concludes when $\zeta > \sqrt{K\sigma_n}$:

$$\mathbb{E}\left[F^2_{3,\zeta}(X,Y,Q)\right]$$

$$\begin{aligned}
&\mathbb{E}\left[\left|\left[-2\left(Y-X^\top\beta_0^s\right)X^\top\delta_0^s+(X^\top\delta_0^s)^2\right]\right|^2\right.\\
&\qquad\left.\times\sup_{d_*(\theta,\theta_0^s)\le\zeta}\left|\left\{K\left(\frac{Q^\top\psi_0^s}{\sigma_n}\right)-K\left(\frac{Q^\top\psi}{\sigma_n}\right)\right\}\right|^2\right]\\
&\lesssim\mathbb{E}\left[m_4(Q)\sup_{d_*(\theta,\theta_0^s)\le\zeta}\left|\left\{K\left(\frac{Q^\top\psi_0^s}{\sigma_n}\right)-K\left(\frac{Q^\top\psi}{\sigma_n}\right)\right\}\right|^2\right]\\
&\lesssim\zeta^2. \qquad\text{(A.8)}
\end{aligned}$$

Combining (A.7), (A.8) with equation (A.4) we have:

$$\begin{aligned}
\sqrt{n}\mathbb{E}\left[\sup_{\theta:d_*(\theta,\theta_0^s)\le\zeta}\left|(\mathbb{P}_n-P)\left(f_\theta-f_{\theta_0^s}\right)\right|\right]&\lesssim\sqrt{\zeta}\sigma_n^{1/4}\mathbb{1}_{\zeta\le\sqrt{\mathcal{K}\sigma_n}}+\zeta\mathbb{1}_{\zeta>\sqrt{\mathcal{K}\sigma_n}}\\
&:=\phi_n(\zeta).
\end{aligned}$$

Hence to obtain rate we have to solve $r_n^2\phi_n(1/r_n)\le\sqrt{n}$, i.e. (ignoring $\mathcal{K}$ as this does not affect the rate)

$$r_n^{3/2}\sigma_n^{1/4}\mathbb{1}_{r_n\ge\sigma_n^{-1/2}}+r_n\mathbb{1}_{r_n\le\sigma_n^{-1/2}}\le\sqrt{n}.$$

Now if $r_n\le\sigma_n^{-1/2}$ then $r_n=\sqrt{n}$ which implies $\sqrt{n}\le\sigma_n^{-1/2}$ i.e. $n\sigma_n\to0$ and hence contradiction. On the other hand, if $r_n\ge\sigma_n^{-1/2}$ then $r_n=n^{1/3}\sigma_n^{-1/6}$. This implies $n^{1/3}\sigma_n^{-1/6}\ge\sigma_n^{-1/2}$, i.e. $n^{1/3}\ge\sigma_n^{-1/3}$, i.e. $n\sigma_n\to\infty$ which is okay. This implies:

$$n^{2/3}\sigma_n^{-1/3}d^2(\hat\theta^s,\theta_0^s)=O_p(1).$$

Now as $n^{2/3}\sigma_n^{-1/3}\gg\sigma_n^{-1}$, we have:

$$\frac{1}{\sigma_n}d^2(\hat\theta^s,\theta_0^s)=o_p(1),$$

which further indicates $\|\hat\psi^s-\psi_0^s\|/\sigma_n=o_p(1)$. This, along with the fact that $\|\psi_0^s-\psi_0\|/\sigma_n=o(1)$ (from Lemma 2.7), establishes that $\|\hat\psi_0^s-\psi_0\|/\sigma_n=o_p(1)$. This completes the proof. □

Supplementary Material

Supplement of "Asymptotic normality of a linear threshold estimator in fixed dimension with near-optimal rate"
(doi: 10.1214/23-EJS2144SUPP; .pdf).

References

[1] Joseph E Aldy. An environmental Kuznets curve analysis of us state-level carbon dioxide emissions. *The Journal of Environment & Development*, 14(1):48–72, 2005.

[2] Theophile Azomahou, François Laisney, and Phu Nguyen Van. Economic development and co2 emissions: A nonparametric panel approach. *Journal of Public Economics*, 90(6-7):1347–1363, 2006.
[3] Luisito Bertinelli and Eric Strobl. The environmental Kuznets curve semi-parametrically revisited. *Economics Letters*, 88(3):350–357, 2005. MR2155494
[4] Bilal Boubellouta and Sigrid Kusch-Brandt. Cross-country evidence on environmental Kuznets curve in waste electrical and electronic equipment for 174 countries. *Sustainable Production and Consumption*, 25:136–151, 2021.
[5] Elizabeth Burton. The compact city: just or just compact? A preliminary analysis. *Urban studies*, 37(11):1969–2006, 2000.
[6] Roberta Capello and Roberto Camagni. Beyond optimal city size: an evaluation of alternative urban growth patterns. *Urban Studies*, 37(9):1479–1496, 2000.
[7] Limin Du, Chu Wei, and Shenghua Cai. Economic development and carbon dioxide emissions in china: Provincial panel data analysis. *China Economic Review*, 23(2):371–384, 2012.
[8] Ailin Fan, Rui Song, and Wenbin Lu. Change-plane analysis for subgroup detection and sample size calculation. *Journal of the American Statistical Association*, 112(518):769–778, 2017. MR3671769
[9] Gene M Grossman and Alan B Krueger. Economic growth and the environment. *The Quarterly Journal of Economics*, 110(2):353–377, 1995.
[10] Bruce E Hansen. Sample splitting and threshold estimation. *Econometrica*, 68(3):575–603, 2000. MR1769379
[11] Bruce E Hansen. Threshold autoregression in economics. *Statistics and its Interface*, 4(2):123–127, 2011. MR2812805
[12] Joel L Horowitz. A smoothed maximum score estimator for the binary response model. *Econometrica: Journal of the Econometric Society*, pages 505–531, 1992. MR1162997
[13] Kosuke Imai, Marc Ratkovic, et al. Estimating treatment effect heterogeneity in randomized program evaluation. *The Annals of Applied Statistics*, 7(1):443–470, 2013. MR3086426
[14] Chae Ryon Kang. New statistical learning methods for chemical toxicity data analysis, 2011.
[15] Yan Lan, Moulinath Banerjee, George Michailidis, et al. Change-point estimation under adaptive sampling. *The Annals of Statistics*, 37(4):1752–1791, 2009. MR2533471
[16] Sokbae Lee, Yuan Liao, Myung Hwan Seo, and Youngki Shin. Factor-driven two-regime regression. *The Annals of Statistics*, 49(3):1656–1678, 2021. MR4298876
[17] Yoonseok Lee, Debasri Mukherjee, and Aman Ullah. Nonparametric estimation of the marginal effect in fixed-effect panel data models. *Journal of Multivariate Analysis*, 171:53–67, 2019. MR3886393
[18] Jialiang Li, Yaguang Li, Baisuo Jin, and Michael R Kosorok. Multithreshold change plane model: Estimation theory and applications in subgroup identification. *Statistics in Medicine*, 40(15):3440–3459, 2021. MR4269063

[19] Wei Liang and Ming Yang. Urbanization, economic growth and environmental pollution: Evidence from China. *Sustainable Computing: Informatics and Systems*, 21:1–9, 2019.

[20] Brant Liddle and Sidney Lung. Age-structure, urbanization, and climate change in developed countries: revisiting stirpat for disaggregated population and consumption-related environmental impacts. *Population and Environment*, 31(5):317–343, 2010.

[21] John A List and Craig A Gallet. The environmental Kuznets curve: does one size fit all? *Ecological Economics*, 31(3):409–423, 1999.

[22] Daniel L Millimet, John A List, and Thanasis Stengos. The environmental Kuznets curve: real progress or misspecified models? *Review of Economics and Statistics*, 85(4):1038–1047, 2003.

[23] Debarghya Mukherjee, Moulinath Banerjee, and Ya'acov Ritov. On robust learning in the canonical change point problem under heavy tailed errors in finite and growing dimensions. *Electronic Journal of Statistics*, 16(1):1153–1252, 2022. MR4381059

[24] Perry Sadorsky. The effect of urbanization on co2 emissions in emerging economies. *Energy Economics*, 41:147–153, 2014.

[25] Myung Hwan Seo and Oliver Linton. A smoothed least squares estimator for threshold regression models. *Journal of Econometrics*, 141(2):704–735, 2007. MR2413485

[26] Nemat Shafik and Sushenjit Bandyopadhyay. *Economic Growth and Environmental Quality: Time-Series and Cross-Country Evidence*, volume 904. World Bank Publications, 1992.

[27] Susan Sunila Sharma. Determinants of carbon dioxide emissions: empirical evidence from 69 countries. *Applied Energy*, 88(1):376–382, 2011.

[28] Fatma Taskin and Osman Zaim. Searching for a Kuznets curve in environmental efficiency using kernel estimation. *Economics Letters*, 68(2):217–223, 2000.

[29] Aad W Van Der Vaart and Jon A Wellner. Weak convergence. In *Weak Convergence and Empirical Processes*, pages 16–28. Springer, 1996. MR1385671

[30] Susan Wei and Michael R Kosorok. Latent supervised learning for estimating treatment effect heterogeneity, 2014. MR3174676