NBER WORKING PAPER SERIES

ON TESTING CONTINUITY AND THE DETECTION OF FAILURES

Matthew Backus
Sida Peng

Working Paper 26016
http://www.nber.org/papers/w26016

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
June 2019

We are grateful to Timothy Armstrong, Simon Lee, Francesca Molinari, Serena Ng, and many conference and seminar participants for thoughtful comments. All remaining errors are our own. The views expressed herein are those of the authors and do not necessarily reflect the views of the National Bureau of Economic Research.

On Testing Continuity and the Detection of Failures
Matthew Backus and Sida Peng
NBER Working Paper No. 26016
June 2019
JEL No. C01,C20,C52

ABSTRACT

Estimation of discontinuities is pervasive in applied economics: from the study of sheepskin effects to prospect theory and "bunching" of reported income on tax returns, models that predict discontinuities in outcomes are uniquely attractive for empirical testing. However, existing empirical methods often rely on assumptions about the number of discontinuities, the type, the location, or the underlying functional form of the model. We develop a nonparametric approach to the study of arbitrary discontinuities —point discontinuities as well as jump discontinuities in the nth derivative, where n = 0,1,... — that does not require such assumptions. Our approach exploits the development of false discovery rate control methods for lasso regression as proposed by G'Sell et al. (2015). This framework affords us the ability to construct valid tests for both the null of continuity as well as the significance of any particular discontinuity without the computation of nonstandard distributions. We illustrate the method with a series of Monte Carlo examples and by replicating prior work detecting and measuring discontinuities, in particular Lee (2008), Card et al. (2008), Reinhart and Rogoff (2010), and Backus et al. (2018b).

Matthew Backus
Graduate School of Business
Columbia University
3022 Broadway, Uris Hall 619
New York, NY 10027
and NBER
matthew.backus@columbia.edu

Sida Peng
Microsoft Research
sp947@cornell.edu

1 Introduction

This paper introduces a method for detecting discontinuities for when the econometrician knows little about the underlying parametric form, the number of discontinuities, the location of discontinuities, or their type (point, jump, kink, etc). Detection of these discontinuities is sometimes of direct economic interest: e.g., the study of tipping points in residential segregation (Card et al., 2008) or cheap-talk signaling conventions (Backus et al., 2018b). Moreover, as a specification test it is a step in the direction of "Sherlock Holmes" inference (Leamer, 1983): by allowing the researcher to be agnostic, it leaves open the possibility that they detect anomalies. In particular, it is useful as a placebo test for regression discontinuity and regression kink designs.

Our agnosticism implies that we are testing against a large set of alternative hypotheses – many possible breaks – and so our procedure is adaptive. We are openly engaging in "data snooping," i.e. model selection, and it is well-known that this is problematic for post-estimation inference. Inference after model selection, respecting these limitations, is an area of recent interest. Our solution eludes these problems (or, alternatively, the need for sample-splitting) by building our chosen inferential guarantee *into the model selection procedure itself*, so that there is no "post" selection. To be precise, we develop an estimator that detects discontinuities while controlling the False Discovery Rate (FDR) – i.e. the ratio of type I errors to the total number of rejections – at a pre-specified level (Benjamini and Hochberg, 1995).[1]

The model selection component of our algorithm uses a lasso framework to construct an ordered sequence of hypothesis tests, i.e. potential discontinuities. We construct conditionally valid test statistics for each – that is, we explicitly condition on the event the hypothesis test was selected. Fithian et al. (2015) demonstrates that this conditioning is closely related to – and more powerful than – sample splitting. Because each hypothesis test is conditional on the last, to control the FDR we must then solve a sequential multiple comparison problem (MCP). To this end we take advantage of the Forward Stop algorithm of G'Sell et al. (2015) which extends FDR control from the simultaneous MCP setting of Benjamini and Hochberg (1995) to the sequential MCP setting. In simulations we are able to show that this sequential approach is more powerful than using standard FDR control with confidence intervals proposed by other recent work (van de Geer et al., 2014; Belloni et al., 2014). Intuitively, this is because that approach ignores correlation between the

[1] FDR control has not seen wider adoption as an inferential standard in economics because it raises some philosophic questions around the interpretation of covariates, which is always conditional on controls. For instance, the effect of education on wages is different from the effect of education on wages conditional on a measure of ability. It is unclear how to interpret the conditional effect when the conditioning argument, here ability, is a false detection with some pre-specified likelihood. These issues will not arise in our context: the detection, false or otherwise, of one discontinuity, is irrelevant to the interpretation of another located elsewhere.

test statistics, which ours exploits via the nested structure of the hypothesis test. We somewhat incidentally contribute, therefore, to an emerging literature on using that correlation in FDR control (see Fan et al., 2012; Fan and Han, 2016; Basu et al., 2017).

We also draw on a large literature on the detection of structural breaks. There are a number of salient technical challenges that this literature has confronted, beginning with the multiple comparison problem (MCP) implicit in testing at many potential break locations. One solution, to formalize the MCP as an order statistic problem, was proposed by Hawkins (1987) and generalized by Andrews (1993). A second thread in this literature has focused on the construction of confidence intervals for the parameter governing the size of a break at the most likely location (Hansen, 2000, 2017). This is a nonstandard inference problem because nesting the null of continuity implies a discontinuity in the parameter set, and these papers develop methods for simulating from the distribution of the test statistic to obtain critical values. This literature suffers from the same failure of uniform validity that plagues all "data-snooping" or adaptive estimators (Leeb and Pötscher, 2005). Applied work has therefore relied on sample splitting, with the attendant loss of power.

With respect to this literature we make several applied contributions: first, we are able to simultaneously test for n-th order discontinuities (points, jumps, kinks, etc). We also allow for multiple unknown discontinuities. In this spirit the estimator is related to work on detection of multiple structural breaks. We compare our procedure to existing procedures for looking for multiple jumps (sequential sample splitting as in Bai (1997a), the sequential sup-F test of Bai and Perron (1998), and application of the BIC criterion as in Yao (1988)). Prior work has focused on consistency results without explicitly addressing the implicit multiple comparison problem, instead assuming that $\alpha \to 0$ asymptotically. This leaves little guidance for the applied researcher, with finite data, in choosing the size of the test. We highlight this point with a Monte Carlo simulation that shows the attendant risk of false discoveries when applying these methods in finite sample (i.e., real) applications. Moreover, as we show in Section 4.4, even when the parametric form of the continuous part of the relationship is known and employed by the econometrician, sequential application of existing methods will often fail in the presence of multiple breaks. We emphasize this to highlight the significance of being nonparametric in the treatment of that form: even if one knows the true functional form of the continuous part, in searching for the first break the model is interim-misspecified, which can lead to erroneous results.

Finally and most importantly for applied work, our method is tractable and transparent, which we illustrate in applications to Lee (2008), Card et al. (2008), Reinhart and Rogoff (2010), and Backus et al. (2018b). Moreover, code for the estimator is publicly available.[2]

[2] As of this writing, the most recent code can be found at https://github.com/psdsam/lasso_break.

These applied contributions rest on several technical innovations that allow us to extend innovations in FDR control lasso, due to G'Sell et al. (2015), to our proposed method. First, we show the conditions under which the set of potential discontinuities, formulated as a design matrix, satisfy the irrepresentable condition (Theorem 1). This is key for our asymptotic results. Then, we extend results in Lockhart et al. (2014) to show that their proposed covariance test can accommodate the interim misspecification error that emerges in our setting (Theorem 2). Finally, we close the argument by demonstrating the asymptotic consistency of the proposed estimator (Theorem 3). In a slight detour, we prove two additional results for the discontinuities located by our estimator that allow us to make a direct comparison to prior work. These results, like that prior work, depend on a substantially more restrictive environment. There we show that the location of a single jump (Theorem 4) and a single kink (Theorem 5) enters the lasso active set is the max of two Wiener processes. In the jump case this is similar to that for the standard sup-F test derived in Bai (1997b). However we also derive the asymptotic distribution for the kink discontinuity case as $\beta \to 0$ sufficiently slowly, which is stronger than Hansen (2017). As in prior work, these two results assume that the continuous part of the function is a constant zero.

2 Model

Our model decomposes the relationship between two variables, y and x, into a continuous part and a finite set of discontinuities. Let x be drawn from a continuous distribution (up to a set of mass points) with support $[\underline{\omega}, \overline{\omega}]$, a compact subset of $\mathbb{R}$, and

$$y = g(x) + \sum_{s=1\ldots S} \sum_{k=0,\ldots,K} d_{sk}(x)\psi_{sk} + \epsilon. \tag{1}$$

In this setting $g(x)$ is bounded, continuous, and differentiable up to order K, while $\{d_{sk}(x)\}$ is a set of violations of continuity or differentiability and ϵ is an i.i.d. error term.[3] With respect to notation, s indexes the location of the violation and k indexes the degree. In particular,

$$d_{sk}(x) = \begin{cases} \mathbb{1}(x = z_s) & \text{(point discontinuity)} \\ \mathbb{1}(x \geq z_s) & \text{(jump discontinuity)} \\ \mathbb{1}(x \geq z_s)(x - z_s) & \text{(discontinuous first derivative)} \\ \mathbb{1}(x \geq z_s)(x - z_s)^2 & \text{(discontinuous second derivative)} \\ \quad \vdots & \quad \vdots \end{cases} \tag{2}$$

[3] We suppress here and where possible the index of the observations.

In this schema, z_s denotes the location of a break point and ψ_s its magnitude. So far, we assume no knowledge on the magnitude of breaks, the number of breaks or the type of breaks.

In empirical applications these discontinuities often have particular economic meaning:

Example 1: Jump discontinuities in Lee (2008). In this classic paper, x is the Democrat vote share in the prior election, normalized to a margin of victory (or loss, for $x < 0$); y is the Democrat vote share in the current election, and the paper is interested measuring in a single jump discontinuity at $z = 0$ which is taken to represent the causal effect of incumbency on subsequent outcomes.

Example 2: Jump discontinuities in Card et al. (2008). This paper studies discontinuities in the dynamics of neighborhood racial composition in order to study Schelling-style tipping models of segregation. Here x is the current fraction of non-white or hispanic residents, y is the rate of change of white residents, and the location and magnitude of a *single* jump discontinuity, which is assumed to exist, are unknown.

Example 3: Kink discontinuities in Reinhart and Rogoff (2010) and Hansen (2017). The former paper proposes proposes that economic growth is discontinuously lower when the ratio of debt to output exceeds a particular threshold, modeled as a single kink discontinuity with an unknown location and magnitude. Hansen (2017) develops a new method for finding a single kink discontinuity at an unknown location and, in an application to the prior paper's data, finds evidence of a growth slowdown at a debt to GDP ratio of 44%.

Example 4: Point discontinuities in Backus et al. (2018b). This paper studies bargaining on an online platform. They take y to be an expected bargaining outcome (e.g., negotiated price or average first buyer offer) and x the original asking price of the seller (who always moves first). The use of round number values in the asking price (e.g., 100, 200, ...) is interpreted as a cheap-talk signal of the seller's bargaining type, which elicits discontinuously different behavior by prospective buyers, represented by point discontinuities in $\mathbb{E}[y|x]$. In an extension that prompted the present investigation, they take the set of discontinuities to be unknown and use lasso regression to detect them (see their Appendix B4).

These examples highlight the wide range of interpretations of discontinuities in economic modeling. We replicate each of them in Section 6 to illustrate the application of our estimator, to compare results to those obtained using pre-existing methods, and to highlight the value of the added generality. The procedure is agnostic as to the meaning of any discontinuities – therefore it is applicable in each of these examples but it does not, on its own, imply any structural or causal interpretation.

3 Assumptions and Setup

Here we develop the assumptions and lasso framework of our estimator. Section 3.1 introduces the central assumptions on the data-generating process. Next, Section 3.2 introduces the our "semi-lasso" specification and presents a preliminary result that maps it back into a standard lasso regression. Finally, Section 3.3 outlines restrictions on the design matrix – which summarizes the possible set of discontinuities – that guarantee the irrepresentable condition, the key condition that affords both the distribution of the test statistic we employ as well as the desirable asymptotic properties of the lasso.

3.1 Assumptions

For a given finite sample $\{(x_i, y_i)\}_{i=1}^n$, let $(x_{(1)}, x_{(2)}, \cdots, x_{(n)})$ be order statistics. Note that point discontinuities are only identified at mass points of the distribution. Moreover, detection in the generic class of possible discontinuities is also problematic with finite data. For example, consider a jump discontinuity of known size β, and two possible locations, z and z'. If, for some i, $x_{(i-1)} < z < z' < x_{(i)}$, then the two discontinuities are equipollent, and therefore empirically indistinguishable. Therefore, for such a sample, the maximal meaningful set of jump discontinuities is $n-1$, and will be strictly larger than n, e.g. if the econometrician is interested in multiple types (point, jump, kink, etc.). We cannot, therefore, simply add those discontinuities as regressors and obtain consistent estimates of their size and location using OLS. Instead, we adopt a model selection approach, and to that end we require four strong assumptions:

Assumption 1. *(Number of Discontinuities) The number of discontinuities is finite.*

This assumption is reasonable for most applied work given the bounded support of x. Sparseness is typically to be motivated by the economic intuition for the existence of the breaks in the first place, e.g. signaling conventions or institutional rules. However, when the economist is agnostic as the to existence or character of the discontinuities, for instance, when the method is used as a generalized placebo test to motivate a regression discontinuity design, this assumption can be strong.[4]

Let $S_m(x)$ denote the spline or power series, where m indexes the order (we omit this index in all but the following assumption), so that $\hat{g}_m(x) \equiv S_m(x)'\hat{\beta}$ and the approximation error can be written $r_{mi} \equiv g(x_i) - \hat{g}_m(x_i)$. Define $S_{mi} \equiv S_m(x_i)$, $S_m \equiv [S_{m1}, S_{m2}, \cdots, S_{mn}]'$, $\varphi_m^2 \equiv \mathbb{E}r_{mi}^2$, $Q_m \equiv \mathbb{E}(S_{mi}S_{mi}')$ and $h_{mi} \equiv S_{mi}'(\sum_{i=1}^n S_{mi}S_{mi}')^{-1}S_{mi}$.

[4] We do not allow the number of discontinuities to grow as a function of n to rule out cases with a sequence of discontinuities that are placed arbitrarily closely as n goes to infinity.

Assumption 2. *(Eigenvalue Restriction) Let K_m be the number of terms in power series or spline expansion. For every K_m, there is a nonsingular constant K_m by K_m matrix B such that smallest eigenvalue $\Lambda^2_{f,min}$ of $\mathbb{E}(BS_m(x)S_m'(x)B')$ is bounded away from 0 uniformly in K_m and there is a sequence of constant $\xi_0(K_m)$ satisfying $\sup_z \|BS_m(x)\| \leq \xi_0(K_m)$ and $K_m = K_m(n)$ such that $\xi_0(K_m)^2 K_m/n \to 0$ as $n \to 0$*

Assumption 2 is the standard Newey (1997) assumption for series estimators. The bounds on the smallest and largest eigenvalues will allow us to look for discontinuities in the null space of the basis functions.

Assumption 3. *(Variance) The variance $\sigma^2 \equiv \mathbb{E}(\epsilon_i^2)$ is finite and bounded away from 0.*

Assumption 3 is standard in the literature, see Newey (1997) and Hansen (2014b). We require the variance to be bounded away from 0 so the test statistic is well defined.

We also assume $g(x)$ satisfies smoothness requirements that allow it to be approximated nonparametrically with splines or a power series. In order to derive the rate of convergence for the integrated mean squared error (IMSE), other technical conditions are required. Assumption 4 is equivalent to those required in Donald and Newey (1994) and Hansen (2014b).

Assumption 4. *(Smoothness)*

Let $S_m(x)$ be either a spline or power series, and nested.

1. *$g(x)$ has s times continuous derivatives on X, with $s > K$, where K is the maximum order of discontinuities to test.*
2. *Recall K_m is the number of terms in power series or spline expansion and define M_n as the maximum order of the spline or power series considered. Then let $\max_{m<M_n} K_m^4/n = O(1)$ for a power series or $\max_{m<M_n} K_m^3/n = O(1)$ for splines sieve.*
3. *$\varphi_m^2 > 0$ for all $m < \infty$.*

Standard nonparametric approximation only requires $g(x)$ to be continuously differentiable while Assumption 4.1 is stronger and assumes the degree of continuity of $g(x)$ to be greater than the degree of discontinuity we are testing. Assumption 4.2 prevents overfitting of the model. Assumption 4.3 bounds away the approximation error from 0.

3.2 Semi-Lasso and Lasso Regression

Our analysis is general to discontinuities of arbitrary order, but for the sake of exposition we will focus on points, jumps, and kinks. The design matrix $D(x_i)$ is comprised of a series of vectors characterizing discontinuities under consideration, e.g.:

$$\left(1_{x_i=x_{(1)}}, 1_{x_i=x_{(2)}}, \cdots, 1_{x_i=x_{(n)}}\right); \qquad \text{(point discontinuity)}$$

$$\left(\frac{1_{x_i>x_{(1)}}}{\sqrt{n-1}}, \frac{1_{x_i>x_{(2)}}}{\sqrt{n-2}}, \cdots, \frac{1_{x_i>x_{(n-1)}}}{1}\right), \text{ and} \qquad \text{(jump discontinuity)}$$

$$\left(\frac{(x_i - x_{(1)}) \times 1_{x_i>x_{(1)}}}{\phi_1}, \frac{(x_i - x_{(2)}) \times 1_{x_i>x_{(2)}}}{\phi_2}, \cdots, \frac{(x_i - x_{(n)}) \times 1_{x_i>x_{(n-1)}}}{\phi_{n-1}}\right) \quad \text{(kink discontinuity)}$$

– where $\phi_k = \sqrt{\sum_{i=k+1}^{n}(x_{(i)} - x_{(k)})^2}$. We normalize all the regressors to have norm 1 so they are balanced when penalized by lasso.

The matrix D may be comprised of multiple types of discontinuities. For example, when the econometrician wishes to test for both jumps and kinks,

$$D(x_i) = \left(\frac{1_{x_i>x_{(1)}}}{\sqrt{n-1}}, \frac{1_{x_i>x_{(2)}}}{\sqrt{n-2}}, \cdots, \frac{1_{x_i>x_{(n-1)}}}{1}, \frac{(x_i - x_{(1)}) \times 1_{x_i>x_{(1)}}}{\phi_1}\right.$$

$$\left.\frac{(x_i - x_{(2)}) \times 1_{x_i>x_{(2)}}}{\phi_2}, \cdots, \frac{(x_i - x_{(n)}) \times 1_{x_i>x_{(n-1)}}}{\phi_{n-1}}\right).$$

Our procedure concerns itself with the following "semi-lasso" regression which, unlike a traditional lasso, penalizes only the vector ψ and not β:

$$(\hat{\beta}, \hat{\psi}) = \arg\min_{\beta,\psi} \sum_n (y_i - S(x_i)\beta - D(x_i)\psi)^2 + \lambda|\psi|_1 \tag{3}$$

– where, recall, $S(x_i)$ is a vector of linear basis functions for the space of continuous functions on $[\underline{\omega}, \overline{\omega}]$, e.g. basis splines, $D(x_i)$ is the set of potential discontinuities of interest, and λ is the lasso penalty parameter. Note that this lasso specification is unique in that the continuous part β is entirely unpenalized – this is why we say our approach maintains the null of continuity.

Finally, we show that minimization problem above is equivalent to a standard lasso problem in the

null space of the basis functions.

Lemma 1. *[Invariance of LASSO under Projection] The problem of (3) is equivalent to the following standard lasso problem:*

$$(\hat{\psi}) = \arg\min_{\psi} \frac{1}{2}\|M_m Y - M_m D\psi\|_2^2 + \lambda|\psi|_1 \tag{4}$$

– where M_m is the projection matrix $I_{n\times n} - S_m(S_m'S_m)^{-1}S_m'$.

The proof, as for all that follows, is available in Appendix A. Having transformed our problem back into the standard lasso framework, we next turn to the condition that guarantees that our lasso is well-behaved.

3.3 Irrepresentable Condition

Write $D = (D(x_1)', D(x_2))', \cdots, D(x_n)')'$, and denote D_j the jth column of D. Denote ψ_i as the ith entry of vector ψ. Let A_0 be the index set of the true discontinuities and D_{A_0} represents a sub-matrix of D with columns corresponding to those indexed in A_0. Let $\tau_{A_0} = sign(\psi_{A_0})$.

Definition 1 (Irrepresentable Condition)**.** *We say that the irrepresentable condition holds for $\eta < 1$, if*

$$\max_{j\notin A_0} \sup_{\|\tau_{A_0}\|_\infty \le 1} \left| D_j' D_{A_0} (D_{A_0}' D_{A_0})^{-1} \tau_{A_0} \right| < \eta.$$

In a jump discontinuity design matrix, D is a lower triangular matrix after rearranging the rows. For example, the kth column is:

$$D_k = \left(0, 0, 0, \cdots, \frac{1}{\sqrt{n-k}}, \frac{1}{\sqrt{n-k}}, \cdots, \frac{1}{\sqrt{n-k}}\right)'.$$

Assume $A_0 = \{k\}$, that is we only have 1 single jump discontinuity. Thus $D_{A_0}' D_{A_0} = 1$ and for all $j \neq k$,

$$D_j' D_{A_0} = \frac{\min\{(n-k), (n-j)\}}{\sqrt{(n-k)(n-j)}} < 1.$$

For asymptotic, we assume the location of the discontinuities represented by the ordered statistics are fixed when n goes to infinity. Thus the irrepresentable condition holds. We show in Theorem 1 below for the general cases:

Theorem 1. *[Design Matrices for Detecting Discontinuities]*

Let p be the number of columns in the design matrix D [5].
(a) For any set $A_0 \subset \{1, 2, \cdots, p\}$, the design matrix for point discontinuities satisfies the irrepresentable condition.
(b) For any set $A_0 \subset \{1, 2, \cdots, p\}$, the design matrix for jump discontinuities satisfies the irrepresentable condition.
(c) For any set $A_0 \in \{1, 2, \cdots, p\}$ as a singleton, the design matrix for any combinations of point and 0 to Kth order discontinuities satisfies the irrepresentable condition.

Corollary 1 (Invariant of Irrepresentable Condition under Projection)**.** *Let P_z be a projection matrix such that $P_z D$ has full column rank. If the design matrix D satisfies the irrepresentable condition, then $P_z D$ also satisfies the irrepresentable condition.*

While these results are encouraging, it is also possible to demonstrate the the irrepresentable condition will not hold for many cases of interest.

Corollary 2 (Kink Violation)**.** *The irrepresentable condition does not hold when there are two or more kinks.*

For the case where the econometrician wishes to allow for multiple kinks in arbitrary locations, the irrepresentable condition does not hold. Recall that the irrepresentable condition requires that there is no regressor in the complement of the active set that approximates a combination of regressors in the active set "too well" – i.e., such that with increasing data and a sharper regularization parameter, they still cannot be distinguished. Our negative result on this point hinges on the fact that for too arbitrarily close kinks, the error induced by combining them into one is second-order, and therefore lasso cannot be guaranteed to differentiate them.

For such cases, stronger assumptions are required to guarantee consistency. Corollary 3 accomplishes this by assuming that there exits a partition on the support of X such that each segment contains at most one kink discontinuity. This is sufficient to guarantee the irrepresentable condition.

Corollary 3 (Irrepresentable Condition under Partition)**.** *Let $\{B_1, B_2, \cdots, B_s\}$ be a partition of the support of X. Define a block diagonal matrix $\Pi = diag[\Pi_1, \Pi_2, \cdots, \Pi_s]$, where Π_i is the design matrix in partition i. If the irrepresentable condition is satisfied in each Π_i, the design matrix Π also satisfies the irrepresentable condition.*

[5] Notice that $p = n$ in the jump discontinuity design matrix, but it can be different from n when we consider multiple types of discontinuities or if we want to consider discontinuities on a subset of the support.

Though it imposes a stronger restriction, Corollary 3 has a natural applied interpretation: our method can be applied on each segment of the partition, which can be accomplished with appropriately flexible splines.

4 Detecting Discontinuities

Our procedure considers a nested sequence of models: starting from a null of continuity, each step in the sequence adds one additional discontinuity. The sequence of models is determined by knots of the coefficient path following the LARS algorithm of Efron et al. (2004).[6] For each step in the sequence, we construct a test statistic. Conditional on a model, the "covariance test" is an interim hypothesis test of the null hypothesis that the true model is contained in the current model.

This generates a sequence of interim p values associated with each marginal discontinuity, a sequence which can be used to control the false discovery rate of the procedure. We apply the sequential MCP approach of G'Sell et al. (2015) to select a critical threshold that guarantees the false discovery rate. In this section we explain each of these steps in detail, and conclude with a brief discussion of of the advantages of this approach *vis-à-vis* standard mean-squared error-based approaches.

4.1 Nested Hypotheses

Recall that from Lemma 1 our problem has been transformed into a standard lasso problem. Therefore, the model selection is determined by the choice of the lasso penalty term λ. The ordinary approach would be to choose λ by either cross-validation or by using an estimate of σ_ϵ to approximate the rate-optimal λ. While convenient, these leave the researcher neither an inferential framework nor control over the Type I error rate – in our setting, the likelihood of falsely detecting discontinuities. Instead, we employ the FDR lasso framework of G'Sell et al. (2015) in order to control the probability of false inclusion of variables in the model. In particular, we use their *forward stop* algorithm, which proceeds path-wise along a sequence of covariance test statistics that offer interim significance tests for the marginal included variable (Lockhart et al., 2014). The solution $\psi(\lambda)$ of (4) is a continuous and piecewise linear function and can be computed via the LARS algorithm as Efron et al. (2004).

Define $\lambda_1 \geq \lambda_2 \geq \cdots$ as the critical values of the penalty parameter λ, i.e. the points associated with the entry of a new member in the active set, and therefore knots (change in slope) in the

[6] These knots correspond to intercepts of the coefficient path of the lasso regression as we vary λ.

piecewise linear function $\psi(\lambda)$. Let $A_k = \{j|\psi_j(\lambda_k) \neq 0\}$ be the lasso active set associated with λ_k. This implies that $A_1 = \emptyset$ and A_1 is a singleton, and $A_1, A_2, \cdots$ is the sequence of models induced by the lasso path.

Lasso may add or remove variables from its active set. At every time a variable is added to the active set, we want to test $A_k \supseteq A_0 = supp(\psi_0)$ where ψ_0 is the true parameter. This setting leads us to a sequence of nested hypothesis $H_1, H_2, \cdots$ such that each hypothesis H_k can be rejected if all previous hypotheses $H_1, H_2, \cdots H_{k-1}$ are rejected. In the next two subsections, we discuss how to construct test statistics at each step and how to control FDR under sequential setting.

Remark. The existing literature has considered a different null hypothesis. The structural breaks literature considers the so called "incremental null", i.e. that the kth discontinuity does not improve the fit, irrespective of whether A_k contains the true model. We instead consider the so-called "complete null" i.e., whether the true model is already contained in A_k. The complete null implies the incremental null, and in this sense our approach is more conservative. See Fithian et al. (2017) for further discussion on this point.

4.2 Covariance Test

Let $< \cdot, \cdot >$ denote the dot product. Let A_k be the active set before the knot λ_k, and suppose predictor j enters at λ_k. Define $\hat{\psi}_{A_k}(\lambda_{k+1})$ be the lasso solution at λ_{k+1} but constrained to the set A_k:

$$\hat{\psi}_{A_k}(\lambda_{k+1}) = \arg\min_{\psi_{A_k}} \frac{1}{2}\|M_m Y - M_m D_{A_k}\psi_{A_k}\|_2^2 + \lambda_{k+1}\|\psi_{A_k}\|_1.$$

Then the covariance test statistic as proposed in Lockhart et al. (2014) is:

$$T_k = \left(< M_m Y, M_m D\hat{\psi}(\lambda_{k+1}) > - < M_m Y, M_m D_{A_k}\hat{\psi}_{A_k}(\lambda_{k+1}) >\right)/\sigma^2. \tag{5}$$

Let the sign vector $s_{A_0} \in \{-1, 1\}^{|A_0|}$ encode the active set in vector form, and consider the event:

$$B = \Big\{ \text{The solution at step } k_0 \text{ in the lasso path has active set } A_k = A_0,$$
$$sign(s_{A_k}) = sign((D_{A_0})^+ Y) = s_{A_0}, \text{ and the next two knots are given by}$$
$$\lambda_{k_0+1} = \max_{j \notin A_{k_0} \cup \{j_{k_0}\}, s \in \{-1,1\}} \frac{D_j'(I - P_{A_{k_0}})Y}{s - D_j'(D'_{A_{k0}})^+ s_{A_{k_0}}}, \textit{and}$$
$$\lambda_{k_0+2} = \max_{j \notin A_{k_0+1}, s \in \{-1,1\}} \frac{D_j'(I - P_{A_{k_0+1}})Y}{s - D_j'(D'_{A_{k0+1}})^+ s_{A_{k_0+1}}} \cdot \mathbf{1}\left(\frac{D_j'(I - P_{A_{k_0+1}})Y}{s - D_j'(D'_{A_{k0+1}})^+ s_{A_{k_0+1}}} < \lambda_{k_0+1} \right) \Big\}.$$

Event B requires more than the event of consistent selection i.e $\{A_{k_0} = A_0\}$. It also requires the least squares estimate on A_0 has the same signs as this lasso estimate and the next two models in the sequence A_{k_0+1} and A_{k_0+2} are adding variables. A necessary and sufficient condition for $\mathbb{P}(B) \to 1$ is the irrepresentable condition in Zhao and Yu (2006).

Theorem 2 (Covariance Test under Measurement Error)**.** *At any given step k on the LASSO path, let the covariance test statistics being defined as in equation* (5)*. Under Assumption 1, 2, and 3, assume the approximation error satisfies $\varphi_m^2 = o(1/(\log(p)))$ and $\lambda_{k+1} \geq 4\sigma\sqrt{\log p}$ where σ. Then*

$$\lim_{n \to \infty} \mathbb{P}(T_k > t) \leq e^{-t}.$$

Thus the standard exponential distribution serves as a conservative bound.

Remark. The standard SEIVE estimates implies a rate of $O(K_m^{-2s})$ on φ_m^2, where K_m is the number of terms in S_m, s is the degree of the smoothness of $g(x_i)$. The tuning parameter for the nonparametric part K_m is chosen to increase as $n \to \infty$. Heuristically, $\varphi_m^2 = o(1/n^2)$ is required in order for the distribution of the test statistics not to be affected by the empirical process. One advantage of the covariance test is its asymptotic relies on p goes to infinity instead of n. As a result, the approximation error can be further relaxed to be order $o(1/(n \log(p)))$.

4.3 Sequential False Discovery Rate Control

At each step on the lasso path where a covariate is added to the active set, we can construct test statistic T_k and the p−value p_k associated with it. This leads to a sequential multiple comparison problem. Recall the False Discovery Rate (FDR) is defined as $\mathbb{E}\Big[V(\hat{k})/\max(1, \hat{k})\Big]$, where $V(\hat{k})$ is the number of null hypotheses among all the rejected hypotheses. The standard Benjamini-Hochberg method for FDR control can not be applied directly here because it is only valid for *simultaneous*

multiple comparison problems, which are sortable. We can not sort all of the p−values as our hypothesis are nested and have a natural order: the $\hat{k}$th hypothesis will only be rejected if the previous $\hat{k}-1$ hypothesis are all rejected. G'Sell et al. (2015) propose the forward stopping rule $\hat{k}_F$ to control FDR in a sequential MCP problem:

$$\hat{k}_F = \max\left\{k \in \{1, \cdots, n\} : -\frac{1}{k}\sum_{i=1}^{k}\log(1-\mathrm{p}_i) \leq \alpha\right\}.$$

This completes the estimator. To summarize, we set up the problem in the semi-lasso form of equation (3). This is transformed into a standard lasso using the projection matrix of the linear representation of $g(\cdot)$, as characterized by Lemma 1. Next, we compute the sequence of nested models using the LARS algorithm. For each addition of a discontinuity we compute the covariance test statistic described in Section 4.2. Finally, we select the kth model (and therefore all discontinuities contained) according to the forward stopping rule with a pre-specified false discovery rate chosen by the econometrician.

4.4 Advantages of the LASSO Approach

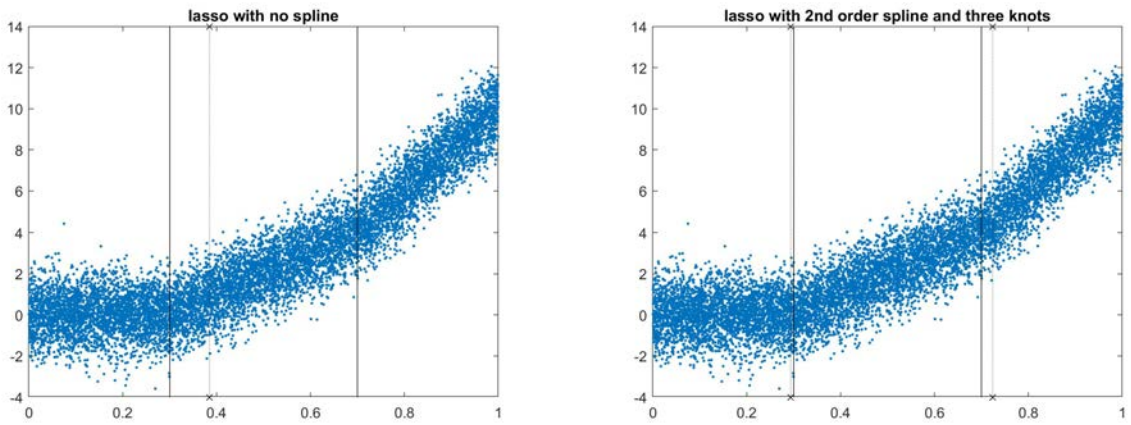
Our approach differs from mean-squared error minimization-based approaches in prior work. Instead, we have gone out of our way to linearize the problem in order to apply lasso regression methods. This has a number of advantages.

First, we are able to take advantage of developments in inference for lasso regression to construct uniformly valid confidence intervals. van de Geer et al. (2014) circumvents the post-selection uniformity issues raised in Leeb and Pötscher (2005) by computing approximate confidence intervals pre-selection.

Second, the linearity of our problem also allows us to treat $g(\cdot)$ flexibly while still retaining the properties of a lasso, as we establish in Lemma 1. In this way we are careful to only use *local* variation to identify discontinuities. This is more than just a matter of being agnostic about $g(\cdot)$. Even if the true functional form of $g(\cdot)$ is known to the econometrician, the model is still misspecified in the interim when it tests for the first of multiple actual discontinuities.[7] We highlight this point with a simulation of the following model, which consists of a linear $g(\cdot)$ and two positive kinks:

[7] At face value, this seems to contradict the intuition that an estimator is more efficient when known parametric forms are exploited. But this is a straw man – here, the known parametric form is simply used incorrectly.

Figure 1: Detection with Inflexible and Flexible $\hat{g}(\cdot)$



(a) (b)

Notes: This figure depicts a Monte Carlo simulation with 10,000 observations, two kinks at 0.3 and 0.7, and a constant $g(\cdot)$. First, we apply the standard covariance test and the true, constant functional form, with no flexible treatment of $\hat{g}(\cdot)$. One discontinuity is detected at a wrong location as in panel (a). Second, we use a basis spline approximation of $\hat{g}(\cdot)$ and both discontinuities are detected, see panel (b).

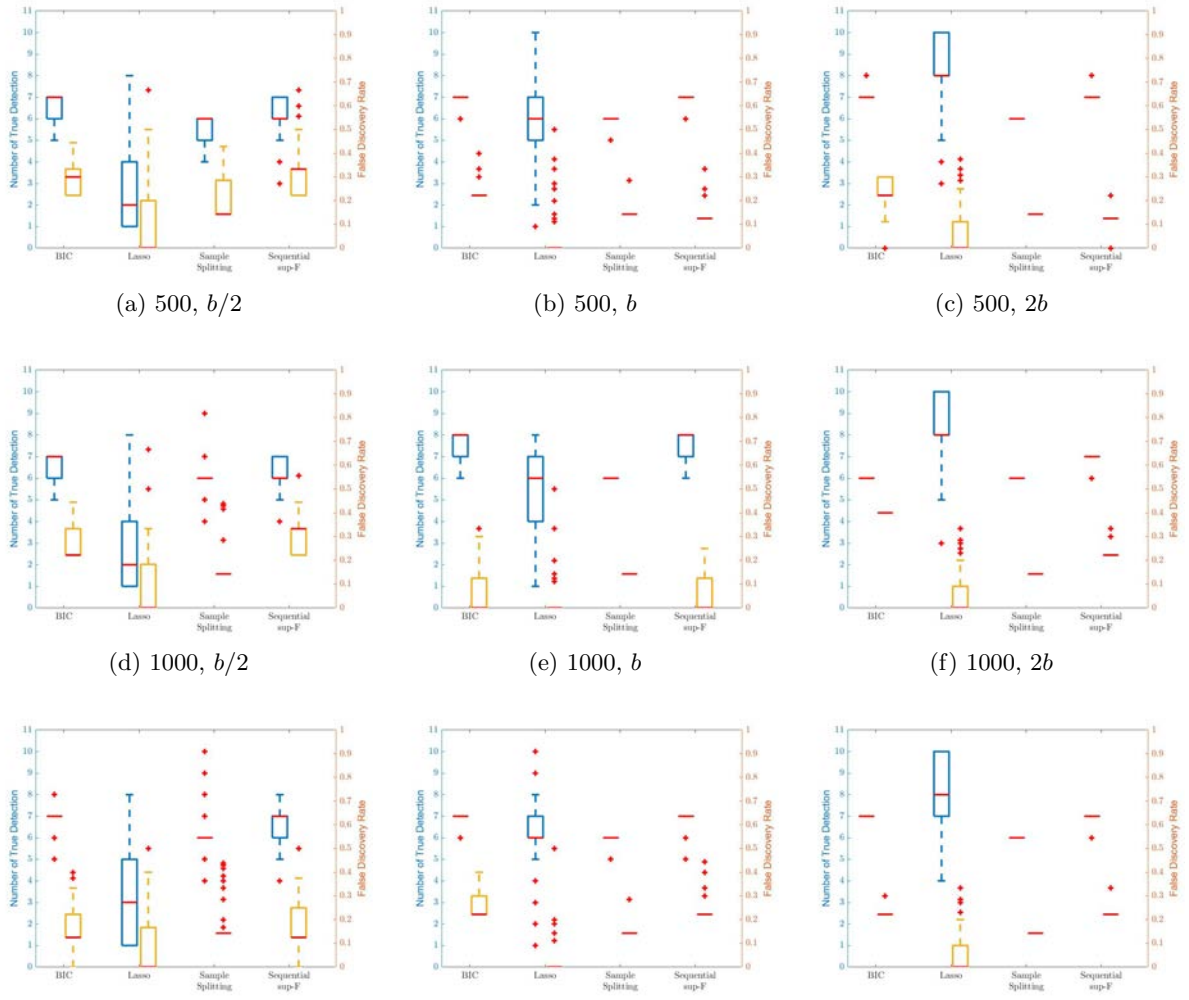
$$Y = \mathbf{1}(X > z_1)(X - z_1)\psi_1 + \mathbf{1}(X > z_2)(X - z_2)\psi_2 + \epsilon. \tag{6}$$

In our simulation, we let $(z_1, z_2) = (0.3, 0.7)$, $\psi_1 = 10$ and $\psi_2 = 10$ and $\epsilon \sim N(0, 1)$, and we take 10,000 independent draws. Results from the application of our method are presented in Figure 1. In Panel (a) we allow the econometrician to use the known functional form of $g(\cdot)$. Using that form leads them the erroneous conclusion that there is one break in between the two. In Panel (b), the econometrician uses a basis spline of 2nd order and five knots and identifies the two discontinuities. This difference comes from the fact that the flexible form for $\hat{g}(\cdot)$ allows the estimator to use only local variation to identify the first kink, ignoring the misspecification introduced by the fact that the second is not yet in the model.

Third, there are several advantages related to the detection of multiple breaks. In particular, the consistency of Bai (1997a) and Bai and Perron (1998) depends on $\alpha_T \to 0$ sufficiently slowly to control the multiple comparison problem asymptotically.[8] This asymptotic guarantee gives little guidance to finite-sample implementation, because the rate at which α should be set in applied

[8]From Bai (1997a) Theorem 11: "suppose that the size of the test α_T converges to zero slowly ($\alpha_T \to 0$ yet $\liminf_{t\to\infty} T_{\alpha_T} > 0$)" and from Bai and Perron (1998) Proposition 8, "If α_T converges to 0 slowly enough (for the test based on $F_T(l+1|l)$ to remain consistent)..."

Figure 2: Simulations: Comparing True Detections and FDR



(a) 500, $b/2$ (b) 500, b (c) 500, $2b$

(d) 1000, $b/2$ (e) 1000, b (f) 1000, $2b$

(g) 2000, $b/2$ (h) 2000, b (i) 2000, $2b$

Notes: These panels compare four different methods (BIC, lasso, sample splitting, and sequential sup-F) of detecting discontinuities across different sample sizes and sizes of discontinuities. Sample sizes vary between 100, 200 and 500. The size of discontinuities varies between b = [16, 8, 24, -16, -24, 24, 8, 16, 8, 24], $b/2$ and $2b$. The blue (left) box plots describe the number of true detections among the 10 discontinuities while the red (right) box plots describe the FDR.

exercises is unspecified. In contrast, the FDR guarantee of our test has a natural interpretation for applied work in real (i.e., finite) datasets. To highlight this point, we conducted a number of simulations.

The simulation is constructed with varying sample sizes ($n = 100, 200, 500$) and 10 jump discontinuities at z = (-1.5, -1, -0.5, -0.2, 0, 0.1, 0.2, 0.5, 1, 1.5). The data generating process is given

as

$$y_i = \sum_{j=1}^{10} b_j \cdot \mathbf{1}(x_i > z_j) + \epsilon_i$$

with x drawn i.i.d from a $N(0,1)$ process and $\epsilon_i \sim N(0,1)$ and b = [16, 8, 24, -16, -24, 24, 8, 16, 8, 24]. In this environment we consider four estimators: A BIC approach, our lasso-FDR estimator, sample splitting (Bai, 1997a), and the sequential sup-F test (Bai and Perron, 1998). The first approach, the BIC method, detects the most discontinuities among the four methods, between 6 and 8, but at a cost of overdetection and no control on the FDR, which hovers around 30%. The second, our lasso-FDR method, detects between 2 and 8 discontinuities among the 10 and with a controlled FDR rate below 5%. Lasso-FDR method outperforms all other when the size of the discontinuities are $2b$. Third, the sample splitting method uses classical Sup-F test as in Andrews (1993) at each step. Once a discontinuity is detected, the sample is split into two by the discontinuity and two Sup-F tests will be carried on the subsamples. In our case, sample splitting method detects around 3 - 6 discontinuities with a FDR in the range of 10% to 20%. Sample splitting may suffer from significant loss of efficiency when the sample size is small and when discontinuities are not evenly distributed. Fourth and fnally, the sequential sup-F method is based on the test proposed in Bai and Perron (1998). It is also based on the simultaneous minimization at each given number of discontinuities. The conditional sup-F test is testing the incremental null at each given step, as a result it should be less the conservative than the Lasso test at each step. Consistent with this, we find that it detects more of the true discontinuities, between 6 and 8, but it also has a much higher false discovery rate similar to the BIC method, often around 30%.

As a final point of comparison, in terms of computational efficiency, BIC method and Sequential Sup-F method both require dynamic programming, which is of order $O(n^2)$. On the other hand, the scale of Lasso method is of order $O(n)$ and it is much faster when n is large.

A fourth advantage of our LASSO approach is that it allows *backward steps* – that is, the lasso path can remove, as well as add, variables – which is related to our second point above concerning interim misspecification. Asymptotically this is irrelevant if $g(\cdot)$ is sufficiently flexible, but in finite samples we found that the algorithm would often attempt to smooth two nearby breaks with the addition of a single large one, later introducing the two finer ones and replacing the original. That finite-sample misstep is a feature of the "one-at-a-time" approach.

Fifth and finally, sup-F test-based approaches such as Andrews (1993), Bai and Perron (1998), Bai (1997a), Hansen (2017), use a trimming parameter to rule out boundaries of the support of X. More precisely, consider the first test of no discontinuity versus at least one discontinuity. The sup-F statistic can be written in LASSO notation as

$$T_{supF} = \lambda_1^2/\hat{\sigma}^2 = \sup_j Y' X_j X_j' Y/\hat{\sigma}^2.$$

It is then easy to show that

$$\mathbb{P}(T_{supF} > t) \to 1 \quad \text{as } n \to \infty \quad \forall t.$$

This result is shown with Corollary 1 in Andrews (1993). A trimming is required such that

$$T^*_{supF} = \sup_{j \in \Pi} Y' X_j X_j' Y,$$

– for $\Pi = [\pi_0 N \cdots, (1-\pi_0)N]$, T^*_{supF} converges to a Brownian bridge, where π_0 is the tuning parameter. The trimming may yield a boot in power for discontinuities inside $[\pi_0 N, (1-\pi_0)N]$, but lose power against discontinuities outside the interval.

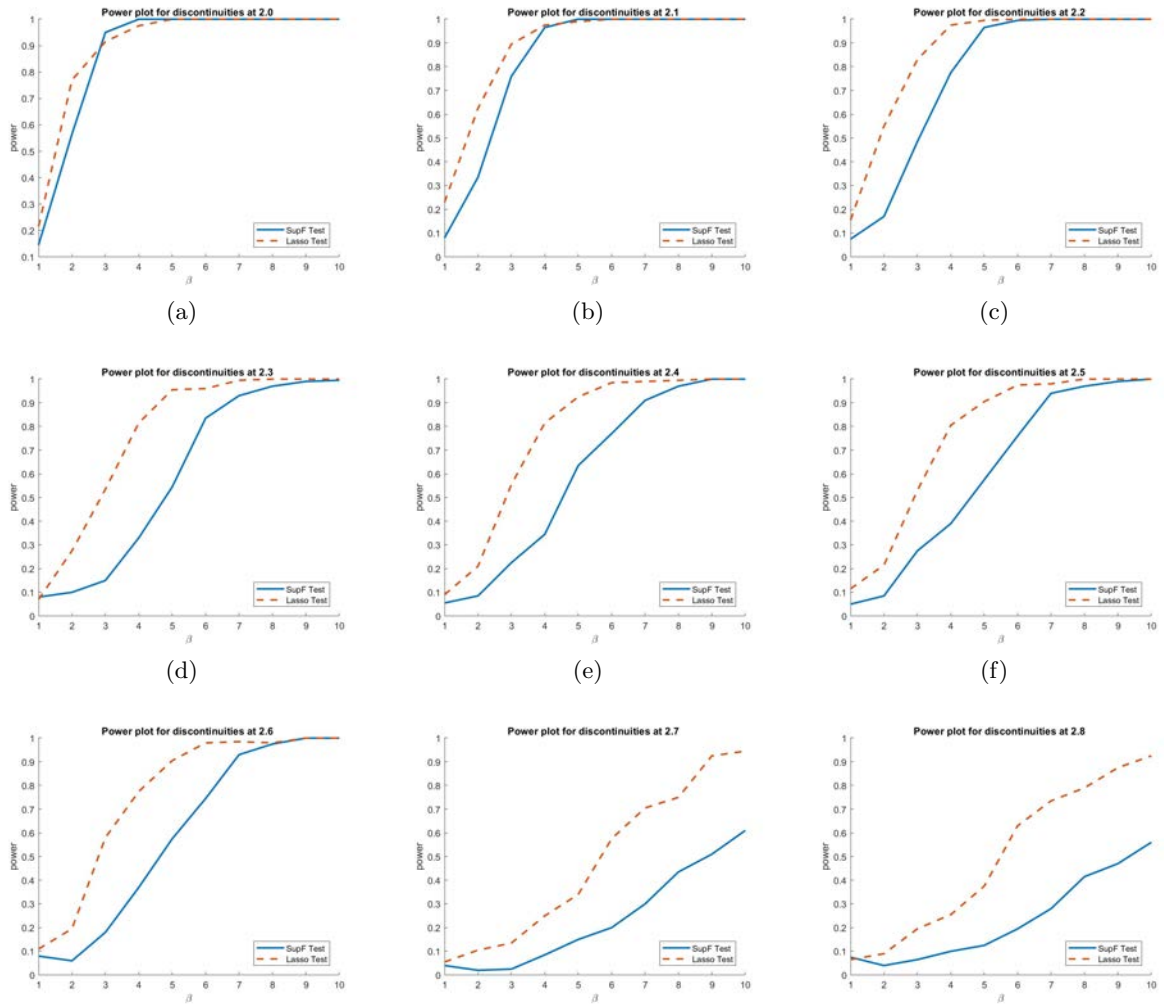
In contrast, our FDR lasso approach does not require the econometrician to choose a tuning parameter. This is particularly important if the discontinuity is located close to the boundary of the trimmed support. In the following simulation, we compare the power curve of sup-F test and LASSO test when the location of a jump discontinuity is moving towards the edge. We consider the data generating process as:

$$y_i = 1 + b \cdot \mathbf{1}(x_i > z) + \epsilon_i$$

where X and ϵ both generated under a standard normal distribution. We allow the location of the jump discontinuity z to vary from 2 to 2.8 and the size of the discontinuity to vary from 1 to 10. We choose 0.05 as the trimming parameter for the sup-F test, which is consistent with convention in this literature. The following graphs are generated with sample sizes $n = 200$ and with 200 simulations. As shown in Figure 3, as the discontinuity moves closer to the boundary, the LASSO test demonstrates better power comparing to the sup-F test.

Two last notes about the character of the test: First, it should not be surprising that it exhibits variable power, depending on the quality of the local approximation to $g(x)$. In regions where there is little data, our approach will fail to detect discontinuities. While it is tempting to interpret this

Figure 3: Simulations: Comparing power



(a) (b) (c)

(d) (e) (f)

(g) (h) (i)

Notes: This figure compares the power of the Sup-F test against our lasso FDR approach. We consider the case with jump discontinuity moving from 2.0 to 2.8, towards the boundary of the trimmed support. The magnitude of the jump is varying from 1 to 10 in all cases. The sup-F test is trimmed at 0.05.

as a bias in favor of flagging discontinuities where data is abundant, say near x', rather than where it is not, say near x'', this is not precisely correct: it is a question of variable power. Rather, one might say that it is a bias in favor of flagging discontinuities near x' rather than x'' *conditional on the existence of discontinuities near both.*

Second, as in the derivation of the rate-optimal λ, our approach requires an estimate of σ_ϵ at every stage of the forward-stop algorithm. We estimate this object using OLS conditional on the

discontinuities included so far, so that the estimate is consistent under the interim null of the sequential MCP.

5 Asymptotic Properties

5.1 Integrated Mean Squared Error

The selection consistency for our procedure can be guaranteed by allowing FDR $\to 0$ as $n \to \infty$. This is a similar argument to that of Bai and Perron (1998) who let the family-wise error rate converge to 0.

Prior work has assumed that we are given the correct specification for the continuous function $g(x)$. In this section, we explore the prediction consistency based on our estimator, holding the FDR fixed, and propose a procedure that is asymptotically valid to determine the specification for $g(x)$.

Formally, let (y_i, x_i), $i = 1, \cdots, n$ be the sample we observed. We suspect the conditional mean is $g(x) = E(y|x)$ with breaks at $z_1 < z_2 < \cdots < z_{s_0}$. Let ϵ_i be a mean 0 process with variance σ^2 unknown. The data generating process can be summarized as:

$$y_i = g(x_i) + \sum_{j=1}^{s_0} \psi_j d_j(x_i) + \epsilon_i \tag{7}$$

where $d_j(x_i)$ represents jump discontinuities.

Let $(\hat{\beta}, \hat{\psi})$ be the estimator from (3) where $g(x)$ is approximated with spline or power series. Define $\hat{I} = \{i, \hat{\psi}_i \neq 0\}$. Recall $D_l(x_i)$ as the lth term in $D(x_i)$. The integrated mean squared error (IMSE) for our estimator can be written as:

$$IMSE_n(m) = \int \mathbb{E}\left(\hat{g}_m(x) + \sum_{l \in I} D_l(x)\hat{\psi}_l - g(x) - \sum_{j=1}^{s_0} d_j(x)\psi_j \right)^2 f(x)dx$$

where

$$\hat{g}_m(x) + \sum_{l \in I} D_l(x)\hat{\psi}_l = S_m(x)'\hat{\beta}_m + \sum_{l \in I} D_l(x)\hat{\psi}_l$$
$$g(x) + \sum_{j=1}^{s_0} d_j(x)\psi_j = S_m(x)'\beta_m + \sum_{j=1}^{s_0} d_j(x)\psi_j + r_m(x)$$

Theorem 3 (Convergence Rate for IMSE)**.** *Under Assumption 1, 2, 3 and 4, there exist a constant C such that*

$$IMSE_n(m) \leq 2\varphi_m^2 + 2\sigma^2 \frac{K_m}{n} + C \cdot \frac{\log(p)}{n}$$

The constant C depends on the sparsity s_0 and the compatibility constant ω. The compatibility constant is defined precisely in the proof (see Appendix A). For intuition, the lower bound on this constant is restricting collinearity between the spline approximation of $g(\cdot)$ and the design matrix D_n. This would fail if, for instance, the econometrician sought to identify kinks but approximated $g(\cdot)$ with a piecewise linear spline, which has, by construction, kinks.

The first two terms in the IMSE expression are approximation error from the spline approximation to $g(\cdot)$, while the last term is the error introduced by the detection of, or failures to detect, discontinuities in the lasso procedure. We show that the empirical process for LASSO approximation can be bounded using the Glivenko-Cantelli and P-Glivenko-Cantelli theorems (see van der Vaart (1998)).

From Theorem 3, $IMSE_n(m) \to 0$ as $\phi_m^2 \to 0$ and $\lambda \to 0$. Consider the following cross-validation process for a set of models $m = \{0, 1, \cdots, M\}$. We first split the sample into training and prediction. We use the training data to detect the number of discontinuities and the parameters in $\hat{g}(x)$ and then we use the prediction data to compute $\hat{IMSE}_n(m)$ under each model. We choose the model that gives the smallest $\hat{IMSE}_n(m)$.

This procedure is asymptotically valid as when we shrink FDR to 0, $\lambda \to 0$, as $n \to \infty$, $\varphi_m^2 \to 0$ and $\frac{K_m}{n} \to 0$. Thus $\hat{IMSE}_n(m)$ will be minimized under the true parameters at 0.

5.2 Distribution of Break Point

The consistency of the selected break points follows from the irrepresentable condition. In this section, we compare our estimators with those in Bai (1997a) and consider the limiting distribution of the location of the break when the magnitude of the break is shrinking to 0 as $n \to \infty$. We derive the distribution for the standard case when only one jump (or kink) discontinuity exists and $g(\cdot)$ is constant. We show that the lasso-detected location has similar distribution as the traditional mean square error estimates. However, as in that prior literature, the assumptions are more restrictive than in the above.

Theorem 4 (Distribution of Jump Detection)**.** *Assume there is only one jump break such that the*

data generating process is:

$$y_i = d_1^0(x_i)\psi_1 + \epsilon_i, \ \textit{with}$$
$$d_1^0(x_i) = \begin{cases} 0 & \textit{if } x_i \leq z, \\ 1 & \textit{if } x_i > z. \end{cases}$$

Let $k > 0$ be an integer such that

$$x_{(k)} \leq z < x_{(k+1)}$$

Let $\rho_n = O_p(\|\psi_1\|)$ and $\hat{k}$ be the lasso estimator from (3). Assume $\rho_n \to 0$ but $\sqrt{n}\rho_n \to \infty$. There exist a constant v, such that

$$\rho_n^2(\hat{k} - k) \to_d \arg\max_v(-|v| + 2W(|v|)),$$

–where $W(v)$ is a wiener process of degree v.

In the kink discontinuity case, the convergence rate is slower than the jump discontinuity case. When the size of the discontinuity is constant with respect to the sample size n, we have a square-root-n convergence rate to a normal distribution, similar to the result in Hansen (2017).

Theorem 5 (Distribution of Kink Detection)**.** *Assume there is only one kink break such that the data generating process is:*

$$y_i = d_1^1(x_i)\psi_1 + \epsilon_i,$$
$$d_1^1(x_i) = \begin{cases} 0 & \textit{if } x_i \leq z \\ (x_i - z) & \textit{if } x_i > z \end{cases}.$$

Let $k > 0$ be an integer such that

$$x_{(k)} \leq z < x_{(k+1)}.$$

Assume $\kappa_1 = \mathbb{E}(x - x_{(k)})$ and $\kappa_2 = \mathbb{E}(x - x_{(k)})^2$ both exit. Let $\psi_1 = \frac{\rho_n}{\sqrt{n-k}}$ and $\hat{k}$ be the lasso estimator from (3). For any $\rho_n = O(1)$,

$$\rho_n(x_{(k)} - x_{(\hat{k})}) \to_d N\left(0, \frac{1}{2\left(1 - \frac{\kappa_1^2}{\kappa_2}\right)}\right).$$

The results in both Theorem 4 and Theorem 5 hold when the size of the kink discontinuity shrinks towards 0 at a speed slower than $1/\sqrt{n}$. When the shrinkage speed is faster than $1/\sqrt{n}$, the selection consistency results of LASSO estimator no longer holds and the distributions converge to

a non-standard process. As a result, no uniformly valid inference for the location of discontinuity exits.

6 Applications

6.1 Placebo Tests for Structural Breaks

A natural application of our procedure is to placebo testing for discontinuities in the RDD setting. While the location of a break is often known in advance from institutional details – for instance, majority rule voting implies a threshold at fifty percent – empirical researchers in this area would like to validate the existence of these discontinuities and know whether there are other, unanticipated ones that may affect results. This exercise takes advantage of the fact that we allow a flexible specification of $g(\cdot)$ and multiple potential breaks.

By way of illustration, we replicate the electoral RD design of Lee (2008) using our estimator. Data for this exercise are available from the online data archive of Angrist and Pischke (2009).[9] Consistent with the design of that paper, we detect the discontinuity at the vote share margin of winning of zero (to be precise, 0.0003), and no further discontinuities, as depicted in Figure 4.

6.2 A Jump Discontinuity with an Unknown Location

Next, we follow Card et al. (2008) and consider our test in the context of detecting neighborhood "tipping." The theory of Schelling tipping points argues that intolerant white preferences generate residential dynamics that lead inexorably to neighborhood segregation. Let m_t denotes the minority share of a neighborhood at time t. Their model suggests a discontinuity in the expected innovation in minority share conditional on m_{t-1} so that

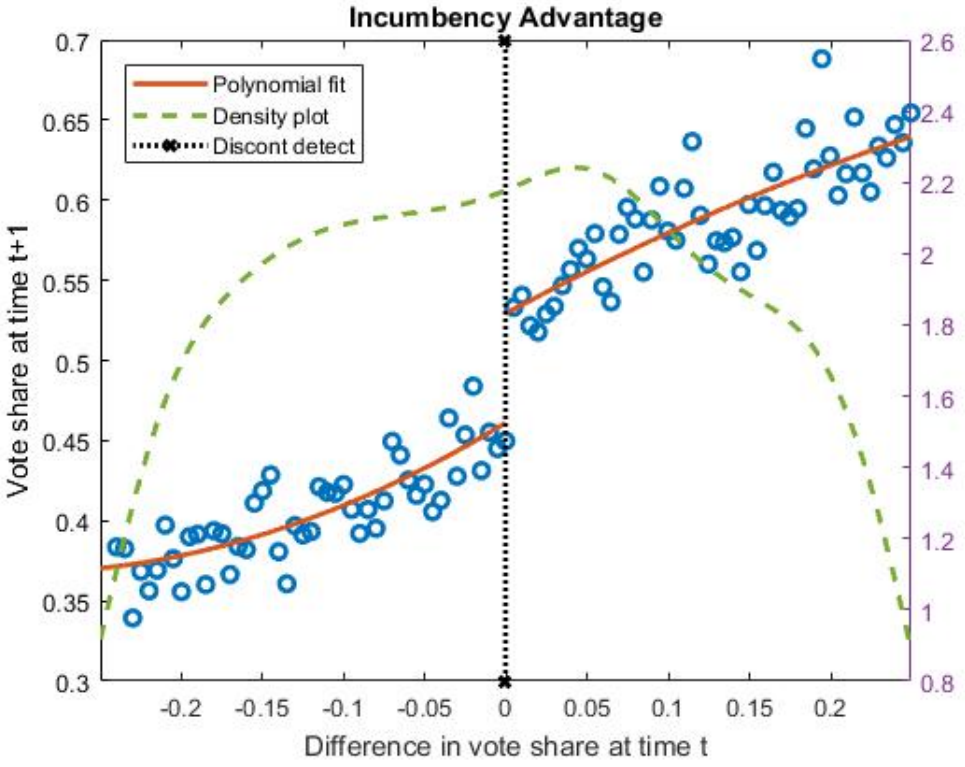
$$\mathbb{E}(\Delta m_t | m_{t-1}) = \mathbf{1}(m_{t-1} < m^*) g(m_{t-1}) + \mathbf{1}(m_{t-1} \geq m^*) h(m_{t-1}),$$

– for some unknown threshold m^* and some functions g and h, with a jump discontinuity at the threshold.

Our design mirrors that in Card et al. (2008). We obtained data the Neighborhood Change Database (NCDB) from 1970-2010. As in their work, the unit of observation is the census tract, the change

[9] See https://economics.mit.edu/faculty/angrist/data1/mhe.

Figure 4: Regression Discontinuity Design in Lee (2008)



Notes: We use data from Lee (2008) and apply our method as a placebo test for regression discontinuity design. The x-axis represents the democratic vote share margin of victory (or loss) in period t. The y-axis represents the vote share in the subsequent election at $t+1$. The red line indicates a quadratic polynomial fit determined by cross-validation. The green dash line represents the density plot and the marked black dotted line indicates the only detected discontinuity at 0.

in minority shares are calculated using ten-year window, and we run our estimator separately for each time period and metropolitan area.

Their paper assumes the existence of a single break and searches over the range $[0, 2/3]$ to find it, assuming that $g(\cdot)$ is a constant function. Here we weaken the assumptions of their approach in two ways: first, we allow for multiple (or no) breaks. Second, we allow for more flexible forms for the $g(\cdot)$ function, including our preferred specification, which chooses the degree of the polynomial series using cross-validation.

Results are reported in Table 1. In the first part of that table, we report the number of detections allowing for more or less flexibility in the specification of $g(\cdot)$. For instance, in the top-left part of the table, we see that the sup-F test finds breaks in 87 or the 93 metropolitan areas, while the lasso

approach finds only 51.

Several features are apparent: first, especially in the presence of a more flexible functional form, many metropolitan areas appear to exhibit no evidence of a structural break or "tipping point." For between a quarter and half of the cases where a structural break is detected, the lasso approach actually detects two rather than one. Second, many of these breaks, especially once we allow for a more flexible functional form for $g(\cdot)$, are positive jumps rather than negative jumps, which is inconsistent with the model of Schelling-style tipping points. These inconsistencies are obscured when we assume the existence of a single jump, and so the generality afforded by our approach is especially important as we seek to assess the fit of the model. We do not believe this undermines the research agenda. Rather, in the spirit of "Sherlock Holmes inference" (Leamer, 1983), these results suggest new directions – perhaps the salient question is understanding why some metropolitan areas exhibit tipping behavior and others do not.

In addition, variation in the results highlights the first-order significance of assumptions on the parametric form of $g(\cdot)$ for the performance of the detection algorithm. Much of the power of the Sup-F test to identify detections seems to be coming from the restrictive functional form assumption, that $g(\cdot)$ is a constant function on $[0, 2/3]$.

The second panel of Table 1 reports results for the lasso approach where the order of approximation of $g(\cdot)$ is chosen by cross-validation. We note that in this, our preferred specification, there is substantial heterogeneity in the choice of order, and second, that only in substantially less than half of metropolitan areas are we able to find any evidence of a structural break.

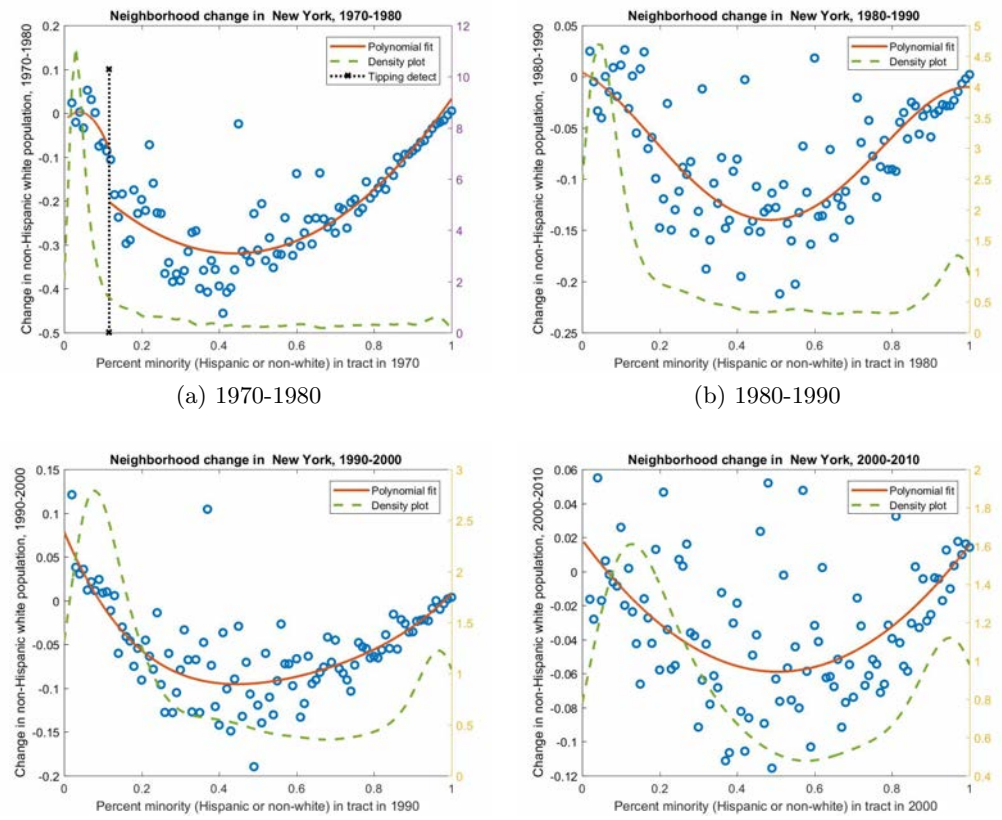
Figures 5 and 6 present results for two cities, New York and Washington, DC respectively. For New York, the procedure detects a tipping point only for changes over 1970–1980, not for any other interval. Moreover, the importance of the quadratic fit of $g(\cdot)$, here chosen by cross-validation, is visually evident. In the absence of sufficient flexibility, it would be possible to detect a discontinuity generated primarily by misspecification. And, given the locally negative slope, that false detection would be likely appear to be a negative jump, equipollent with the predictions of the theoretical model of tipping points. Likewise in DC, for half of the time periods no discontinuity is detected. Moreover, for the period 1980–1990, the algorithm instead detects a *positive* jump, which has no interpretation in the theory.

Table 1: Detection of Tipping Points

	1970-1980		1980-1990		1990-2000		2000-2010	
	sup-F	Lasso	sup-F	Lasso	sup-F	Lasso	sup-F	Lasso
Full sample:								
Constant:								
MSAs with detections	90	48	91	66	92	63	70	47
Mean detected location	[0.1057]	[0.0834]	[0.1289]	[0.1103]	[0.1335]	[0.1077]	[0.1601]	[0.1286]
(Multiple/Postive detections)	(-/4)	(20/3)	(-/2)	(20/7)	(-/1)	(24/3)	(-/0)	(11/3)
Linear:								
MSAs with detections	69	42	78	42	73	51	41	33
Mean detected location	[0.0901]	[0.0718]	[0.1209]	[0.1135]	[0.1258]	[0.1116]	[0.1598]	[0.1236]
(Multiple/Postive detections)	(-/23)	(9/15)	(-/8)	(5/8)	(-/1)	(4/5)	(-/1)	(4/4)
Qudratic:								
MSAs with detections	55	35	54	39	38	29	11	20
Mean detected location	[0.0873]	[0.1090]	[0.1113]	[0.0989]	[0.1137]	[0.1198]	[0.1540]	[0.1094]
(Multiple/Postive detections)	(-/25)	(9/19)	(-/26)	(9/16)	(-/12)	(4/7)	(-/1)	(2/8)
3rd order:								
MSAs with detections	48	30	51	43	25	24	3	16
Mean detected location	[0.0990]	[0.0952]	[0.1050]	[0.0796]	[0.1049]	[0.1312]	[0.1442]	[0.1157]
(Multiple/Postive detections)	(-/25)	(10/12)	(-28)	(6/22)	(-/7)	(3/6)	(-/1)	(3/7)
4th order:								
MSAs with detections	50	34	46	48	25	25	2	16
Mean detected location	[0.1040]	[0.0916]	[0.1086]	[0.1030]	[0.1049]	[0.0882]	[0.1452]	[0.1123]
(Multiple/Postive detections)	(-/27)	(8/17)	(-/17)	(8/21)	(-/10)	(4/9)	(-/0)	(4/8)
Cross-validation:								
MSAs with detections	62	48	63	51	47	37	18	20
Mean detected location	[0.0932]	[0.0834]	[0.1061]	[0.1042]	[0.1134]	[0.1347]	[0.1601]	[0.1615]
(Multiple/Postive detections)	(-/26)	(20/21)	(-/20)	(11/22)	(-/11)	(7/12)	(-/0)	(4/8)
Percentage of models selected:								
Constant	20.62%	7.22%	15.32%	10.81%	22.12%	7.96%	18.58%	13.27%
1st order	13.40%	13.40%	12.61%	10.81%	14.16%	15.93%	15.04%	17.70%
2st order	18.56%	24.74%	19.82%	23.42%	24.78%	21.24%	26.55%	27.43%
3st order	19.59%	17.53%	22.52%	23.42%	19.47%	29.20%	25.66%	21.24%
4st order	27.84%	37.11%	29.73%	31.53%	19.47%	25.66%	14.16%	20.35%
# of MSAs in sample	97		111		113		113	

Notes: Here we compare the performance of the Sup-F test and our lasso FDR approach using the NCDB data to detect tipping points in patterns of neighborhood demographics. The upper portion of the table uses the four period windows and five different models for the functional form of $\hat{g}(\cdot)$. The lower portion reports results for our preferred procedure, which uses cross-validation to determine the order of the approximation.

Figure 5: Change in Minority Share: New York



(a) 1970-1980

(b) 1980-1990

(c) 1990-2000

(d) 2000-2010

6.3 Regression Kink with an Unknown Threshold

As an opportunity to compare our approach to the traditional structural breaks literature, we follow the replication of Reinhart and Rogoff (2010) in Hansen (2017), and achieve a similar size and power in the detection of the first discontinuity. In this model, the discontinuity represents a structural break in the relationship between the ratio of debt to GDP and the growth rate of GDP.

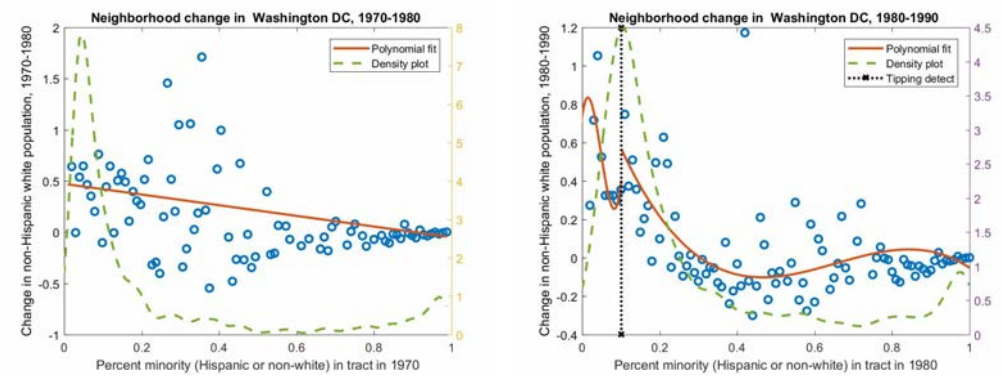
Our setup mirrors Hansen (2017); the data are entirely simulated. The data generating process is the following:

$$y_t = \beta_1(x_t - \gamma)_- + \beta_2(x_t - \gamma)_+ + \beta_3 y_{t-1} + \beta_4 + \epsilon_t \tag{8}$$

– where $\epsilon_t \sim N(0, \sigma^2)$.

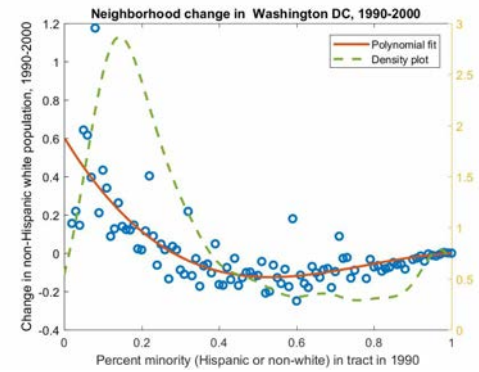
To evaluate size, we set $\beta_1 = \beta_2 = 0$, $\beta_3 = 0.3$, $\beta_4 = 3$, and $\sigma^2 = 16$ to match the empirical

Figure 6: Change in Minority Share: Washington DC

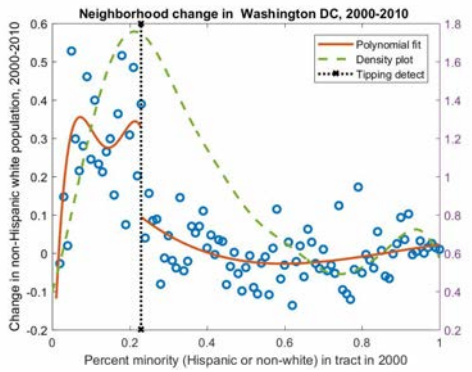


(a) 1970-1980

(b) 1980-1990



(c) 1990-2000



(d) 2000-2010

estimates from Reinhart and Rogoff (2010).

In Table 2 we report results from one thousand simulations with a bootstrap size of one thousand, and compare our results to the MSE-based detection algorithm of Hansen (2017). Both tests exhibit no meaningful size distortion.

Next, in Table 3 we compare the power of the two tests. We consider a range of values for β_2, from -0.16 to -0.02, and set the kink point at $\gamma = 40$. Nominal size is set at $\alpha = 0.05$. The power curves are similar, however the lasso approach performs slightly better when β is small and slightly worse when β is large.

Table 4 documents the coverage of the two approaches. We do not condition on detection. This makes clear the failure of uniform validity – as $\beta \to 0$, the coverage rate of the standard approach erodes, while the FDR lasso approach holds at 90%, even when there is no discontinuity.

Table 2: Type I error of Lasso test vs threshold F test

	α				
	0.05	0.10	0.15	0.20	0.25
threshold-F	0.051	0.0990	0.1469	0.1970	0.2460
Lasso	0.071	0.1120	0.1520	0.2040	0.2450

Notes: We compare the size of our lasso FDR test with threshold F test when there is no discontinuity in the true data generating process. We vary the nominal level of control from 5% to 25% and report the empirical incorrect rejection rate for 1000 simulations.

Finally, in Table 5, we show bootstrap results for the break location $\hat{\gamma}$ with comparison between the standard and lasso method. For standard method, we follow Hansen (2017) and report coverage and length of 95% confidence interval using native ($\pm 1.645 s(\hat{\gamma})$), percentile, inverse percentile, symmetric percentile, and C_γ, C_{γ^*} methods. For the lasso method, we report coverage and length of 95% confidence interval using the native ($\pm 1.645 s(\hat{\gamma})$), percentile, inverse percentile, and symmetric percentile method. Let $\theta^*_{(\alpha/2)}$ denote the $(\alpha/2)$ percentile of the bootstrap statistic θ^*. The percentile confidence interval is represented by $(\theta^*_{(\alpha/2)}, \theta^*_{(1-\alpha/2)})$. The inverse percentile confidence interval is represented by $(2\theta - \theta^*_{(1-\alpha/2)}, 2\theta - \theta^*_{(\alpha/2)})$. The symmetric percentile is represented by $(\theta - \Delta^*_{(\alpha/2)}, \theta + \Delta^*_{(\alpha/2)})$, where $\Delta^*_{(\alpha/2)})$ is the $\alpha/2$ quantile of $|\theta^* - \theta|$. Note that C_γ, C_{γ^*} is not available under the lasso method because the test specified in our setting can not be inverted in a similar way to the standard MSE test. While the coverage between MSE and lasso methods are very similar, the length of confidence interval is smaller under the lasso method.

Table 3: Power of of Lasso test vs threshold F test with nominal size 10%

	β							
	-0.02	-0.04	-0.06	-0.08	-0.10	-0.12	-0.14	-0.16
Hansen	0.0691	0.1130	0.2010	0.3110	0.4680	0.6200	0.7500	0.8540
Lasso	0.0920	0.1540	0.2626	0.3740	0.4690	0.5460	0.6690	0.6890

Notes: We compare the power of our lasso FDR test with threshold F test when there is one kink discontinuity in the true data generating process. We vary the magnitude of discontinuity from -0.02 to -0.16 and report the empirical rate of detection under 1000 simulations. The nominal level of rejection is controlled at 10% level.

Table 4: Coverage rate of nominal 90% CI for magnitude of discontinuity

	β				
	0.00	-0.01	-0.02	-0.04	-0.16
$\hat{\beta} \pm 1.645 s(\hat{\beta})$	0.8250	0.7950	0.8050	0.8100	0.8600
Percentile	0.8100	0.8250	0.8150	0.8450	0.9200
Inverse Percentile	0.8500	0.8650	0.8700	0.9000	0.8750
Symmetric Percentile	0.8600	0.8650	0.8650	0.8950	0.8950
lasso de-biased	0.8950	0.9000	0.9000	0.8950	0.9000

Notes: We compare the coverage probability for the magnitude of the discontinuity using post fitting and de-biased methods. For post fitting, we use the full sample to estimate the location of discontinuity and then use the full sample to construct confidence interval conditioning on the detected discontinuity. For lasso de-biased method, we first use the full sample to detect the discontinuity and its lasso estimator. We then computer the bias for lasso estimator based on its analytic form and add it back to the lasso estimator.

6.4 Signaling in Online Bargaining

Finally, we replicate a result from Backus et al. (2018b), that the use of a round-number asking price in Best Offer listings on eBay.com elicits lower offers from buyers. In their paper, this result is a component of the argument that round numbers are signals of sellers' willingness to bargain. We are interested in the following regression:

$$\text{Mean First Buyer Offer}_i = g(\text{Asking Price}_i) + \sum_{z \in \mathcal{Z}} \beta \mathbb{1}_z(\text{Asking Price}_i) + \epsilon_i.$$

Here $\mathcal{Z}$ is a set of discontinuities, $\mathbb{1}_z(x)$ is an indicator function for the event $x \in [z, z+1)$, and we are interested in selecting the correct active set of point (or rather, unit-length) discontinuities. Our data are drawn from Collectibles listings using the publicly available Best Offer bargaining dataset introduced by Backus et al. (2018a).[10] We use observations where the asking price is in $[x-25, x+25]$ for $x \in \{100, 200, 300, 400, 500\}$.

Results are presented in Figure 7. Solid vertical lines represent the selected set – in each scenario, the unit interval round to the nearest 100 is selected, and in the 400's case, we also select 413. This is consistent with the round-number effects identified by that paper.

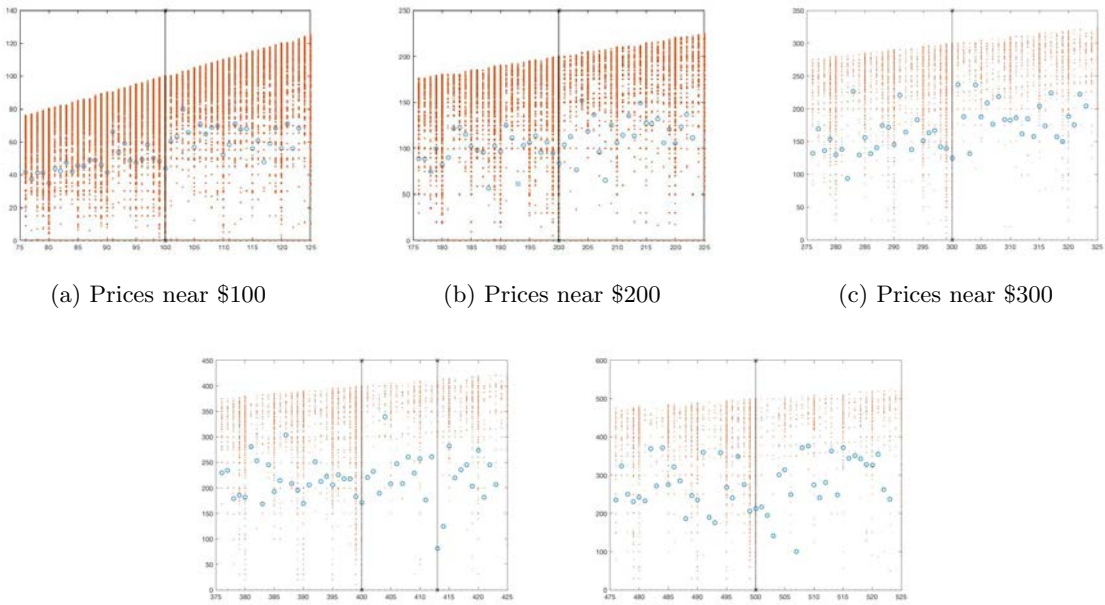
[10] See http://www.nber.org/data/bargaining/.

Table 5: Coverage rate of nominal 90% CI for the location of discontinuity

	α				
	0.05	0.10	0.15	0.20	0.25
Threshold F (MSE) method :					
$\hat{\gamma} \pm 1.645 s(\hat{\gamma})$	0.7900	0.7200	0.6850	0.6450	0.6050
	(37.9457)	(31.8450)	(27.8699)	(24.8114)	(22.2712)
Percentile	0.9550	0.8900	0.8650	0.8000	0.7650
	(40.7975)	(32.5250)	(27.0200)	(23.2675)	(20.0700)
Inverse Percentile	0.8250	0.7950	0.7350	0.7100	0.6800
	(40.7975)	(32.5250)	(27.0200)	(23.2675)	(20.0700)
Symmetric Percentile	0.9000	0.8450	0.7800	0.7500	0.7150
	(41.9400)	(32.1950)	(26.5050)	(22.5850)	(19.5100)
C_γ	0.9050	0.8400	0.7800	0.7150	0.6450
	(30.6950)	(24.6100)	(20.9950)	(18.4050)	(16.0800)
C_{γ^*}	0.9200	0.8450	0.8050	0.7700	0.7250
	(33.3950)	(27.3200)	(23.5250)	(20.9050)	(18.6600)
Lasso (Covariance) method:					
Percentile	0.9500	0.8650	0.8000	0.7700	0.7250
	(28.0704)	(22.3397)	(18.9569)	(16.6049)	(14.5919)
Inverse Percentile	0.9200	0.8550	0.7850	0.7350	0.6850
	(28.0704)	(22.3397)	(18.9569)	(16.6049)	(14.5919)
Symmetric Percentile	0.9500	0.8600	0.7950	0.7550	0.6900
	(28.0731)	(22.2410)	(18.9754)	(16.4391)	(14.4893)

Notes: Here we compare the coverage probabilities. The location of discontinuity is selected using either mean-square-error criterion or largest covariance between outcome and regressors, the basis for our lasso approach. We construct confidence intervals and report the empirical coverage under different methods. The length of confidence interval is reported in parentheses.

Figure 7: Signaling in Online Bargaining



(a) Prices near $100 (b) Prices near $200 (c) Prices near $300

(d) Prices near $400 (e) Prices near $500

Notes: Here we apply our lasso FDR method to the best offer bargaining dataset, in intervals 25 above and below the conjectured discontinuities. The x-axis are the price offered initially and y-axis is the final sell price. The orange dots are original data and blue circles depict the mean. The black line is the detected discontinuities using LASSO method.

7 Discussion

This paper has endeavored to introduce a new method to detect discontinuities using lasso regression. The method is robust to an unknown number, type, location, and magnitude of discontinuities.

Searching over a function to detect discontinuities is, by nature, an exercise in data snooping and so post-selection inference is problematic, here as in prior work. This problem is complicated further by allowing for multiple breaks, which makes the multiple selection problem sequential in character and, as we showed, raises substantial problems for existing approaches. Our solution to these problems is to avoid post-selection inference altogether, by building the desired inferential guarantee, false discovery rate control, into the selection process itself.

Our lasso-based approach has the added advantage of linearity, so that it easily accommodates flexible estimation of the continuous part of the function. We showed that this flexibility is more

than mere agnosticism. Even if the empiricist knows the true functional form, the sequential character of testing for discontinuities introduces interim misspecification that can lead to false detections.

We believe that the value of such an approach is twofold, and we highlighted both in our applications. First, when the location of a theoretically-motivated discontinuity is unknown, as in Card et al. (2008) and Reinhart and Rogoff (2010), our approach offers a flexible approach to detection. But second, even when we posit the existence of a discontinuity in a particular location, as in Lee (2008) and Backus et al. (2018b), the method is a useful placebo test for additional discontinuities, a specification test that can alert the econometrician to either a failure of their assumed model, or the location of additional discontinuities that may complicate estimation.

Finally, we hope that we have made a small contribution to the growing literature on the adoption of machine learning techniques in a way that maintains interpretable inferential guarantees, an important area for future work as empirical researchers economics adopt those tools.

References

Andrews, D. W. K. (1993). Tests for parameter instability and structural change with unknown change point. *Econometrica*, 61(4):821–856.

Angrist, J. D. and Pischke, J.-S. (2009). *Mostly Harmless Econometrics: An Empiricist's Companion*. Princeton University Press.

Backus, M., Blake, T., Larsen, B., and Tadelis, S. (2018a). Sequential bargaining in the field: Evidence from millions of online bargaining threads. NBER Working Paper No. 24306.

Backus, M., Blake, T., and Tadelis, S. (2018b). On the empirical content of cheap-talk signaling: An application to bargaining. forthcoming, *Journal of Political Economy*.

Bai, J. (1997a). Estimating multiple breaks one at a time. *Econometric Theory*, 13(3):315–352.

Bai, J. (1997b). Estimation of a change point in multiple regression models. *The Review of Economics and Statistics*, 79(4):551–563.

Bai, J. and Perron, P. (1998). Estimating and testing linear models with multiple structural changes. *Econometrica*, 66(1):47–78.

Basu, P., Cai, T. T., Das, K., and Sun, W. (2017). Weighted false discovery rate control in large-scale multiple testing. *Journal of the American Statistical Association*.

Belloni, A., Chernozhukov, V., and Kato, K. (2014). Uniform post selection inference for lad regression and other z-estimation problems. Working Paper.

Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (methodological)*, 57(1):289–300.

Bühlmann, P. and van de Geer, S. (2011). *Statistics for High-Dimensional Data*. Springer.

Card, D., Mas, A., and Rothstein, J. (2008). Tipping and the dynamics of segregation. *The Quarterly Journal of Economics*, 123(1):177–218.

Donald, S. and Newey, W. K. (1994). Series estimation of semilinear models. *Journal of Multivariate Analysis*, 50(1):30–40.

Efron, B., Hastie, T., Johnstone, I., and Tibshirani, R. (2004). Least angle regression. *The Annals of Statistics*, 32(2):407–499.

Fan, J. and Han, X. (2016). Estimation of the false discovery proportion with unknown dependence. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, pages n/a–n/a.

Fan, J., Han, X., and Gu, W. (2012). Estimating false discovery proportion under arbitrary covariance dependence. *Journal of the American Statistical Association*, 107(499):1019–1035.

Fithian, W., Sun, D., and Taylor, J. (2015). Optimal inference after model selection. Working Paper.

Fithian, W., Taylor, J., Tibshirani, R., and Tibshirani, R. (2017). Selective sequential model selection.

G'Sell, M. G., Wager, S., Chouldechova, A., and Tibshirani, R. (2015). Sequential selection procedures and false discovery rate control. Working Paper.

Hansen, B. E. (2000). Sample splitting and threshold estimation. *Econometrica*, 68(3):575–603.

Hansen, B. E. (2014a). *Nonparametric Sieve Regression: Least Squares, Averaging Least Squares, and Cross-Validation*. Oxford Handbooks. Oxford Handbook of Applied Nonparametric and Semiparametric Econometrics and Statisics.

Hansen, B. E. (2014b). *Oxford Handbook of Applied Nonparametric and Semiparametric Econometrics and Statistics*. Oxford University Press.

Hansen, B. E. (2017). Regression kink with an unknown threshold. *Journal of Business and Economic Statistics*, 35:228–240.

Hawkins, D. L. (1987). A test for a change point in a parametric model based on a maximal wald-type statistic. *Sankhyā: The Indian Journal of Statistics*, 49(3):368–376.

Leamer, E. E. (1983). Let's take the con out of econometrics. *American Economic Review*, 73(1):31–43.

Lee, D. S. (2008). Randomized experiments from non-random selection in u.s. house elections. *Journal of Econometrics*, 142:675–697.

Leeb, H. and Pötscher, B. M. (2005). Model selection and inference: Facts and fiction. *Econometric Theory*, 21:21–59.

Lockhart, R., Taylor, J., Tibshirani, R. J., and Tibshirani, R. (2014). A significance test for the lasso. *The Annals of Statistics*, 42(2):413–468.

Newey, W. K. (1997). Convergence rates and asymptotic normality for series estimators. *Journal of Econometrics*, 79(1):147–168.

Reinhart, C. M. and Rogoff, K. S. (2010). Growth in a time of debt. *American Economic Review: Papers and Proceedings*, 100:573–578.

Tian, X. and Taylor, J. (2017). Asymptotics of selective inference. *Scandinavian Journal of Statistics*, 44(2):480–499. 10.1111/sjos.12261.

van de Geer, S., Bühlmann, P., Ritov, Y., and Dezeure, R. (2014). On asymptotically optimal confidence regions and tests for high-dimensional models. *Annals of Statistics*, 42(3):1166–1202.

van de Geer, S. A. and Bühlmann, P. (2009). On the conditions used to prove oracle results for the lasso. *Electronic Journal of Statistics*, 3:1360–1392.

van der Vaart, A. (1998). *Asymptotic Statistics*. Cambridge University Press.

Xie, H. and Huang, J. (2009). Scad-penalized regression in high-dimensional partially linear models. *The Annals of Statistics*, 37(2):673–696.

Yao, Y.-C. (1988). Estimating the number of change-points via schwarz' criterion. *Statistics and Probability Letters*, 6(3):181–189.

Zhao, P. and Yu, B. (2006). On model selection consistency of lasso. *Journal of Machine Learning Research*, 7:2541–2563.

Appendix-1

Appendices

A Proofs

A-1 Proof of Lemma 1 [Invariance of LASSO under Projection]

Let $L = \|Y - S_m\beta - D\psi\|_2^2 + \lambda|\psi|_1$. Then,

$$\frac{\partial L}{\partial \beta} = -2S_m'Y + 2S_m'S_m\beta + 2S_m'D\psi. \tag{A1}$$

$\hat{\beta}$ must satisfy $\frac{\partial L}{\partial \beta} = 0$. Thus,

$$\hat{\beta} = (S_m'S_m)^{-1}S_m'(Y - D\psi) \tag{A2}$$

Substituting equation (A2) into equation (3), we have:

$$\begin{aligned} L &= \|(I - P_Z)(Y - D\psi)\|_2^2 + \lambda|\psi|_1 \\ &= \|M_mY - M_mD\psi\|_2^2 + \lambda|\psi|_1, \end{aligned}$$

– where $M_m = I_n - P_Z$. □

A-2 Proof of Theorem 1 [Design Matrices for Detecting Discontinuities]

For point discontinuities in part (a), the design matrix is exactly the identity matrix I_n and thus the irrepresentable condition holds.

Consider next jump discontinuities in part (b). For the general case, let $D_{A_0} = (D_{j_1}, D_{j_2}, \cdots, D_{j_s})$ such that $j_1 > j_2 > \cdots j_s$. Thus,

$$D_{A_0}'D_{A_0} = \begin{bmatrix} 1 & \sqrt{\frac{n-j_1}{n-j_2}} & \sqrt{\frac{n-j_1}{n-j_3}} & \cdots & \sqrt{\frac{n-j_1}{n-j_s}} \\ \sqrt{\frac{n-j_1}{n-j_2}} & 1 & \sqrt{\frac{n-j_2}{n-j_3}} & & \vdots \\ \sqrt{\frac{n-j_1}{n-j_3}} & \sqrt{\frac{n-j_2}{n-j_3}} & 1 & & \\ \vdots & & & \ddots & \\ \sqrt{\frac{n-j_1}{n-j_s}} & \cdots & & & 1 \end{bmatrix}.$$

Next, fix k and define the S-vector

$$\kappa \equiv D_k' D_{A_0} (D_{A_0}' D_{A_0})^{-1}.$$

With this notation,

$$\begin{aligned}
\frac{\min\{n-k, n-j_1\}}{\sqrt{(n-k)(n-j_1)}} &= \kappa_1 \sqrt{\frac{n-j_1}{n-j_1}} + \kappa_2 \sqrt{\frac{n-j_1}{n-j_2}} + \kappa_3 \sqrt{\frac{n-j_1}{n-j_3}} + \cdots + \kappa_s \sqrt{\frac{n-j_1}{n-j_s}} \\
\frac{\min\{n-k, n-j_2\}}{\sqrt{(n-k)(n-j_2)}} &= \kappa_1 \sqrt{\frac{n-j_1}{n-j_2}} + \kappa_2 \sqrt{\frac{n-j_2}{n-j_2}} + \kappa_3 \sqrt{\frac{n-j_2}{n-j_3}} \cdots + \kappa_s \sqrt{\frac{n-j_2}{n-j_s}} \\
&\cdots \\
\frac{\min\{n-k, n-j_s\}}{\sqrt{(n-k)(n-j_s)}} &= \kappa_1 \sqrt{\frac{n-j_1}{n-j_s}} + \kappa_2 \sqrt{\frac{n-j_2}{n-j_s}} + \kappa_3 \sqrt{\frac{n-j_3}{n-j_s}} \cdots + \kappa_s \sqrt{\frac{n-j_s}{n-j_s}}.
\end{aligned} \tag{A3}$$

From the above, compute the first line of (A3) minus the product of $\sqrt{\frac{n-j_1}{n-j_2}}$ and the second line to obtain

$$\frac{\min\{n-k, n-j_1\}}{\sqrt{(n-k)(n-j_1)}} - \sqrt{\frac{n-j_1}{n-j_2}} \frac{\min\{n-k, n-j_2\}}{\sqrt{(n-k)(n-j_2)}} = \kappa_1 \frac{j_1 - j_2}{n-j_2}.$$

Thus,

$$\kappa_1 = \begin{cases} \sqrt{\frac{n-k}{n-j_1}} & k \geq j_1 \\ \sqrt{\frac{n-j_1}{n-k}} \frac{k-j_2}{j_1-j_2} & j_1 > k \geq j_2 \\ 0 & j_2 > k \end{cases}.$$

Next we show that the irrepresentable condition holds in each of the three cases. When $k \geq j_1$, notice that $\kappa_2 = \kappa_3 = \cdots = \kappa_s = 0$ is a solution to the system. And since $(D_{A_0}' D_{A_0})$ is full rank, the solution is unique. Thus

$$\sup_{\|\tau_{A_0}\|_\infty \leq 1} |\kappa \tau_{A_0}| = \sqrt{\frac{n-k}{n-j_1}} < 1,$$

– and therefore the irrepresentable condition holds.

When $j_1 > k \geq j_2$, use the second line of (A3) minus the product of $\sqrt{\frac{n-j_2}{n-j_3}}$ and the third line to obtain

$$\frac{\min\{n-k, n-j_2\}}{\sqrt{(n-k)(n-j_2)}} - \sqrt{\frac{n-j_2}{n-j_3}}\frac{\min\{n-k, n-j_3\}}{\sqrt{(n-k)(n-j_3)}}$$
$$= \kappa_2 \frac{j_2 - j_3}{n-j_3} + \kappa_1 \left(\sqrt{\frac{n-j_1}{n-j_2}} - \frac{\sqrt{(n-j_1)(n-j_2)}}{n-j_3} \right).$$

Next, substitute κ_1 to get

$$\frac{\min\{n-k, n-j_2\}}{\sqrt{(n-k)(n-j_2)}} - \sqrt{\frac{n-j_2}{n-j_3}}\frac{\min\{n-k, n-j_3\}}{\sqrt{(n-k)(n-j_3)}}$$
$$= \kappa_2 \frac{j_2 - j_3}{n-j_3} + \kappa_1 \left(\sqrt{\frac{n-j_1}{n-j_2}} - \frac{\sqrt{(n-j_1)(n-j_2)}}{n-j_3} \right).$$

Thus,

$$\kappa_2 = \sqrt{\frac{n-j_2}{n-k}}\frac{j_1 - k}{j_1 - j_2}.$$

Notice that $\kappa_3 = \kappa_4 = \cdots = \kappa_s = 0$ is a solution to the system.

$$\sup_{\|\tau_{A_0}\|_\infty \le 1} |\kappa \tau_{A_0}| = \sqrt{\frac{n-j_1}{n-k}}\frac{k - j_2}{j_1 - j_2} + \sqrt{\frac{n-j_2}{n-k}}\frac{j_1 - k}{j_1 - j_2} < 1,$$

– and therefore the irrepresentable condition holds.

Finally, when $j_2 > k$, since $\kappa_1 = 0$, we can rewrite the system as:

$$\frac{\min\{n-k, n-j_2\}}{\sqrt{(n-k)(n-j_2)}} = \kappa_2\sqrt{\frac{n-j_2}{n-j_2}} + \kappa_3\sqrt{\frac{n-j_2}{n-j_3}} \cdots + \kappa_s\sqrt{\frac{n-j_2}{n-j_s}}$$
$$\frac{\min\{n-k, n-j_3\}}{\sqrt{(n-k)(n-j_3)}} = \kappa_2\sqrt{\frac{n-j_2}{n-j_3}} + \kappa_3\sqrt{\frac{n-j_3}{n-j_3}} \cdots + \kappa_s\sqrt{\frac{n-j_3}{n-j_s}}$$
$$\cdots$$
$$\frac{\min\{n-k, n-j_s\}}{\sqrt{(n-k)(n-j_s)}} = \kappa_2\sqrt{\frac{n-j_2}{n-j_s}} + \kappa_3\sqrt{\frac{n-j_3}{n-j_s}} \cdots + \kappa_s\sqrt{\frac{n-j_s}{n-j_s}},$$

–and we are back to the initial system with $s-1$ equations. By induction, we have $\sup_{\|\tau_{A_0}\|_\infty \leq 1} |\kappa\tau_{A_0}|$, thus the irrepresentable condition holds.

Finally, for case (c), we prove by induction. First we show that the design matrix for point and jump discontinuities satisfies the irreprenstable condition, then we assume it holds till the $K-1$th discontinuities and prove for the Kth discontinuity

when there is only one discontinuity but its type (kink or jump) is unknown, the design matrix D is a combination of both point and jump dummies :

$$D_k^{00} = (0, 0, 0, \cdots, 1, \cdots, 0)',$$

$$D_k^{0} = \left(0, 0, 0, \cdots, \frac{1}{\sqrt{n-k}}, \frac{1}{\sqrt{n-k}}, \cdots, \frac{1}{\sqrt{n-k}}\right)'$$

Since only one break exists, $D'_{A_0} D_{A_0} = 1$. First, assume the break is a jump:

$$D_j^{0'} D_{A_0} = \frac{\min\{(n-k),(n-j)\}}{\sqrt{(n-k)(n-j)}} < 1, \text{ and}$$
$$D_j^{00'} D_{A_0} = \frac{1}{\sqrt{n-k}} < 1.$$

Next, instead assume the break is a point. Then,

$$D_j^{0'} D_{A_0} = \frac{1}{\sqrt{n-j}} < 1, \text{ and}$$
$$D_j^{00'} D_{A_0} = 0 < 1.$$

By Cauchy-Schwarz, a useful proposition for the following proof is:

$$\left(\sum_{i=\max\{k+1,j+1\}}^{n} (x_{(i)} - x_{(k)})^L (x_{(i)} - x_{(j)})^K\right)^2 < \left(\sum_{i=\max\{k+1,j+1\}}^{n} (x_{(i)} - x_{(k)})^{2L}\right)\left(\sum_{i=\max\{k+1,j+1\}}^{n} (x_{(i)} - x_{(j)})^{2K}\right)$$
$$\leq \left(\sum_{i=j+1}^{n} (x_{(i)} - x_{(k)})^{2L}\right)\left(\sum_{i=k+1}^{n} (x_{(i)} - x_{(j)})^{2K}\right).$$

The proof proceeds by induction. Assume the irrepresentable condition holds when there is only one discontinuity up to the $(K-1)$th order.

First assume the break is a Lth order where $L \leq (K-1)$:

$$D_j^{K'} D_{A_0} = \frac{\sum_{i=\max\{k+1,j+1\}}^{n} (x_{(i)} - x_{(k)})^L (x_{(i)} - x_{(j)})^K}{\phi_k^L \phi_j^K} < 1.$$

Next assume the break is Kth order, for any $L \leq (K-1)$:

$$D_j^{L'} D_{A_0} = \frac{\sum_{i=\max\{k+1,j+1\}}^{n} (x_{(i)} - x_{(k)})^K (x_{(i)} - x_{(j)})^L}{\phi_k^K \phi_j^L} < 1,$$

– and for $L = K$,

$$D_j^{K'} D_{A_0} = \frac{\sum_{i=\max\{k+1,j+1\}}^{n} (x_{(i)} - x_{(k)})^K (x_{(i)} - x_{(j)})^K}{\phi_k^K \phi_j^K} < 1.$$

□

A-3 Proof of Corollary 1: [Invariant of Irrepresentable Condition under Projection]

Under the projection of P_Z, the irrepresentable condition becomes:

$$\begin{aligned}
&\left| D_j' P_Z D_{A_0} (D_{A_0}' P_Z D_{A_0})^{-1} \tau_{A_0} \right| \\
&= \left| trace \left(P_Z D_{A_0} (D_{A_0} P_Z D_{A_0}')^{-1} \tau_{A_0} D_j' \right) \right| \\
&= \left| trace \left((D_{A_0}' D_{A_0})^{-1} D_{A_0} D_{A_0}' P_Z D_{A_0} (D_{A_0} P_Z D_{A_0}')^{-1} \tau_{A_0} D_j' \right) \right| \\
&= \left| trace \left((D_{A_0}' D_{A_0})^{-1} D_{A_0} \tau_{A_0} D_j' \right) \right| \\
&= \left| trace \left(\tau_{A_0} D_j' (D_{A_0}' D_{A_0})^{-1} D_{A_0} \right) \right| \\
&= \left| trace \left(\tau_{A_0} D_j' (D_{A_0}' D_{A_0})^{-1} D_{A_0} D_{A_0}' D_{A_0} (D_{A_0}' D_{A_0})^{-1} \right) \right| \\
&= \left| trace \left(\tau_{A_0} D_j' D_{A_0} (D_{A_0}' D_{A_0})^{-1} \right) \right| \\
&= \left| D_j' D_{A_0} (D_{A_0}' D_{A_0})^{-1} \tau_{A_0} \right| < 1.
\end{aligned}$$

□

A-4 Proof of Corollary 2 [Kink Violation]

Consider the two-kink case with, WLOG, $k_1 > k_2$, so that

$$D^{1'}_{A_0} D^{1}_{A_0} = \begin{bmatrix} 1 & \frac{\sum_{i=k_1+1}^{n}(x_{(i)}-x_{(k_1)})(x_{(i)}-x_{(k_2)})}{\phi_{k_1}\phi_{k_2}} \\ \frac{\sum_{i=k_1+1}^{n}(x_{(i)}-x_{(k_1)})(x_{(i)}-x_{(k_2)})}{\phi_{k_1}\phi_{k_2}} & 1 \end{bmatrix},$$

– and,

$$(D^{1'}_{A_0} D^{1}_{A_0})^{-1} = \frac{1}{W} \begin{bmatrix} 1 & -\frac{\sum_{i=k_1+1}^{n}(x_{(i)}-x_{(k_1)})(x_{(i)}-x_{(k_2)})}{\phi_{k_1}\phi_{k_2}} \\ -\frac{\sum_{i=k_1+1}^{n}(x_{(i)}-x_{(k_1)})(x_{(i)}-x_{(k_2)})}{\phi_{k_1}\phi_{k_2}} & 1 \end{bmatrix},$$

–where $W = 1 - \left(\frac{\sum_{i=k_1+1}^{n}(x_{(i)}-x_{(k_1)})(x_{(i)}-x_{(k_2)})}{\phi_{k_1}\phi_{k_2}} \right)^2$.

Now pick j such that $k_1 > j > k_2$, so

$$D^{1'}_{j} D^{1}_{A_0} = \begin{bmatrix} \frac{\sum_{i=k_1+1}^{n}(x_{(i)}-x_{(k_1)})(x_{(i)}-x_{(j)})}{\phi_j\phi_{k_1}} & \frac{\sum_{i=j+1}^{n}(x_{(i)}-x_{(k_2)})(x_{(i)}-x_{(j)})}{\phi_j\phi_{k_2}} \end{bmatrix}.$$

Then,

$$\begin{aligned} & D^{1'}_{j} D^{1}_{A_0} (D^{1'}_{A_0} D^{1}_{A_0})^{-1} [1,1]' \\ & = \frac{1}{D} \Big(\frac{\sum_{i=k_1+1}^{n}(x_{(i)}-x_{(k_1)})(x_{(i)}-x_{(j)})}{\phi_j\phi_{k_1}} \\ & - \frac{\sum_{i=k_1+1}^{n}(x_{(i)}-x_{(k_1)})(x_{(i)}-x_{(k_2)})}{\phi_{k_1}\phi_{k_2}} \frac{\sum_{i=j+1}^{n}(x_{(i)}-x_{(k_2)})(x_{(i)}-x_{(j)})}{\phi_j\phi_{k_2}} \\ & - \frac{\sum_{i=k_1+1}^{n}(x_{(i)}-x_{(k_1)})(x_{(i)}-x_{(k_2)})}{\phi_{k_1}\phi_{k_2}} \frac{\sum_{i=k_1+1}^{n}(x_{(i)}-x_{(k_1)})(x_{(i)}-x_{(j)})}{\phi_j\phi_{k_1}} \\ & + \frac{\sum_{i=j+1}^{n}(x_{(i)}-x_{(k_2)})(x_{(i)}-x_{(j)})}{\phi_j\phi_{k_2}} \Big) \end{aligned}$$

$$= \frac{1}{W}\Big(\frac{\sum_{i=k_1+1}^{n}(x_{(i)}-x_{(k_1)})(x_{(i)}-x_{(j)})}{\phi_j\phi_{k_1}} + \frac{\sum_{i=j+1}^{n}(x_{(i)}-x_{(k_2)})(x_{(i)}-x_{(j)})}{\phi_j\phi_{k_2}}\Big)$$
$$\Big(1 - \frac{\sum_{i=k_1+1}^{n}(x_{(i)}-x_{(k_1)})(x_{(i)}-x_{(k_2)})}{\phi_{k_1}\phi_{k_2}}\Big)$$
$$= \Big(\frac{\sum_{i=k_1+1}^{n}(x_{(i)}-x_{(k_1)})(x_{(i)}-x_{(j)})}{\phi_j\phi_{k_1}} + \frac{\sum_{i=j+1}^{n}(x_{(i)}-x_{(k_2)})(x_{(i)}-x_{(j)})}{\phi_j\phi_{k_2}}\Big)$$
$$\Big(1 + \frac{\sum_{i=k_1+1}^{n}(x_{(i)}-x_{(k_1)})(x_{(i)}-x_{(k_2)})}{\phi_{k_1}\phi_{k_2}}\Big)^{-1}.$$

Define $f(j) = \frac{\sum_{i=k_1+1}^{n}(x_{(i)}-x_{(k_1)})(x_{(i)}-x_{(j)})}{\phi_j\phi_{k_1}} + \frac{\sum_{i=j+1}^{n}(x_{(i)}-x_{(k_2)})(x_{(i)}-x_{(j)})}{\phi_j\phi_{k_2}}$.

Notice that $f(k_1) = f(k_2) = 1 + \frac{\sum_{i=k_1+1}^{n}(x_{(i)}-x_{(k_1)})(x_{(i)}-x_{(k_2)})}{\phi_{k_1}\phi_{k_2}}$ and f is concave. Thus $D_j^{1\prime} D_{A_0}(D'_{A_0}D_{A_0})^{-1}[1,1]' > 1$ and the irrepresentable condition fails. □

A-5 Proof of Corollary 3 [Irrepresentable Condition under Partition]

Note that $\Pi_{A_0} = diag(\Pi^1_{A_0}, \Pi^2_{A_0}, \cdots, \Pi^s_{A_0})$ is block diagonal, so that

$$(\Pi'_{A_0}\Pi_{A_0})^{-1} = diag((\Pi^{1\prime}_{A_0}\Pi^1_{A_0})^{-1}, (\Pi^{2\prime}_{A_0}\Pi^2_{A_0})^{-1}, \cdots, (\Pi^{s\prime}_{A_0}\Pi^s_{A_0})^{-1}).$$

For the jth column of Π, the non-zero entries are Π^i_j, i.e. the jth column in block i, thus

$$\Pi'_j\Pi_{A_0}(\Pi'_{A_0}\Pi_{A_0})^{-1} = \Pi^i_j\Pi^i_{A_0}(\Pi^{i\prime}_{A_0}\Pi^i_{A_0})^{-1},$$

–and so,

$$\max_{j\notin A_0}\sup_{\|\tau_{A_0}\|_\infty\le 1}\left|\Pi'_j\Pi_{A_0}(\Pi'_{A_0}\Pi_{A_0})^{-1}\tau_{A_0}\right| = \max_{i\in\{1,2,\cdots,s\}}\max_{j\notin A_0}\sup_{\|\tau_{A_0}\|_\infty\le 1}\left|\Pi^i_j\Pi^i_{A_0}(\Pi^{i\prime}_{A_0}\Pi^i_{A_0})^{-1}\right| < 1.$$

□

A-6 Proof of Theorem 2 [Covariance Test under Measurement Error]

Let $Y = S_m\beta + r_m + D_{A_0}\psi_{A_0} + \epsilon$ and $M_m = (I - S_m(S_m'S_m)^{-1}S_m')$, then

$$M_m Y = M_m D_{A_0}\psi_{A_0} + r_m + M_m\epsilon.$$

Recall that r_m is the measurement error in the spline approximation. We can plug this into the definition of the test statistic T_k in (5), yielding

$$\begin{aligned}
T_k &= \left(Y'M_m D\psi(\lambda_{k+1}) - Y'M_m D_{A_k}\psi_{A_k}(\lambda_{k+1})\right)/\sigma^2 \\
&= \left(\psi'_{A_0}D'_{A_0}M_m D\psi(\lambda_{k+1}) - \psi'_{A_0}D'_{A_0}M_m D_{A_k}\psi_{A_k}(\lambda_{k+1})\right)/\sigma^2 \\
&+ \left(r'_m\Big(D\psi(\lambda_{k+1}) - D_{A_k}\psi_{A_k}(\lambda_{k+1})\Big)\right)/\sigma^2 \\
&+ \left(\epsilon'M_m D\psi(\lambda_{k+1}) - \epsilon'M_m D_{A_k}\psi_{A_k}(\lambda_{k+1})\right)/\sigma^2.
\end{aligned}$$

Our strategy is to bound the approximation error, which appears in the second term of the expansion above. Define $P_{A_k} = M_m D_{A_k}(D'_{A_k}M_m D_{A_k})^{-1}D'_{A_k}M_m$ and $P^+_{A_k} = M_m D_{A_k}(D'_{A_k}M_m D_{A_k})^{-1}$. Consider the term

$$\begin{aligned}
\left(r'_m\Big(D\psi(\lambda_{k+1}) - D_{A_k}\psi_{A_k}(\lambda_{k+1})\Big)\right) &= r'_m(P_{A_{k+1}} - P_{A_k})y - \lambda_{k+1}\cdot r'_m(P^+_{A_{k+1}}s_{k+1} - P^+_{A_k}s_k) \\
&= (\lambda_k - \lambda_{k+1})\cdot r'_m(P^+_{A_{k+1}}s_{k+1} - P^+_{A_k}s_k) \\
&\le (\lambda_k - \lambda_{k+1})\cdot \|r_m\|_2\|P^+_{A_{k+1}}s_{k+1} - P^+_{A_k}s_k\|_2.
\end{aligned}$$

The last inequality follows from Cauchy-Schwarz. Notice that $|s_0|$ is a constant with respect to n in our application. From Lemma 1 and Lemma 2 in Tian and Taylor (2017), for some $\Delta \to 0$ as $n \to \infty$, when $\lambda_{k+1} \ge 4\sigma\sqrt{\log p}$ with probability $1 - \Delta$:

$$|s_k|_0 \le C s_0.$$

Thus,

$$\|P^+_{A_k}s_k\|_2^2 \le \|s_k\|_0^2\|P^+_{A_k}\|_\infty^2 \le C^2 s_0^2 \max_{i,j}|M_m D|_{ij}^2.$$

By normalization, $\max_{i,j}|M_m D|_{ij}$ is of order $O(n^{-1/2})$. Thus when $\varphi_m^2 = \mathbb{E}(r_{mi}^2) = o(1/(\log p))$,

$$\mathbb{E}\left(r'_m\Big(D\psi(\lambda_{k+1}) - D_{A_k}\psi_{A_k}(\lambda_{k+1})\Big)\right)^2 \le n\cdot\varphi_m^2\cdot(\lambda_k - \lambda_{k+1})^2\cdot\mathbb{E}\|P^+_{A_{k+1}}s_{k+1} - P^+_{A_k}s_k\|_2^2 = o_p(1).$$

Now consider a new process,

$$Y^* = M_m D_{A_0} \psi_{A_0} + \epsilon.$$

Define the test statistic

$$\begin{aligned} T_k^* &= \left(Y^{*'} M_m D \psi(\lambda_{k+1}) - Y^{*'} M_m D_{A_k} \psi_{A_k}(\lambda_{k+1}) \right) / \sigma^2 \\ &= \left(\psi'_{A_0} D'_{A_0} M_m D \psi(\lambda_{k+1}) - \psi'_{A_0} D'_{A_0} M_m D_{A_k} \psi_{A_k}(\lambda_{k+1}) \right) / \sigma^2 \\ &+ \left(\epsilon' M_m D \psi(\lambda_{k+1}) - \epsilon' M_m D_{A_k} \psi_{A_k}(\lambda_{k+1}) \right) / \sigma^2. \end{aligned}$$

Thus by Theorem 2 in Lockhart et al. (2014),

$$\lim_{n \to \infty} \mathbb{P}(T_k > t) = \lim_{n \to \infty} \mathbb{P}(T_k^* > t) \leq e^{-t}.$$

□

A-7 Proof of Theorem 3 [Convergence Rate for IMSE]

First, a definition: we say that the *compatibility condition* is met for the set A_0 if, for some $\omega_m > 0$ (independent of n), and for all ψ satisfying $\|\psi_{A_0^c}\|_1 \leq 3\|\psi_{A_0}\|_1$, it holds that

$$\|\psi_{A_0}\|_1^2 \leq (\psi' \hat{\Sigma}_m \psi) s_0 / \omega_m^2, \tag{A4}$$

– where s_0 is the number of elements in A_0 and $\hat{\Sigma}_m = (M_m D_n)' M_m D_n$, which is the sample covariance matrix. Define ω_m to be the compatibility constant.

Here we apply Lemma 1 in Xie and Huang (2009) and show that the compatibility constant for $\hat{\Sigma}_m$ is asymptotically transformation invariant for spline and polynomial basis as $n \to \infty$, i.e.

$$\frac{1}{n} \|\hat{\Sigma}_m - D'_n D_n\|_\infty \to 0 \qquad \text{in probability.}$$

Thus

$$\left| \frac{\psi' \hat{\Sigma}_m \psi}{\|\psi_{A_0}\|_1^2} - \frac{\psi' D_n' D_n \psi}{\|\psi_{A_0}\|_1^2} \right| \to 0 \qquad \text{in probability.}$$

By Theorem 1, the irrepresentable condition holds for the design matrix $M_n' D_n$. Since the irrepresentable condition implies compatibility condition as shown in Theorem 9.1 in van de Geer and Bühlmann (2009), there exits some $\omega_0 > 0$ such that

$$\frac{(\psi' D_n' D_n \psi)}{\|\psi_{A_0}\|_1^2} \geq s_0/\omega_0^2.$$

As a result, there exist an absolute constant $\omega > 0$ such that for some $m_0 > 0$,

$$\frac{(\psi' \hat{\Sigma}_m \psi)}{\|\psi_{A_0}\|_1^2} \geq s_0/\omega^2.$$

The compatibility assumptions leads to the following results as shown in Lemma 6.1 in Bühlmann and van de Geer (2011):

$$\|M_m D_n(\hat{\psi} - \psi)\|_2^2/n \leq 4\lambda^2 s_0/\omega_m^2. \tag{A5}$$

$$\|(\hat{\psi} - \psi)\|_2^2 \leq \|(\hat{\psi} - \psi)\|_1^2 \leq 16\lambda^2 s_0^2/\omega_m^4 \tag{A6}$$

Let $(\hat{\beta}, \hat{\psi})$ be estimates from the lasso as in (3). Notice that

$$\begin{aligned}
\hat{\beta} &= (S_m' S_m)^{-1} S_m'(Y_n - D_n\psi) = (S_m' S_m)^{-1} S_m' \left(g(X) + \sum_{k=1}^{s_0} d_k(X)\psi_k + \epsilon - D_n\hat{\psi} \right) \\
&= (S_m' S_m)^{-1} S_m' \left(g(X) + \epsilon \right) + (S_m' S_m)^{-1} S_m' \left(\sum_{k=1}^{s_0} d_k(X)\psi_k - D_n\hat{\psi} \right) \\
&= \ddot{\beta}_m - (S_m' S_m)^{-1} S_m' \left(D_n(\hat{\psi} - \psi) \right),
\end{aligned}$$

– where $\ddot{\beta} = (S_m' S_m)^{-1} S_m' \left(g(X) + \epsilon \right)$ are coefficients for a standard SEIVE regression on $g(X) + \epsilon$.

$$
\begin{aligned}
IMSE_n(m) &= \int \mathbb{E}\left(\hat{g}_m(x) - g(x) + D_n(x)'\hat{\psi} - \sum_{k=1}^{s_0} d_k(x)\psi_k\right)^2 f(x)dx \\
&= \int \mathbb{E}\left(\hat{g}_m(x) - g(x) + D_n(x)'(\hat{\psi} - \psi)\right)^2 f(x)dx \\
&= \int \mathbb{E}\left(S_m(x)'(\ddot{\beta}_m - \beta_m) - r_m(x) - S_m(x)(S_m'S_m)^{-1}S_m'D_n(\hat{\psi} - \psi) + D(x)'(\hat{\psi} - \psi)\right)^2 f(x)dx \\
&= \underbrace{\int \mathbb{E}(S_m(x)(\ddot{\beta}_m - \beta_m) - r_m(x))^2 f(x)dx}_{(\equiv B1)} \\
&+ \underbrace{\int \mathbb{E}\left(D_n(x)'(\hat{\psi} - \psi) - S_m(x)(S_m'S_m)^{-1}S_m'D_n(\hat{\psi} - \psi)\right)^2 f(x)dx}_{(\equiv B2)} \\
&+ 2\underbrace{\int \mathbb{E}\Big(S_m(x)(\tilde{\beta}_m - \beta_m) - r_m(x)\Big)\left(D_n(x)'(\hat{\psi} - \psi) - S_m(x)(S_m'S_m)^{-1}S_m'D_n(\hat{\psi} - \psi)\right) f(x)dx}_{(\equiv B3)}.
\end{aligned}
$$

First we look at (B3):

$$
\begin{aligned}
&2\int \mathbb{E}\Big(S_m(x)(\tilde{\beta}_m - \beta_m) - r_m(x)\Big)\left(D_n(x)'(\hat{\psi} - \psi) - S_m(x)(S_m'S_m)^{-1}S_m'D_n(\hat{\psi} - \psi)\right) f(x)dx \\
&\leq \int \mathbb{E}\Big(S_m(x)(\tilde{\beta}_m - \beta_m) - r_m(x)\Big)^2 + \mathbb{E}\left(D_n(x)'(\hat{\psi} - \psi) - S_m(x)(S_m'S_m)^{-1}S_m'D_n(\hat{\psi} - \psi)\right)^2 f(x)dx \\
&= B1 + B2
\end{aligned}
$$

Notice that part (B1) is the standard SEIVE term and as shown in theorem 1 in Hansen (2014a),

$$
(B1) = \varphi_m^2 + \sigma^2 \frac{K_m}{n}.
$$

Now consider part (B2).

$$
\begin{aligned}
(B2) &= \int \left(D_n(x)'(\hat{\psi} - \psi) - S_m(x)'(S_m'S_m)^{-1}S_m'D_n(\hat{\psi} - \psi)\right)^2 f(x)dx \\
&= (\hat{\psi} - \psi)' \underbrace{\left(\int (D_n(x)' - S_m(x)'(S_m'S_m)^{-1}S_m'D_n)'(D_n(x)' - S_m(x)'(S_m'S_m)^{-1}S_m'D_n) f(x)dx\right)}_{(\equiv B4)} (\hat{\psi} - \psi).
\end{aligned}
$$

Define $\mathbb{E}(S_m(x)'S_m(x)) = Q_m$. Define

$$DD_n = \int D_n(x)D_n(x)'f(x)dx = \begin{bmatrix} \frac{\mathbb{P}(x>x_{(1)})}{n-1} & \frac{\mathbb{P}(x>x_{(2)})}{\sqrt{n-1}\sqrt{n-2}} & \cdots & \frac{\mathbb{P}(x>x_{(n-1)})}{\sqrt{n-1}\sqrt{1}} \\ \frac{\mathbb{P}(x>x_{(2)})}{\sqrt{n-1}\sqrt{n-2}} & \frac{\mathbb{P}(x>x_{(2)})}{n-2} & \cdots & \frac{\mathbb{P}(x>x_{(n-1)})}{\sqrt{n-2}\sqrt{1}} \\ \vdots & & & \vdots \\ \frac{\mathbb{P}(x>x_{(n-1)})}{\sqrt{n-1}\sqrt{1}} & \frac{\mathbb{P}(x>x_{(n-1)})}{\sqrt{n-2}\sqrt{1}} & \cdots & \frac{\mathbb{P}(x>x_{(n-1)})}{1} \end{bmatrix}.$$

Notice that

$$D_n'D_n = \begin{bmatrix} \frac{\mathbf{1}(x_1>x_{(1)})}{\sqrt{n-1}} & \frac{\mathbf{1}(x_2>x_{(1)})}{\sqrt{n-1}} & \cdots & \frac{\mathbf{1}(x_n>x_{(1)})}{\sqrt{n-1}} \\ \frac{\mathbf{1}(x_1>x_{(2)})}{\sqrt{n-2}} & \frac{\mathbf{1}(x_2>x_{(2)})}{\sqrt{n-2}} & \cdots & \frac{\mathbf{1}(x_n>x_{(2)})}{\sqrt{n-2}} \\ \vdots & & & \vdots \\ \frac{\mathbf{1}(x_1>x_{(n-1)})}{\sqrt{1}} & \frac{\mathbf{1}(x_2>x_{(n-1)})}{\sqrt{1}} & \cdots & \frac{\mathbf{1}(x_n>x_{(n-1)})}{\sqrt{1}} \end{bmatrix} \begin{bmatrix} \frac{\mathbf{1}(x_1>x_{(1)})}{\sqrt{n-1}} & \frac{\mathbf{1}(x_1>x_{(2)})}{\sqrt{n-2}} & \cdots & \frac{\mathbf{1}(x_1>x_{(n-1)})}{\sqrt{1}} \\ \frac{\mathbf{1}(x_2>x_{(1)})}{\sqrt{n-1}} & \frac{\mathbf{1}(x_2>x_{(2)})}{\sqrt{n-2}} & \cdots & \frac{\mathbf{1}(x_2>x_{(n-1)})}{\sqrt{1}} \\ \vdots & & & \vdots \\ \frac{\mathbf{1}(x_n>x_{(1)})}{\sqrt{n-1}} & \frac{\mathbf{1}(x_n>x_{(2)})}{\sqrt{n-2}} & \cdots & \frac{\mathbf{1}(x_n>x_{(n-1)})}{\sqrt{1}} \end{bmatrix}$$

$$= \begin{bmatrix} 1 & \sqrt{\frac{n-2}{n-1}} & \cdots & \sqrt{\frac{1}{n-1}} \\ \sqrt{\frac{n-2}{n-1}} & 1 & \cdots & \sqrt{\frac{1}{n-2}} \\ \vdots & & & \vdots \\ \sqrt{\frac{1}{n-1}} & \sqrt{\frac{1}{n-2}} & \cdots & 1 \end{bmatrix},$$

– and

$$DD_n - \frac{1}{n}D_n'D_n = \begin{bmatrix} \frac{\mathbb{P}(x>x_{(1)})-\frac{n-1}{n}}{n-1} & \frac{\mathbb{P}(x>x_{(2)})-\frac{n-2}{n}}{\sqrt{n-1}\sqrt{n-2}} & \cdots & \frac{\mathbb{P}(x>x_{(n-1)})-\frac{1}{n}}{\sqrt{n-1}\sqrt{1}} \\ \frac{\mathbb{P}(x>x_{(2)})-\frac{n-2}{n}}{\sqrt{n-1}\sqrt{n-2}} & \frac{\mathbb{P}(x>x_{(2)})-\frac{n-2}{n}}{n-2} & \cdots & \frac{\mathbb{P}(x>x_{(n-1)})-\frac{1}{n}}{\sqrt{n-2}\sqrt{1}} \\ \vdots & & & \vdots \\ \frac{\mathbb{P}(x>x_{(n-1)})-\frac{1}{n}}{\sqrt{n-1}\sqrt{1}} & \frac{\mathbb{P}(x>x_{(n-1)})-\frac{1}{n}}{\sqrt{n-2}\sqrt{1}} & \cdots & \frac{\mathbb{P}(x>x_{(n-1)})-\frac{1}{n}}{1} \end{bmatrix}.$$

Let $\hat{\mathbb{P}}$ be the empirical distribution. By Glivenko-Cantelli, $\hat{\mathbb{P}}(x > x_{(i)}) = \frac{n-i}{n} + o_p(1/\sqrt{n})$, and therefore

$$\|DD_n - \frac{1}{n}D_n'D_n\|_\infty \to 0 \quad \text{with probability 1.}$$

Second, Define

$$DS_n = \int D_n(x) S_m(x)' f(x) dx = \begin{bmatrix} \frac{\int \mathbf{1}(x > x_{(1)}) S_m(x)' f(x) dx}{\sqrt{n-1}} \\ \frac{\int \mathbf{1}(x > x_{(2)}) S_m(x)' f(x) dx}{\sqrt{n-2}} \\ \vdots \\ \frac{\int \mathbf{1}(x > x_{(n-1)}) S_m(x)' f(x) dx}{\sqrt{1}} \end{bmatrix} = \begin{bmatrix} \frac{\mathbb{E}(S_m(x)' | x > x_{(1)})}{\sqrt{n-1}} \\ \frac{\mathbb{E}(S_m(x)' | x > x_{(2)})}{\sqrt{n-2}} \\ \vdots \\ \frac{\mathbb{E}(S_m(x)' | x > x_{(n-1)})}{\sqrt{1}} \end{bmatrix}.$$

$$D_n' S_m = \begin{bmatrix} \frac{\mathbf{1}(x_1 > x_{(1)})}{\sqrt{n-1}} & \frac{\mathbf{1}(x_2 > x_{(1)})}{\sqrt{n-1}} & \cdots & \frac{\mathbf{1}(x_n > x_{(1)})}{\sqrt{n-1}} \\ \frac{\mathbf{1}(x_1 > x_{(2)})}{\sqrt{n-2}} & \frac{\mathbf{1}(x_2 > x_{(2)})}{\sqrt{n-2}} & \cdots & \frac{\mathbf{1}(x_n > x_{(2)})}{\sqrt{n-2}} \\ \vdots & & & \vdots \\ \frac{\mathbf{1}(x_1 > x_{(n-1)})}{\sqrt{1}} & \frac{\mathbf{1}(x_2 > x_{(n-1)})}{\sqrt{1}} & \cdots & \frac{\mathbf{1}(x_n > x_{(n-1)})}{\sqrt{1}} \end{bmatrix} \begin{bmatrix} S_m(x_1)' \\ S_m(x_2)' \\ \vdots \\ S_m(x_n)' \end{bmatrix}$$

$$= \int D_n(x) S_m(x)' f(x) dx = \begin{bmatrix} \frac{\sum_{i=1}^n \mathbf{1}(x_i > x_{(1)}) S_m(x_i)'}{\sqrt{n-1}} \\ \frac{\sum_{i=1}^n \mathbf{1}(x_i > x_{(2)}) S_m(x_i)'}{\sqrt{n-2}} \\ \vdots \\ \frac{\sum_{i=1}^n \mathbf{1}(x_i > x_{(n-1)}) S_m(x_i)'}{\sqrt{1}} \end{bmatrix}.$$

Notice first that

$$\frac{1}{n} \sum_{i=1}^n \mathbf{1}(x_i > x_{(1)}) S_{mj}(x_i) \to_p \mathbb{E}(S_{mj}(x) | x > x_{(1)}).$$

And notice $\|\mathbf{1}(x_i > x_{(1)}) S_{mj}(x_i)\| \leq \xi_0(m)$. By P-Glivenko-Cantelli, (see van der Vaart (1998)),

$$\sup_j \left| \mathbb{P}_n \mathbf{1}(x_i > x_{(1)}) S_{mj}(x_i) - \mathbb{P} \mathbf{1}(x_i > x_{(1)}) S_{mj}(x_i) \right| \to 0 \quad \text{a.s.}$$

And finally,

$$Q_n = \int S_m(x) S_m(x)' f(x) dx = \left[\int S_{mi}(x) S_{mj}(x) f(x) dx \right]_{i=1:n, j=1:n} = I_{n \times n},$$

– and,

$$S_{mi}(x_k)S_{mj}(x_k) = \begin{cases} 1 & i = j \\ 0 & \text{otherwise.} \end{cases}$$

We can further decompose (B4) into the empirical processes below:

$$\begin{aligned}
(B4) &= \Big(\int (D_n(x)' - S_m(x)'(S_m'S_m)^{-1}S_m'D_n)'(D_n(x)' - S_m(x)'(S_m'S_m)^{-1}S_m'D_n)f(x)dx \Big) \\
&= \Big(\int (D_n(x)D_n(x)' - D_n(x)S_m(x)'(S_m'S_m)^{-1}S_m'D_n \\
&- D_nS_m(S_m'S_m)^{-1}S_m(x)D_n(x)' + D_n'S_m(S_m'S_m)^{-1}S_m(x)S_m(x)'(S_m'S_m)^{-1}S_m'D_n)f(x)dx \Big) \\
&= \Big(\int D_n(x)D_n(x)'f(x)dx \Big) - \Big(\int D_n(x)S_m(x)'f(x)dx(S_m'S_m)^{-1}S_m'D_n \Big) \\
&- \Big(D_n'S_m(S_m'S_m)^{-1} \int S_m(x)D_n(x)'f(x)dx \Big) + \Big(D_n'S_m(S_m'S_m)^{-1} \int S_m(x)S_m(x)'f(x)dx(S_m'S_m)^{-1}S_m'D_n \Big) \\
&= DD_n - DS_n(S_m'S_m)^{-1}S_m'D_n - D_n'S_m(S_m'S_m)^{-1}(DS_n') + \Big(D_n'S_m(S_m'S_m)^{-1}Q_n(S_m'S_m)^{-1}S_m'D_n \Big) \\
&= \underbrace{\frac{1}{n}D_n'(I - S_m(S_m'S_m)^{-1}S_m')D_n}_{(\equiv B5)} + \underbrace{\Big(DD_n - \frac{1}{n}D_n'D_n \Big)}_{(\equiv B6)} - \underbrace{\Big(DS_n - \frac{1}{n}D_nS_m \Big)(S_m'S_m)^{-1}S_m'D_n}_{(\equiv B7)} \\
&- \underbrace{D_n'S_m(S_m'S_m)^{-1}\Big(DS_n - \frac{1}{n}D_nS_m \Big)'}_{(\equiv B8)} + \underbrace{\Big(D_n'S_m(S_m'S_m)^{-1}\Big(Q_n - \frac{1}{n}S_m'S_m \Big)(S_m'S_m)^{-1}S_m'D_n \Big)}_{(\equiv B9)}.
\end{aligned}$$

For part (B5),

$$(\hat{\psi} - \psi)'\frac{1}{n}D_n'(I - S_m(S_m'S_m)^{-1}S_m')D_n(\hat{\psi} - \psi) = \|M_mD_n(\hat{\psi} - \psi)\|_n^2 \leq 4\lambda^2 s_0/\omega_m^2 \leq 4\lambda^2 s_0/\omega^2.$$

The last inequality follows from equation (A5).

For part (B6),

$$\begin{aligned}
(\hat{\psi} - \psi)'\Big(DD_n - \frac{1}{n}D_n'D_n \Big)(\hat{\psi} - \psi) &\leq \Big\| DD_n - \frac{1}{n}D_n'D_n \Big\|_\infty \cdot \|(\hat{\psi} - \psi)\|_2^2 \\
&\leq o_p(1/\sqrt{n})16\lambda^2 s_0^2/\omega_m^4 \leq o_p(1/\sqrt{n}).
\end{aligned}$$

Similarly, (B7) and (B8) can be shown to be order $o_p(1/\sqrt{n})$. And (B9) is 0.

As a result,

$$IMSE_n(m) \leq 2\varphi_m^2 + 2\sigma^2 \frac{K_m}{n} + 4\lambda^2 s_0/\omega^2.$$

□

A-8 Proof of Theorem 4 [Distribution of Jump Detection]

Let $\{(y_i, x_i)\}_{i=1}^n$ be observed data. Assume $y_i = d_i^0 \psi_1 + \epsilon_i$, such that

$$d_i^0 = \begin{cases} 0 & \text{if } x_i \leq z \\ 1 & \text{if } x_i > z. \end{cases} \tag{A7}$$

Let $\{(x_{(1)}, x_{(2)}, \cdots, x_{(n)})\}$ be the ordered statistic of $\{x_i\}_{i=1}^n$. Let $k > 0$ be an integer such that

$$x_{(k)} \leq z < x_{(k+1)}.$$

Now define

$$D_i^{(j)} = \begin{cases} 0 & \text{if } x_i \leq x_{(j)} \\ \frac{1}{\sqrt{n-j}} & \text{if } x_i > x_{(j)}. \end{cases} \tag{A8}$$

We can rewrite $y_i = D_i^{(k)} \tilde{\psi}_1 + \epsilon_i$, where $\tilde{\psi}_1 = \sqrt{n-k}\psi_1$. Thus:

$$\begin{aligned} D^{(j)\prime} Y &= \sum_{i=1}^n D_i^{(j)} D_i^{(k)} \tilde{\psi}_1 + D_i^{(j)} \epsilon \\ &= \begin{cases} \frac{(n-k)\tilde{\psi}_1}{\sqrt{n-j}} + \frac{\sum_{i=j+1}^n \epsilon_i}{\sqrt{n-j}} & \text{if } j \leq k \\ \\ \frac{(n-j)\tilde{\psi}_1}{\sqrt{n-j}} + \frac{\sum_{i=j+1}^n \epsilon_i}{\sqrt{n-j}} & \text{if } j > k. \end{cases} \end{aligned} \tag{A9}$$

Define $V(j) = (D^{(j)'}Y)^2$, and $\hat{k} = \arg\max_j V(j)$

$$\begin{aligned} V(j) - V(k) &= (D^{(j)'}Y)^2 - (D^{(k)'}Y)^2 \\ &= Y'(D^{(j)}D^{(j)'} - D^{(k)}D^{(k)'})Y \\ &= \tilde{\psi}_1 D^{(k)'}\Big(D^{(j)}D^{(j)'} - D^{(k)}D^{(k)'}\Big)D^{(k)}\tilde{\psi}_1 + h(\epsilon, \tilde{\psi}_1) \\ &= \tilde{\psi}_1\Big(D^{(k)'}D^{(j)}D^{(j)'}D^{(k)} - 1\Big)\tilde{\psi}_1 + h(\epsilon, \tilde{\psi}_1). \end{aligned}$$

$$h(\epsilon, \tilde{\psi}_1) = 2\tilde{\psi}_1 D^{(k)'}(D^{(j)}D^{(j)'} - D^{(k)}D^{(k)'})\epsilon + \epsilon'(D^{(j)}D^{(j)'} - D^{(k)}D^{(k)'})\epsilon.$$

Define

$$\gamma(j) = \frac{\tilde{\psi}_1\Big(1 - D^{(k)'}D^{(j)}D^{(j)'}D^{(k)}\Big)\tilde{\psi}_1}{|j-k|}.$$

if $j > k$

$$\gamma(j) = \frac{\tilde{\psi}_1^2}{n-k} = \psi_1^2.$$

if $j < k$

$$\gamma(j) = \frac{\tilde{\psi}_1^2}{n-j} = \frac{n-k}{n-j}\psi_1^2.$$

By selection consistency of LASSO, for any fixed constant C, we have $P(|\hat{k} - k| > C) < \xi$:

$$
\begin{aligned}
P(|\hat{k}-k|>C) &= P(\sup_{|j-k|>C} V(j) \geq V(k)) \\
&\leq P(\sup_{|j-k|>C} |h(\epsilon,\tilde{\psi}_1)| \geq \inf_{|j-k|>C} |j-k|\gamma(j)) \\
&\leq P\left(\sup_{|j-k|>C} \left|\frac{h(\epsilon,\tilde{\psi}_1)}{j-k}\right| \geq \inf_{|j-k|>C} \gamma(j)\right) \\
&\leq P\left(\sup_{|j-k|>C} \left|\frac{h(\epsilon,\tilde{\psi}_1)}{j-k}\right| \geq \gamma_K\right),
\end{aligned}
$$

– where $\gamma_K = \psi_1^2 > 0$.

When $j > k$,

$$
\begin{aligned}
\epsilon'(D^{(j)}D^{(j)'})\epsilon - \epsilon'(D^{(k)}D^{(k)'})\epsilon &= \frac{1}{n-j}\Big(\sum_{l=j+1}^{n}\epsilon_l\Big)^2 - \frac{1}{n-k}\Big(\sum_{l=k+1}^{n}\epsilon_l\Big)^2 \\
&= \frac{j-k}{(n-j)(n-k)}\Big(\sum_{l=j+1}^{n}\epsilon_l\Big)^2 - \frac{2}{n-k}\Big(\sum_{l=k+1}^{j}\epsilon_l\Big)\Big(\sum_{l=j+1}^{n}\epsilon_l\Big) \\
&\quad - \frac{1}{n-k}\Big(\sum_{l=k+1}^{j}\epsilon_l\Big)^2 \\
&= (j-k)o(1).
\end{aligned}
$$

Similarly, one can show the same rate of convergence when $j < k$.

Next consider $2\tilde{\psi}_1 D^{(k)'}(D^{(j)}D^{(j)'} - D^{(k)}D^{(k)'})\epsilon$.

When $k > j$,

$$
\begin{aligned}
2\tilde{\psi}_1 D^{(k)'}D^{(j)}D^{(j)'}\epsilon - 2\tilde{\psi}_1 D^{(k)'}D^{(k)}D^{(k)'}\epsilon &= 2\psi_1\frac{n-k}{n-j}\sum_{l=j+1}^{n}\epsilon_l - 2\psi_1\sum_{l=k+1}^{n}\epsilon_l \\
&= 2\psi_1\frac{j-k}{n-j}\sum_{l=k+1}^{n}\epsilon_l + 2\psi_1\frac{n-k}{n-j}\sum_{l=j+1}^{k}\epsilon_l \\
&= (j-k)o(1).
\end{aligned}
$$

and for any ξ, we can find a C such that $P(|\hat{k} - k| > C) < \xi$.

For Asymptotic, consider $\psi_1 \to 0$ but $\sqrt{n}\psi_1 \to \infty$. In this case, $\gamma(j)$ can no longer be treated as $O_p(1)$ but $O_p(\psi_1^2)$. Thus, the convergence of $\hat{k}$ is

$$\hat{k} = k + O_p(\|\psi_1\|^{-2}).$$

Let

$$\hat{k} = k + v\rho_n^{-2}.$$

where $\rho_n = O_p(\|\psi_1\|)$ and v is a real number in a compact set. Define

$$K(B) = \{w : w = k + [v\rho_n^{-2}], |v| < B\}.$$

we derive the limiting process of $V(w) - V(k)$ for $w \in K(B)$ and then use the continuous mapping theorem for arg max to derive the asymptotic distribution. Recall that $\psi_1 = \tilde{\psi}_1/\sqrt{n-k}$

$$\begin{aligned} V(w) - V(k) = \tilde{\psi}_1\Big(D^{(k)\prime}D^{(w)}D^{(w)\prime}D^{(k)} - 1\Big)\tilde{\psi}_1 + 2\tilde{\psi}_1 D^{(k)\prime}(D^{(w)}D^{(w)\prime} - D^{(k)}D^{(k)\prime})\epsilon \\ + \epsilon'(D^{(w)}D^{(w)\prime} - D^{(k)}D^{(k)\prime})\epsilon. \end{aligned}$$

When $w > k$,

$$D^{(k)\prime}D^{(w)}D^{(w)\prime}D^{(k)} = \frac{n-w}{n-k}.$$

Thus,

$$\tilde{\psi}_1\Big(D^{(k)\prime}D^{(w)}D^{(w)\prime}D^{(k)} - 1\Big)\tilde{\psi}_1 = -\tilde{\psi}_1^2\Big(\frac{w-k}{n-k}\Big) = -\tilde{\psi}_1^2\Big(\frac{[v\rho_n^{-2}]}{n-k}\Big) = -v.$$

Next,

$$D^{(k)\prime}(D^{(w)}D^{(w)\prime})\epsilon = \frac{\sum_{l=w+1}^{n}\epsilon_l}{\sqrt{n-k}}, \text{ and}$$

$$D^{(k)'}(D^{(k)}D^{(k)'})\epsilon = \frac{\sum_{l=k+1}^{n}\epsilon_l}{\sqrt{n-k}}.$$

So

$$\tilde{\psi}_1 D^{(k)'}(D^{(w)}D^{(w)'})\epsilon - \tilde{\psi}_1 D^{(k)'}(D^{(k)}D^{(k)'})\epsilon = -\psi_1 \sum_{l=k+1}^{w}\epsilon_l = -\psi_1 \frac{v\rho_n^{-2}}{w-k}\sum_{l=k+1}^{w}\epsilon_l \to_d -W_1(v),$$

– where W_1 is a wiener process of degree v.

And finally,

$$\begin{aligned}
\epsilon'(D^{(w)}D^{(w)'} - D^{(k)}D^{(k)'})\epsilon &= \frac{1}{n-w}\Big(\sum_{l=w+1}^{n}\epsilon_l\Big)^2 - \frac{1}{n-k}\Big(\sum_{l=k+1}^{n}\epsilon_l\Big)^2 \\
&= \frac{w-k}{(n-w)(n-k)}\Big(\sum_{l=k+1}^{n}\epsilon_l\Big)^2 - \frac{1}{n-k}\Big(\sum_{l=k+1}^{w}\epsilon_l\Big)^2 - \frac{2}{n-k}\Big(\sum_{l=k+1}^{w}\epsilon_l\Big)\Big(\sum_{l=k+1}^{n}\epsilon_l\Big). \\
&= o_p(1)
\end{aligned}$$

When $w < k$

$$D^{(k)'}D^{(w)}D^{(w)'}D^{(k)} = \frac{n-k}{n-w}.$$

Thus,

$$\tilde{\psi}_1\Big(D^{(k)'}D^{(w)}D^{(w)'}D^{(k)} - 1\Big)\tilde{\psi}_1 = -\tilde{\psi}_1^2\Big(\frac{k-w}{n-w}\Big) = \tilde{\psi}_1^2\Big(\frac{[v\rho_n^{-2}]}{n-w}\Big) = \psi_1^2 v\rho_n^{-2}.$$

Next,

$$D^{(k)'}(D^{(w)}D^{(w)'})\epsilon = \frac{\sqrt{n-k}}{n-w}\sum_{l=w+1}^{n}\epsilon_l,$$

$$D^{(k)'}(D^{(k)}D^{(k)'})\epsilon = \frac{\sum_{l=k+1}^{n}\epsilon_l}{\sqrt{n-k}}.$$

So

$$\tilde{\psi}_1 D^{(k)\prime}(D^{(w)}D^{(w)\prime})\epsilon - \tilde{\psi}_1 D^{(k)\prime}(D^{(k)}D^{(k)\prime})\epsilon = \tilde{\psi}_1 \frac{\sqrt{n-k}}{n-w} \sum_{l=w+1}^{k} \epsilon_l + \tilde{\psi}_1 \frac{w-k}{n-w} \frac{\sum_{l=k+1}^{n} \epsilon_l}{\sqrt{n-k}}$$

$$= \psi_1 \frac{n-k}{n-w} \sum_{l=w+1}^{k} \epsilon_l + \sqrt{n-k}\psi_1 \frac{w-k}{n-w} \frac{\sum_{l=k+1}^{n} \epsilon_l}{\sqrt{n-k}} \rightarrow_d W_2(v),$$

– where W_2 is another wiener process of degree v.

Thus

$$V(k + [v\rho_n^{-2}]) - V(k) \rightarrow_d -|v| + 2W(|v|).$$

By the continuous mapping theorem:

$$\rho_n^2(\hat{k} - k) \rightarrow_d \arg\max_v(-|v| + 2W(|v|)). \tag{A10}$$

A-9 Proof of Theorem 5 [Distribution of Kink Detection]

Assume there is only one 1th-order discontinuity such that the data generating process is:

$$y_i = d_i^1 \psi_1 + \epsilon_i,$$

$$d_i^1 = \begin{cases} 0 & \text{if } x_i \leq z \\ (x_i - z) & \text{if } x_i > z. \end{cases}$$

Let $k > 0$ be an integer such that

$$x_{(k)} \leq z < x_{(k+1)}.$$

Define

$$D_i^{(j)} = \begin{cases} 0 & \text{if } x_i \leq x_{(j)} \\ (x_i - x_{(j)})/\phi_j & \text{if } x_i > x_{(j)}, \end{cases}$$

– where $\phi_j = \sqrt{\sum_{i=j+1}^{n}(x_{(i)} - x_{(j)})^2}$

Consider the correlation matrix $D^{(j)\prime}Y$

$$D^{(j)'}Y = \sum_{i=1}^{n} D_i^{(j)'}(x_i - z)\cdot \mathbf{1}(x_i > z)\psi_1 + D_i^{(j)'}\epsilon$$
$$= \begin{cases} \frac{\psi_1}{\phi_j}\sum_{i=k+1}^{n}(x_{(i)} - x_{(k)})(x_{(i)} - x_{(j)}) + \frac{1}{\phi_j}\sum_{i=j+1}^{n}(x_{(i)} - x_{(j)})\epsilon_i & \text{if } j \le k \\ \\ \frac{\psi_1}{\phi_j}\sum_{i=j+1}^{n}(x_{(i)} - x_{(k)})(x_{(i)} - x_{(j)}) + \frac{1}{\phi_j}\sum_{i=j+1}^{n}(x_{(i)} - x_{(j)})\epsilon_i & \text{if } j > k. \end{cases}$$

Define $V(j) = (D^{(j)'}Y)^2$, and $\hat{k} = \arg\max_j V(j)$.

$$V(j) - V(k) = (D^{(j)'}Y)^2 - (D^{(k)'}Y)^2$$
$$= \Big(\frac{1}{\phi_j}\sum_{i=j+1}^{n}(x_{(i)} - x_{(j)})\epsilon_i\Big)^2 - \Big(\frac{1}{\phi_k}\sum_{i=k+1}^{n}(x_{(i)} - x_{(k)})\epsilon_i\Big)^2 + \begin{cases} h_1(\epsilon,\psi_1) & \text{if } j \le k \\ h_2(\epsilon,\psi_1) & \text{if } j > k, \end{cases}$$

– where

$$h_1(\epsilon,\psi_1) = \underbrace{\frac{\psi_1^2}{\phi_j^2}\left(\sum_{i=k+1}^{n}(x_{(i)} - x_{(k)})(x_{(i)} - x_{(j)})\right)^2 - \frac{\psi_1^2}{\phi_k^2}\left(\sum_{i=k+1}^{n}(x_{(i)} - x_{(k)})(x_{(i)} - x_{(k)})\right)^2}_{(h_{11})}$$
$$\underbrace{+ 2\frac{\psi_1}{\phi_j^2}\Big(\sum_{i=j+1}^{n}(x_{(i)} - x_{(j)})\epsilon_i\Big)\Big(\sum_{i=k+1}^{n}(x_{(i)} - x_{(k)})(x_{(i)} - x_{(j)})\Big) - 2\psi_1\Big(\sum_{i=k+1}^{n}(x_{(i)} - x_{(k)})\epsilon_i\Big)}_{h_{12}}.$$

$$h_2(\epsilon,\psi_1) = \underbrace{\frac{\psi_1^2}{\phi_j^2}\left(\sum_{i=j+1}^{n}(x_{(i)} - x_{(k)})(x_{(i)} - x_{(j)})\right)^2 - \frac{\psi_1^2}{\phi_k^2}\left(\sum_{i=k+1}^{n}(x_{(i)} - x_{(k)})(x_{(i)} - x_{(k)})\right)^2}_{h_{21}}$$
$$\underbrace{+ 2\frac{\psi_1}{\phi_j^2}\Big(\sum_{i=j+1}^{n}(x_{(i)} - x_{(j)})\epsilon_i\Big)\Big(\sum_{i=j+1}^{n}(x_{(i)} - x_{(k)})(x_{(i)} - x_{(j)})\Big) - 2\psi_1\Big(\sum_{i=k+1}^{n}(x_{(i)} - x_{(k)})\epsilon_i\Big)}_{h_{22}}.$$

First consider the common terms in $V(j) - V(k)$ when $j \leq k$ and $j > k$. We show that it is of order $O_p(x_{(k)} - x_{(j)})$. This quantity will be shown of order smaller than h_{11}, h_{12}, h_{21} and h_{22}.

$$\Big(\frac{1}{\phi_j}\sum_{i=j+1}^{n}(x_{(i)}-x_{(j)})\epsilon_i\Big)^2 - \Big(\frac{1}{\phi_k}\sum_{i=k+1}^{n}(x_{(i)}-x_{(k)})\epsilon_i\Big)^2$$
$$= \Big(\frac{1}{\phi_j}\sum_{i=j+1}^{n}(x_{(i)}-x_{(j)})\epsilon_i - \frac{1}{\phi_k}\sum_{i=k+1}^{n}(x_{(i)}-x_{(k)})\epsilon_i\Big)\Big(\frac{1}{\phi_j}\sum_{i=j+1}^{n}(x_{(i)}-x_{(j)})\epsilon_i + \frac{1}{\phi_k}\sum_{i=k+1}^{n}(x_{(i)}-x_{(k)})\epsilon_i\Big)$$

$$\begin{aligned}\frac{1}{\phi_j}\sum_{i=j+1}^{n}(x_{(i)}-x_{(j)})\epsilon_i &= \frac{1}{\phi_j}\sum_{i=k+1}^{n}(x_{(i)}-x_{(j)})\epsilon_i + \frac{1}{\phi_j}\sum_{i=min(j+1,k+1)}^{max(j,k)}(x_{(i)}-x_{(j)})\epsilon_i \\ &= \frac{1}{\phi_k}\sum_{i=k+1}^{n}(x_{(i)}-x_{(j)})\epsilon_i + \frac{\phi_k-\phi_j}{\phi_k\phi_j}\sum_{i=k+1}^{n}(x_{(i)}-x_{(j)})\epsilon_i \\ &\quad + \frac{1}{\phi_j}\sum_{i=min(j+1,k+1)}^{max(j,k)}(x_{(i)}-x_{(j)})\epsilon_i.\end{aligned}$$

Then by the irrepresentable condition, $j \to k$ as long as $\psi_1 = O(1/\sqrt{n})$, thus $\frac{1}{\phi_j}\sum_{i=min(j+1,k+1)}^{max(j,k)}(x_{(i)}-x_{(j)})\epsilon_i = o(1)$ and

$$\begin{aligned}\frac{\phi_k-\phi_j}{\phi_k\phi_j}\sum_{i=k+1}^{n}(x_{(i)}-x_{(j)})\epsilon_i &= \frac{\phi_k^2-\phi_j^2}{\phi_k\phi_j(\phi_k+\phi_j)}\sum_{i=k+1}^{n}(x_{(i)}-x_{(j)})\epsilon_i \\ &= O_p((x_{(j)}-x_{(k)})).\end{aligned}$$

This is due to

$$\begin{aligned}\phi_k^2 - \phi_j^2 &= \sum_{i=k+1}^{n} (x_{(i)} - x_{(k)})^2 - \sum_{i=j+1}^{n} (x_{(i)} - x_{(j)})^2 \\ &= \sum_{i=k+1}^{n} (x_{(i)} - x_{(k)})^2 - (x_{(i)} - x_{(j)})^2 - \sum_{i=min(j+1,k+1)}^{max(j,k)} (x_{(i)} - x_{(j)})^2 \\ &= (x_{(j)} - x_{(k)}) \sum_{i=k+1}^{n} (2x_{(i)} - x_{(j)} - x_{(k)}) - \sum_{i=min(j+1,k+1)}^{max(j,k)} (x_{(i)} - x_{(j)})^2.\end{aligned}$$

As a result

$$\begin{aligned}&\Big(\frac{1}{\phi_j} \sum_{i=j+1}^{n} (x_{(i)} - x_{(j)})\epsilon_i - \frac{1}{\phi_k} \sum_{i=k+1}^{n} (x_{(i)} - x_{(k)})\epsilon_i\Big) \\ &= \Big(\frac{(x_{(k)} - x_{(j)})}{\phi_k} \sum_{i=j+1}^{n} \epsilon_i + O((x_{(k)} - x_{(j)}))\Big). \\ &= O_p(x_{(k)} - x_{(j)})\end{aligned}$$

Together with

$$\Big(\frac{1}{\phi_j} \sum_{i=j+1}^{n} (x_{(i)} - x_{(j)})\epsilon_i + \frac{1}{\phi_k} \sum_{i=k+1}^{n} (x_{(i)} - x_{(k)})\epsilon_i\Big) = O_p(1)$$

We have

$$\Big(\frac{1}{\phi_j} \sum_{i=j+1}^{n} (x_{(i)} - x_{(j)})\epsilon_i\Big)^2 - \Big(\frac{1}{\phi_k} \sum_{i=k+1}^{n} (x_{(i)} - x_{(k)})\epsilon_i\Big)^2 = O_p(x_{(k)} - x_{(j)})$$

Next, we focus on h_{11}, h_{12}, h_{21} and h_{22}. Lmma 2 highlights the connection among these four quantities, that is

$$h_{21}(\epsilon, \psi_1) = h_{11}(\epsilon, \psi_1) + O(1), \quad \text{and} \quad h_{22}(\epsilon, \psi_1) = h_{12}(\epsilon, \psi_1) + o_p(1)$$

As a result, we can combine the case of $k < j$ and $k > j$ and only consider h_{11} and h_{12}:

$$
\begin{aligned}
h_{11} &= \frac{\psi_1^2}{\phi_j^2}\left(\sum_{i=k+1}^{n}(x_{(i)}-x_{(k)})(x_{(i)}-x_{(j)})\right)^2 - \psi_1^2\phi_k^2 \\
&= \frac{\psi_1^2}{\phi_j^2}\left[\left(\sum_{i=k+1}^{n}(x_{(i)}-x_{(k)})(x_{(i)}-x_{(j)})\right)^2 - \phi_k^4 - \phi_k^2\left(\phi_j^2-\phi_k^2\right)\right] \\
&= \frac{\psi_1^2}{\phi_j^2}\left[\left(\sum_{i=k+1}^{n}(x_{(i)}-x_{(k)})(x_{(i)}-x_{(j)})\right)^2 - \phi_k^4 - \phi_k^2\left(\phi_j^2-\phi_k^2\right)\right] \\
&= (x_{(k)}-x_{(j)})\frac{\psi_1^2}{\phi_j^2}\left[\left(\sum_{i=k+1}^{n}(x_{(i)}-x_{(k)})\right)\left(\sum_{i=k+1}^{n}(x_{(i)}-x_{(k)})(2x_{(i)}-x_{(j)}-x_{(k)})\right)\right. \\
&\left.-\phi_k^2\left(\sum_{i=k+1}^{n}(x_{(i)}-x_{(k)})+(x_{(i)}-x_{(j)})\right)\right] \\
&= (x_{(k)}-x_{(j)})\frac{\psi_1^2}{\phi_j^2}\left[\left(\sum_{i=k+1}^{n}(x_{(i)}-x_{(k)})\right)\left(\sum_{i=k+1}^{n}(x_{(i)}-x_{(k)})(x_{(i)}-x_{(j)})\right)\right. \\
&\left.-\left(\phi_k^2\sum_{i=k+1}^{n}(x_{(i)}-x_{(k)})+(x_{(k)}-x_{(j)})\right)\right] \\
&= (x_{(k)}-x_{(j)})^2\frac{\psi_1^2}{\phi_j^2}\left[\left(\sum_{i=k+1}^{n}(x_{(i)}-x_{(k)})\right)^2 - \left(\phi_k^2(n-k)\right)\right].
\end{aligned}
$$

Notice that this quantity is of order $O\left((n-k)(x_{(k)}-x_{(j)})^2\psi_1^2\right)$. Next for h_{12},

$$
\begin{aligned}
h_{12} &= 2\frac{\psi_1}{\phi_j^2}\Big(\sum_{i=k+1}^{n}(x_{(i)}-x_{(j)})\epsilon_i\Big)\Big(\sum_{i=k+1}^{n}(x_{(i)}-z)(x_{(i)}-x_{(j)})\Big) - 2\psi_1\Big(\sum_{i=k+1}^{n}(x_{(i)}-x_{(k)})\epsilon_i\Big) \\
&= 2\frac{\psi_1}{\phi_j^2}\left(\sum_{i=k+1}^{n}\Big(\sum_{l=k+1}^{n}(x_{(l)}-z)(x_{(l)}-x_{(j)})(x_{(i)}-x_{(j)}) - (x_{(l)}-x_{(j)})^2(x_{(i)}-x_{(k)})\Big)\epsilon_i\Big)\right) \\
&= 2\frac{\psi_1}{\phi_j^2}(x_{(k)}-x_{(j)})\left(\sum_{i=k+1}^{n}\Big(\sum_{l=k+1}^{n}(x_{(l)}-x_{(i)})(x_{(l)}-x_{(j)})\Big)\epsilon_i\Big)\right) \\
&= 2\frac{\psi_1}{\phi_j^2}(x_{(k)}-x_{(j)})\left(\sum_{i=k+1}^{n}\Big(\sum_{l=k+1}^{n}(x_{(l)}-x_{(i)})(x_{(l)}-x_{(k)})\Big)\epsilon_i\Big)\right) \\
&+ 2\frac{\psi_1}{\phi_j^2}(x_{(k)}-x_{(j)})^2\left(\sum_{i=k+1}^{n}\Big(\sum_{l=k+1}^{n}(x_{(l)}-x_{(i)})\Big)\epsilon_i\Big)\right).
\end{aligned}
$$

We show next that for any $\psi_1 = \frac{\rho_n}{\sqrt{n-k}}$, $x_{(k)} - x_{(j)} = O(\rho_n^{-1})$. From h_{11} and h_{12} above

$$
\begin{aligned}
V(j) - V(k) = (x_{(k)}-x_{(j)})^2\frac{\psi_1^2}{\phi_j^2}&\left[\left(\sum_{i=k+1}^{n}(x_{(i)}-x_{(k)})\right)^2 - \big(\phi_k^2(n-k)\big)\right] \\
&+ 2\frac{\psi_1}{\phi_j^2}(x_{(k)}-x_{(j)})\left(\sum_{i=k+1}^{n}\Big(\sum_{l=k+1}^{n}(x_{(l)}-x_{(i)})(x_{(l)}-x_{(k)})\Big)\epsilon_i\Big)\right) \\
&+ 2\frac{\psi_1}{\phi_j^2}(x_{(k)}-x_{(j)})^2\left(\sum_{i=k+1}^{n}\Big(\sum_{l=k+1}^{n}(x_{(l)}-x_{(i)})\Big)\epsilon_i\Big)\right).
\end{aligned}
$$

And when $V(j) > V(k)$, we have

$$
\begin{aligned}
&|x_{(k)}-x_{(j)}|\rho_n\left[\left(\frac{\phi_k^2}{n-k} - \left(\frac{1}{n-k}\sum_{i=k+1}^{n}(x_{(i)}-x_{(k)})\right)^2\right)\right] \\
&\le 2\frac{1}{\sqrt{n-k}}\left|\sum_{i=k+1}^{n}\Big(\frac{1}{n-k}\sum_{l=k+1}^{n}(x_{(l)}-x_{(i)})(x_{(l)}-x_{(k)})\Big)\epsilon_i\Big)\right| \\
&+ 2|x_{(k)}-x_{(j)}|\frac{1}{\sqrt{n-k}}\left|\sum_{i=k+1}^{n}\Big(\frac{1}{n-k}\sum_{l=k+1}^{n}(x_{(l)}-x_{(i)})\Big)\epsilon_i\Big)\right|.
\end{aligned}
$$

Notice that $\frac{\phi_k^2}{n-k} > \frac{1}{n-k}\left(\sum_{i=k+1}^{n}(x_{(i)}-x_{(k)})\right)^2$ by Cauchy-Schwarz. Thus

$$|x_{(k)} - x_{(j)}| \leq \frac{C_1}{\rho_n + O(1)} \frac{2}{\sqrt{n-k}} \left| \sum_{i=k+1}^{n} \Big(\frac{1}{n-k} \sum_{l=k+1}^{n} (x_{(l)} - x_{(i)})(x_{(l)} - x_{(k)}) \Big) \epsilon_i \right|.$$

Thus $x_{(k)} - x_{(j)} = O(\rho_n^{-1})$. Now let $x_{(k)} - x_{(j)} = v\rho_n^{-1}$, we have

$$\begin{aligned} V(j) - V(k) = & \frac{v^2}{\phi_k^2/(n-k)} \left(\left(\frac{1}{n-k} \sum_{i=k+1}^{n} (x_{(i)} - x_{(k)}) \right)^2 - \frac{\phi_k^2}{n-k} \right) \\ & + \frac{v}{\phi_k^2/(n-k)} \frac{2}{\sqrt{n-k}} \left(\sum_{i=k+1}^{n} \Big(\frac{1}{n-k} \sum_{l=k+1}^{n} (x_{(l)} - x_{(i)})(x_{(l)} - x_{(k)}) \Big) \epsilon_i \right). \end{aligned} \tag{A11}$$

We can further expand the second term in equation A11 as

$$\begin{aligned} & \frac{2}{n-k} \left(\sum_{i=k+1}^{n} \Big(\sum_{l=k+1}^{n} (x_{(l)} - x_{(i)})(x_{(l)} - x_{(k)}) \Big) \epsilon_i \right) \\ & = \frac{1}{n-k} \left(\sum_{i=k+1}^{n} \sum_{l=k+1}^{n} (x_{(l)} - x_{(i)})(x_{(l)} - x_{(k)}) \epsilon_i + (x_{(i)} - x_{(l)})(x_{(i)} - x_{(k)}) \epsilon_l \right). \end{aligned}$$

This is in the form of V-statistic. Define $\phi_{il} = (x_{(l)} - x_{(i)})(x_{(l)} - x_{(k)})\epsilon_i + (x_{(i)} - x_{(l)})(x_{(i)} - x_{(k)})\epsilon_l$. Then by hoeffding decomposition, since

$$\mathbb{E}(\phi_{il}|i) = \mathbb{E}_l \Big((x_{(l)} - x_{(k)})(x_{(l)} - x_{(i)}) \Big) \epsilon_i.$$

we have the variance

$$\begin{aligned} var \left(\frac{1}{(n-k)^{3/2}} \sum_{il} \phi_{il} \right) = var \left(\mathbb{E}(\phi_{il}|i) \right) & = \mathbb{E} \left(\mathbb{E}_l \Big((x_{(l)} - x_{(k)})(x_{(l)} - x_{(i)}) \Big)^2 \right) \\ & = 2 \left(\mathbb{E}(x_{(l)} - x_{(k)})^2 \right) \left(\mathbb{E}(x_{(l)} - x_{(k)})^2 - \Big(\mathbb{E}(x_{(l)} - x_{(k)}) \Big)^2 \right). \end{aligned}$$

On the other hand, the first term in equation A11

$$\left(\left(\frac{1}{n-k}\sum_{i=k+1}^{n}(x_{(i)}-x_{(k)})\right)^2-\frac{\phi_k^2}{n-k}\right)\xrightarrow{p}-\left(\mathbb{E}(x_{(i)}-x_{(k)})^2-\left(\mathbb{E}(x_{(i)}-x_{(k)})\right)^2\right).$$

Thus

$$\rho_n(x_{(k)}-x_{(\hat{k})})\to_d \arg\max_v(v^2C+vZ)=-\frac{Z}{2C},$$

– and

$$-\frac{Z}{2C}\sim N\left(0,\frac{1}{2\left(1-\frac{(\mathbb{E}(x_{(l)}-x_{(k)}))^2}{\left(\mathbb{E}(x_{(l)}-x_{(k)})^2\right)}\right)}\right).$$

Notice that we will not be able to perform the same algebra trick for general k-th order discontinuity as

$$\begin{aligned}
&\frac{\psi_1^2}{\phi_j^2}\left(\sum_{i=k+1}^{n}(x_{(i)}-x_{(k)})^K(x_{(i)}-x_{(j)})^K\right)^2-\psi_1^2\phi_k^2\\
&=\frac{\psi_1^2}{\phi_j^2}\left[\left(\sum_{i=k+1}^{n}(x_{(i)}-x_{(k)})^K(x_{(i)}-x_{(j)})^K\right)^2-\phi_k^4-\phi_k^2\left(\phi_j^2-\phi_k^2\right)\right]\\
&=(x_{(k)}-x_{(j)})\frac{\psi_1^2}{\phi_j^2}\left[\left(\sum_{i=k+1}^{n}\sum_{a=0}^{K-1}(x_{(i)}-x_{(j)})^a(x_{(i)}-x_{(k)})^{2K-a-1}\right)\right.\\
&\cdot\left(\sum_{i=k+1}^{n}(x_{(i)}-x_{(k)})^K(x_{(i)}-x_{(j)})^K+(x_{(i)}-x_{(k)})^{2K}\right)\\
&\left.-\phi_k^2\left(\sum_{i=k+1}^{n}\sum_{a=0}^{K-1}(x_{(i)}-x_{(j)})^a(x_{(i)}-x_{(k)})^{K-a-1}\right)\left(\sum_{i=k+1}^{n}(x_{(i)}-x_{(k)})^K+(x_{(i)}-x_{(j)})^K\right)\right].
\end{aligned}$$

A-10 Lemma 2 [Kink Detection Algebra 1]

Lemma 2. *For any ψ_1 of order greater than $O(1/\sqrt{n})$*

$$h_{21}(\epsilon,\psi_1)=\frac{\psi_1^2}{\phi_j^2}\left(\sum_{i=k+1}^{n}(x_{(i)}-x_{(k)})(x_{(i)}-x_{(j)})\right)^2-\psi_1^2\phi_k^2+O(1)$$

$$h_{22}(\epsilon,\psi_1)=2\frac{\psi_1}{\phi_j^2}\Big(\sum_{i=k+1}^{n}(x_{(i)}-x_{(j)})\epsilon_i\Big)\Big(\sum_{i=k+1}^{n}(x_{(i)}-x_{(k)})(x_{(i)}-x_{(j)})\Big)-2\psi_1\Big(\sum_{i=k+1}^{n}(x_{(i)}-x_{(k)})\epsilon_i\Big)+o_p(1).$$

Proof. Notice that the first equation is equivalent as showing

$$\frac{\psi_1^2}{\phi_j^2}\left(\sum_{i=j+1}^{n}(x_{(i)}-x_{(k)})(x_{(i)}-x_{(j)})\right)^2=\frac{\psi_1^2}{\phi_j^2}\left(\sum_{i=k+1}^{n}(x_{(i)}-x_{(k)})(x_{(i)}-x_{(j)})\right)^2+O(1)\text{ when }k<j.$$

$$\begin{aligned}\frac{\psi_1^2}{\phi_j^2}\left(\sum_{i=k+1}^{n}(x_{(i)}-x_{(k)})(x_{(i)}-x_{(j)})\right)^2&=\frac{\psi_1^2}{\phi_j^2}\left(\sum_{i=j+1}^{n}(x_{(i)}-x_{(k)})(x_{(i)}-x_{(j)})+\sum_{i=k+1}^{j}(x_{(i)}-x_{(k)})(x_{(i)}-x_{(j)})\right)^2\\&=\frac{\psi_1^2}{\phi_j^2}\left(\sum_{i=j+1}^{n}(x_{(i)}-x_{(k)})(x_{(i)}-x_{(j)})\right)^2\\&+\frac{\psi_1^2}{\phi_j^2}\left(\sum_{i=k+1}^{j}(x_{(i)}-x_{(k)})(x_{(i)}-x_{(j)})\right)^2\\&+\frac{\psi_1^2}{\phi_j^2}\left(\sum_{i=j+1}^{n}(x_{(i)}-x_{(k)})(x_{(i)}-x_{(j)})\right)\left(\sum_{i=k+1}^{j}(x_{(i)}-x_{(k)})(x_{(i)}-x_{(j)})\right).\end{aligned}$$

Then by the irrepresentable condition, $j\to k$ as $\psi_1=O(1/\sqrt{n})$, thus $\sum_{i=k+1}^{j}(x_{(i)}-x_{(j)})(x_{(i)}-x_{(k)})=O(1)$ thus the last three terms in the above equation are of order $O(1)$. To show the second equation with respect to h_{22}, notice that it is equivalent as showing

$$\frac{\psi_1}{\phi_j^2}\Big(\sum_{i=j+1}^{n}(x_{(i)}-x_{(j)})\epsilon_i\Big)\Big(\sum_{i=k+1}^{n}(x_{(i)}-x_{(k)})(x_{(i)}-x_{(j)})\Big)$$
$$=\frac{\psi_1}{\phi_j^2}\Big(\sum_{i=j+1}^{n}(x_{(i)}-x_{(j)})\epsilon_i\Big)\Big(\sum_{i=j+1}^{n}(x_{(i)}-x_{(k)})(x_{(i)}-x_{(j)})\Big)+o_p(1).$$

and since

$$\frac{\psi_1}{\phi_j^2}\Big(\sum_{i=j+1}^{n}(x_{(i)}-x_{(j)})\epsilon_i\Big)\Big(\sum_{i=k+1}^{n}(x_{(i)}-x_{(k)})(x_{(i)}-x_{(j)})\Big)$$
$$-\frac{\psi_1}{\phi_j^2}\Big(\sum_{i=j+1}^{n}(x_{(i)}-x_{(j)})\epsilon_i\Big)\Big(\sum_{i=j+1}^{n}(x_{(i)}-x_{(k)})(x_{(i)}-x_{(j)})\Big)$$
$$=\frac{\psi_1}{\phi_j^2}\Big(\sum_{i=j+1}^{n}(x_{(i)}-x_{(j)})\epsilon_i\Big)\Big(\sum_{i=k+1}^{j}(x_{(i)}-x_{(k)})(x_{(i)}-x_{(j)})\Big).$$

Then, by the irrepresentable condition, $j \to k$ as $\psi_1 = O(1/\sqrt{n})$, thus $\frac{1}{\phi_j}\sum_{i=k+1}^{j}(x_{(i)}-x_{(j)})(x_{(i)}-x_{(k)}) = o(1)$ and thus

$$\frac{\psi_1}{\phi_j^2}\Big(\sum_{i=j+1}^{n}(x_{(i)}-x_{(j)})\epsilon_i\Big)\Big(\sum_{i=k+1}^{j}(x_{(i)}-x_{(k)})(x_{(i)}-x_{(j)})\Big)=o_p(1).$$