



AARHUS UNIVERSITY



# Coversheet

---

**This is the accepted manuscript (post-print version) of the article.**

Contentwise, the accepted manuscript version is identical to the final published version, but there may be differences in typography and layout.

## How to cite this publication

Please cite the final published version:

Cattaneo, M. D., Jansson, M., & Ma, X. (2018). Manipulation Testing Based on Density Discontinuity. *The Stata Journal*, 18(1), 234-261. <https://doi.org/10.1177/1536867X1801800115>

## Publication metadata

**Title:** Manipulation Testing Based on Density Discontinuity  
**Author(s):** Cattaneo, M. D., Jansson, M., & Ma, X.  
**Journal:** *The Stata Journal*  
**DOI/Link:** [10.1177/1536867X1801800115](https://doi.org/10.1177/1536867X1801800115)  
**Document version:** Accepted manuscript (post-print)

### General Rights

Copyright and moral rights for the publications made accessible in the public portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognize and abide by the legal requirements associated with these rights.

- Users may download and print one copy of any publication from the public portal for the purpose of private study or research.
- You may not further distribute the material or use it for any profit-making activity or commercial gain
- You may freely distribute the URL identifying the publication in the public portal

If you believe that this document breaches copyright please contact us providing details, and we will remove access to the work immediately and investigate your claim.

If the document is published under a Creative Commons license, this applies instead of the general rights.

## rddensity: Manipulation Testing Based on Density Discontinuity

Matias D. Cattaneo  
University of Michigan  
Ann Arbor, MI  
cattaneo@umich.edu

Michael Jansson  
UC-Berkeley and CREATES  
Berkeley, CA  
mjansson@econ.berkeley.edu

Xinwei Ma  
University of Michigan  
Ann Arbor, MI  
xinweima@umich.edu

**Abstract.** We introduce two *Stata* commands implementing automatic manipulation tests based on density discontinuity, constructed using the results for local polynomial density estimators in [Cattaneo, Jansson, and Ma \(2017a\)](#). These new tests exhibit better size properties (and more power under additional assumptions) than other conventional approaches currently available in the literature. The first command, `rddensity`, implements manipulation tests based on a novel local polynomial density estimation technique that avoids pre-binning of the data (improving size properties) and allows for restrictions on other features of the model (improving power properties). The second command, `rdbwdensity`, implements several bandwidth selectors specifically tailored for the manipulation tests discussed herein. A companion *R* package with the same syntax and capabilities as `rddensity` and `rdbwdensity` is also provided.

**Keywords:** st0001, rddensity, rdbwdensity, falsification test, manipulation test, regression discontinuity.

This version: October 7, 2017

## 1 Introduction

[McCrary \(2008\)](#) introduced the idea of “manipulation testing” in the context of regression discontinuity (RD) designs. Consider a setting where each unit in a random sample from a large population is assigned to one of two groups depending on whether one of their observed covariates exceeds a known cutoff. In this context, the two possible groups are generically referred to as “control” and “treatment” groups, and the observed variable determining group assignment is generically referred to as “score”, “index” or “running variable”. The key idea behind manipulation testing in this context is that, in the absence of systematic manipulation of the unit’s index around the cutoff, the density of units should be continuous near this cutoff value. Thus, a manipulation test seeks to formally determine whether there is evidence of a discontinuity in the density of units at the known cutoff. Presence of such evidence is usually interpreted as empirical evidence of self-selection or non-random sorting of units into control and treatment status.

Manipulation testing is useful for falsification of RD designs: see [Cattaneo and Escanciano \(2017\)](#) for an edited volume with a recent overview of the RD literature, [Cattaneo, Titiunik, and Vazquez-Bare \(2017c\)](#) for a practical introduction to RD designs with a comparison between leading empirical methods, [Calonico, Cattaneo, and Titiunik \(2015a\)](#) for a discussion of graphical presentation and falsification of RD designs, and references therein for other related topics. In addition to providing a formal statistical check for RD designs, a manipulation test can be used substantively whenever the empirical goal is to test for self-selection or endogenous sorting of units exposed to a known hard threshold-crossing assignment rule. Thus, flexible data-driven implementations of manipulation tests, with good size and power properties, are potentially very useful for empirical work in Economics and related social sciences.

To implement a manipulation test, the researcher needs to estimate the density of units near the cutoff to conduct a hypothesis test about whether the density is discontinuous. Three distinct manipulation tests have been proposed in the literature. First, [McCrary \(2008\)](#) introduced a test based on the nonparametric local polynomial density estimator of [Cheng, Fan, and Marron \(1997\)](#), which requires pre-binning of the data and hence introduces additional tuning parameters. Second, [Otsu, Xu, and Matsushita \(2014\)](#) proposed an empirical likelihood method employing boundary corrected kernels. Third, [Cattaneo, Jansson, and Ma \(2017a\)](#) developed a set of manipulation tests based on a novel local polynomial density estimator, which does not require pre-binning of the data and is constructed in an intuitive way based on easy-to-interpret kernel functions. The latter procedures are shown to also provide demonstrable improvements in both size and power, under appropriate assumptions, relative to the other approaches currently available in the literature. Finally, [Frandsen \(2017\)](#) recently proposed a manipulation test in the context of RD designs with a discrete running variable.

This article discusses data-driven implementations of manipulations tests following the results in [Cattaneo, Jansson, and Ma \(2017a\)](#). We introduce two commands that together give several manipulation test implementations, which depend on (i) the restrictions imposed in the underlying data generating process, (ii) the method for bias-correction, (iii) the bandwidth selection approach and (iv) the method to estimate

standard errors, among many other alternatives. Specifically, our first main command `rddensity` implements two distinct manipulation tests given a choice of bandwidth and standard errors estimator: one test is constructed using the basic Wald statistic, which requires undersmoothing, while the other test employs robust bias-correction (Calonico et al. 2017a) to obtain valid statistical inference. This command also allows for both unrestricted models as well as restricted models where the cumulative distribution function and higher-order derivatives are assumed to be equal for both groups (which increase the power of the test). These methods can also be used in the context of RD designs with a discrete running variable, under additional assumptions. We review the main aspects of these different approaches to manipulation testing in Section 2 below, but we refer the reader to Cattaneo, Jansson, and Ma (2017a) and its supplemental appendix for most of the technical and theoretical discussion.

The command `rddensity` also offers a plot of the manipulation test. To implement this plot a density estimate needs to be constructed not only at the cutoff point but also at other nearby evaluation points, which may also be affected by boundary bias. Thus, to construct this plot in a principled way, the command `rddensity` employs the package `lpdensity`, which implements local polynomial based density estimation methods. This density estimation package needs to be installed to construct the density plot: see Cattaneo, Jansson, and Ma (2017b) for further details. If the package `lpdensity` is not installed, the command `rddensity` issues an error message when trying to construct a manipulation test plot.

To complement the command `rddensity`, and because in empirical applications researchers often want to select the bandwidth entering the manipulation test in a data-driven and automatic way, we introduce the companion command `rdbwdensity`, which provides several bandwidth selection methods based on asymptotic mean squared error (MSE) minimization. Our implementation takes into account whether the unrestricted or restricted model is used for inference, and also allows for both different bandwidths on either side of the cutoff (whenever possible) as well as a common bandwidth for both sides. By default, our main command `rddensity` employs the companion command `rdbwdensity` to estimate the bandwidth(s), whenever the user does not provide a specific choice, thereby giving fully automatic and data-driven inference procedures implemented by `rddensity`.

The two commands discuss herein complement the recently introduced `Stata` and `R` commands/functions `rdrobust`, `rdbwselect` and `rdplot`, which are useful for graphical presentation, estimation and inference in RD designs, using nonparametric local polynomial techniques. For an introduction to the latter commands see Calonico, Cattaneo, and Titiunik (2014a, 2015b) and Calonico, Cattaneo, Farrell, and Titiunik (2017b). Together, the five commands offer a complete toolkit for empirical work employing RD designs. In addition, see Cattaneo, Titiunik, and Vazquez-Bare (2016) for `Stata` and `R` commands/functions (`rdrandinf`, `rdwselect`, `rdsensitivity`, `rdrbounds`) implementing randomization-based inference methods for RD designs under a local randomization assumption.

The rest of this article is organized as follows. Section 2 provides a brief review

of the methods implemented in our two commands. The syntax of `rddensity` and `rdbwdensity` are described in Sections 3 and 4, respectively. In Section 5, we illustrate some of the functionalities of our commands employing real data from Cattaneo, Frandsen, and Titiunik (2015). And in Section 6 we report results from a small scale simulation study, tailored to investigate the performance of our testing procedure. A companion R package with the same functionalities and syntax is also provided.

The latest version of this software, as well as other related softwares for RD designs, can be found at: <https://sites.google.com/site/rdpackages/>.

## 2 Methods Overview

This section offers a brief overview of the methods implemented in our commands `rddensity` and `rdbwdensity`. We follow closely the results in Cattaneo, Jansson, and Ma (2017a, CJM hereafter), including those in their supplemental appendix. Regularity conditions and most of the technical details are not discussed here to conserve space and ease the exposition.

### 2.1 Setup and Notation

We assume that  $\{X_1, X_2, \dots, X_n\}$  is a random sample of size  $n$  from the random variable  $X$  with cumulative distribution function (c.d.f.) and probability density function (p.d.f.) given by  $F(x)$  and  $f(x)$ , respectively. The random variable  $X_i$  denotes the score, index or running variable of unit  $i$  in the sample. Each unit is assigned to control or treatment depending on whether their observed index exceeds a known cutoff denoted by  $\bar{x}$ . That is, group or “treatment” assignment is given by:

$$\begin{aligned} \text{unit } i \text{ assigned to control group} & \quad \text{if } X_i < \bar{x} \\ \text{unit } i \text{ assigned to treatment group} & \quad \text{if } X_i \geq \bar{x} \end{aligned}$$

where the cutoff point  $\bar{x}$  is known and, of course, we assume enough observations for each group are available (e.g.,  $f(x) > 0$  near  $\bar{x}$  and the sample is large enough). In the specific case of RD designs, manipulation testing can be used for sharp RD designs (where treatment assignment and treatment status coincide) as well as for fuzzy RD designs (where treatment assignment and treatment status differ). In the later case, of course, the test applies to the intention-to-treat mechanism because units can select into treatment or control status beyond the hard-thresholding rule for treatment assignment (i.e., assigned to control group if  $X_i < \bar{x}$  and assigned to treatment group if  $X_i \geq \bar{x}$ ).

A manipulation test in this context is a hypothesis test on the continuity of the density  $f(\cdot)$  at the cutoff point  $\bar{x}$ . Formally, we are interested in the testing problem:

$$H_0 : \lim_{x \uparrow \bar{x}} f(x) = \lim_{x \downarrow \bar{x}} f(x) \quad \text{vs} \quad H_1 : \lim_{x \uparrow \bar{x}} f(x) \neq \lim_{x \downarrow \bar{x}} f(x).$$

To construct a test statistic for this hypothesis testing problem, we follow CJM and estimate the density  $f(x)$  using a local polynomial density estimator based on the c.d.f.

of the observed sample. This estimator has several interesting properties, including the fact that it does not require pre-binning of the data and is quite intuitive in its implementation (e.g., simple second-order kernels can be used). Importantly, this estimator also permits to incorporate restrictions on the c.d.f. and higher-order derivatives of the density, leading to new manipulation tests with better power properties in applications. For an introduction to conventional local polynomial techniques see, e.g., [Fan and Gijbels \(1996\)](#).

The class of manipulation test statistics implemented in `rddensity` take the form:

$$T_p(h) = \frac{\hat{f}_{+,p}(h) - \hat{f}_{-,p}(h)}{\hat{V}_p(h)}, \quad \hat{V}_p^2(h) = \hat{\mathbb{V}}[\hat{f}_{+,p}(h) - \hat{f}_{-,p}(h)],$$

where  $T_p(h) \stackrel{d}{\sim} \mathcal{N}(0, 1)$  under appropriate assumptions, and the notation  $\hat{\mathbb{V}}[\cdot]$  is meant to denote some plug-in consistent estimator of the population quantity  $\mathbb{V}[\cdot]$ . The parameter  $h$  is the bandwidth(s) used to localize the estimation and inference procedures near the cutoff point  $\bar{x}$ . The statistics may be constructed in several different ways, as we discuss in more detail below. In particular, given a bandwidth(s) choice, two main ingredients to construct the test statistic  $T_p(h)$  are (i) the local polynomial density estimators  $\hat{f}_{+,p}(h)$  and  $\hat{f}_{-,p}(h)$ , and (ii) the corresponding standard error estimator  $\hat{V}_p(h)$ . These estimators also depend on the choice of polynomial order  $p$ , the choice of kernel function  $K(\cdot)$  and the restrictions imposed in the model, among other possibilities. The standard error formulas  $\hat{V}_p(h)$  could be based on either an asymptotic plug-in or jackknife approach, and its specific form will depend on whether additional restrictions are imposed to the model. A crucial ingredient is, of course, the choice of bandwidth  $h$ , which determines which observations near the cutoff  $\bar{x}$  are used for estimation and inference. This choice can either be specified by the user or estimated using the available data. Our commands allow, when possible, for different bandwidth choices on either side of the cutoff  $\bar{x}$ . A common bandwidth on both sides of the cutoff is always possible.

In the following two subsection we discuss the different alternatives for estimation and inference: (i) unrestricted inference, (ii) restricted inference, (iii) standard error estimation in both cases, (iv) bandwidth selection in both cases. In closing this section, we also offer a brief review of the different data-driven inference methods implemented in our `Stata` (and `R`) commands.

## 2.2 Unrestricted testing

In unrestricted testing, the manipulation test becomes a standard two-sample problem where the estimators  $\hat{f}_{+,p}(h)$  and  $\hat{f}_{-,p}(h)$  are unrelated. Thus, the standard errors formula reduces to  $\hat{V}_p^2(h) = \hat{\mathbb{V}}[\hat{f}_{+,p}(h) - \hat{f}_{-,p}(h)] = \hat{\mathbb{V}}[\hat{f}_{+,p}(h)] + \hat{\mathbb{V}}[\hat{f}_{-,p}(h)]$ . To be more concrete, the density estimators take the form:

$$\hat{f}_{-,p}(h) = \mathbf{e}'_1 \hat{\boldsymbol{\beta}}_{-,p}(h) \quad \text{and} \quad \hat{f}_{+,p}(h) = \mathbf{e}'_1 \hat{\boldsymbol{\beta}}_{+,p}(h),$$

with

$$\hat{\beta}_{-,p}(h) = \arg \min_{\beta \in \mathbb{R}^{p+1}} \sum_{i=1}^n \mathbf{1}(X_i < \bar{x}) (\hat{F}(X_i) - \mathbf{r}_p(X_i - \bar{x})' \beta)^2 K_h(X_i - \bar{x}),$$

$$\hat{\beta}_{+,p}(h) = \arg \min_{\beta \in \mathbb{R}^{p+1}} \sum_{i=1}^n \mathbf{1}(X_i \geq \bar{x}) (\hat{F}(X_i) - \mathbf{r}_p(X_i - \bar{x})' \beta)^2 K_h(X_i - \bar{x}),$$

where  $\hat{F}(X_i)$  denotes the (leave-one-out) c.d.f. estimator for the full sample,  $\mathbf{r}_p(x) = (1, x, \dots, x^p)'$ ,  $\mathbf{e}_1 = (0, 1, 0, 0, \dots, 0)' \in \mathbb{R}^{p+1}$  is the second unit vector,  $K_h(u) = K(u/h)/h$  with  $K(\cdot)$  a kernel function,  $h$  is a positive bandwidth sequence, and  $\mathbf{1}(\cdot)$  denotes the indicator function. Following the results in CJM, it can be shown that  $\hat{\beta}_{-,p}(h)$  and  $\hat{\beta}_{+,p}(h)$  roughly approximate, respectively,  $(F_-, f_-, \frac{1}{2!} f_-^{(1)}, \dots, \frac{1}{(p+1)!} f_-^{(p)})$  and  $(F_+, f_+, \frac{1}{2!} f_+^{(1)}, \dots, \frac{1}{(p+1)!} f_+^{(p)})$ , where we employ the notation:

$$f_-^{(s)} = \lim_{x \uparrow \bar{x}} \frac{\partial^s}{\partial x^s} f(x) \quad \text{and} \quad f_+^{(s)} = \lim_{x \downarrow \bar{x}} \frac{\partial^s}{\partial x^s} f(x),$$

$s = 1, 2, \dots, p$ . Thus, the second elements of  $\hat{\beta}_{-,p}(h)$  and  $\hat{\beta}_{+,p}(h)$  are used to construct consistent, boundary-corrected density estimators at the cutoff point  $\bar{x}$  entering the numerator of the manipulation test statistic. Notice that this estimation approach avoids pre-binning of the data and may be constructed using simple and easy-to-interpret kernels  $K(\cdot)$ , such as the uniform or triangular kernels.

Consistency, asymptotic normality and moment approximations for the estimators  $\hat{f}_{-,p}(h)$  and  $\hat{f}_{+,p}(h)$  are derived in CJM, where these results are also used to study the asymptotic properties of the unrestricted manipulation tests implemented in our main command `rddensity`. We discuss the exact implementations at the end of this section.

### 2.3 Restricted testing

The approach described above treats manipulation testing as a traditional two-sample problem, where the left and right approximations to the density  $f(x)$  at the cutoff  $\bar{x}$  are done independently. However, in the context of manipulation testing, it may be argued that the c.d.f.  $F(x)$  and higher-order derivatives  $f^{(s)}(x)$ ,  $s \geq 1$ , are equal for both groups at the cutoff even when  $f_- \neq f_+$  (i.e.,  $H_0$  is false). In this case, researchers may wish to incorporate these restrictions to construct a more powerful testing procedure.

A restricted manipulation test is constructed by solving the above weighted (local) least-squares problem with the additional restrictions ensuring that all but the second element in  $\beta$  are equal in both groups. It follows that this restricted problem can be represented as a single regression problem:

$$\hat{\beta}_p^R(h) = \arg \min_{\beta \in \mathbb{R}^{p+2}} \sum_{i=1}^n (\hat{F}(X_i) - \mathbf{r}_p^R(X_i - \bar{x})' \beta)^2 K_h(X_i - \bar{x}),$$

where  $\mathbf{r}_p^{\mathbf{R}}(x) = (1, x \cdot \mathbf{1}(x < \bar{x}), x \cdot \mathbf{1}(x \geq \bar{x}), x^2, x^3, \dots, x^p)'$ . In words, this estimation approach incorporates the restrictions ensuring that the estimated c.d.f. as well as the estimated higher-order derivatives are equal on both sides of the cutoff point  $\bar{x}$ . This problem involves estimating only  $p+2$  parameters, rather than  $2p+2$  parameters as it is the case for the unrestricted method discussed above. The main advantage of imposing these restrictions is related to power improvements, provided the restrictions are indeed satisfied in the underlying data generating process.

Therefore, a restricted manipulation test employs the density estimators:

$$\hat{f}_{-,p}^{\mathbf{R}}(h) = \mathbf{e}'_1 \hat{\boldsymbol{\beta}}_p^{\mathbf{R}}(h) \quad \text{and} \quad \hat{f}_{+,p}^{\mathbf{R}}(h) = \mathbf{e}'_2 \hat{\boldsymbol{\beta}}_p^{\mathbf{R}}(h),$$

that is, the density estimators from the left and from the right of the cutoff point  $\bar{x}$  are given by the second and third element of in the least-squares vector  $\hat{\boldsymbol{\beta}}_p^{\mathbf{R}}(h)$ ;  $\mathbf{e}_2 = (0, 0, 1, 0, 0, \dots, 0)' \in \mathbb{R}^{p+1}$  denotes the third unit vector. Furthermore, the standard errors formula is different because now cross-restrictions are incorporated in the estimation procedure, leading to a different asymptotic variance for  $(\hat{f}_{-,p}^{\mathbf{R}}(h), \hat{f}_{+,p}^{\mathbf{R}}(h))$  and, in consequence, for  $\hat{f}_{+,p}^{\mathbf{R}}(h) - \hat{f}_{-,p}^{\mathbf{R}}(h)$  as well. This is exactly the source of the power gains of a restricted manipulation test relative to an unrestricted one. In this case, the standard errors formula  $\hat{V}_p(h) = \hat{\mathbb{V}}[\hat{f}_{+,p}^{\mathbf{R}}(h) - \hat{f}_{-,p}^{\mathbf{R}}(h)] \neq \hat{\mathbb{V}}[\hat{f}_{+,p}^{\mathbf{R}}(h)] + \hat{\mathbb{V}}[\hat{f}_{-,p}^{\mathbf{R}}(h)]$ .

Formal asymptotic properties for restricted local polynomial density estimation and inference are also discussed in CJM, where these results are then used to propose an asymptotically valid restricted manipulation test. We also discuss the exact implementations at the end of this section.

## 2.4 Standard Errors

As mentioned above, the asymptotic variance entering the denominator of the manipulation test statistic  $T_p(h)$  will be different depending on whether an unrestricted or a restricted model is used. Our commands `rddensity` and `rdbwdensity` allow for both cases. In addition, for each of these cases, the commands provide two distinct consistent standard errors estimators: (i) a plug-in estimator based on the asymptotic variance of the numerator of  $T_p(h)$ , and (ii) a jackknife estimator based on the leading term of an expansion of asymptotic variance of the numerator of  $T_p(h)$ .

The plug-in estimator is faster because it essentially requires no additional estimation beyond the quantities entering the numerator of  $T_p(h)$ , but it relies on asymptotic approximations. On the other hand, the jackknife estimator is slower because it requires additional estimation and looping over the data, but according to simulation evidence in CJM it appears to provide a more accurate approximation to the finite-sample variability of the numerator of  $T_p(h)$  in both cases (unrestricted model and restricted model). Therefore, our implementations employ the jackknife standard error estimator by default, but we also offer the plug-in estimator for cases involving relatively large sample sizes.

## 2.5 Bandwidth Selection

The command `rddensity` either requires specifying bandwidths for estimation, or employs `rdbwdensity` to construct data-driven bandwidth choices specifically tailored for the manipulation tests discussion in this article. In this subsection we briefly outline the data-driven implementations provided in `rdbwdensity` for automatic bandwidth selection.

In the unrestricted model, the user has the option to specify two distinct bandwidths:  $h_l$  for left estimation and  $h_r$  for right estimation. Of course, one such choice may be equal bandwidths:  $h = h_l = h_r$ . In the restricted model, however, only a common bandwidth  $h$  can be specified because estimation is done jointly by construction. For each of these cases, whenever possible, CJM develops three distinct approaches to select the the bandwidth(s) employing the mean squared error (MSE) criterion function, generically denoted by  $\text{MSE}[\hat{\theta}] = \mathbb{E}[(\hat{\theta} - \theta)^2]$ , where  $\hat{\theta}$  denotes some estimator and  $\theta$  denotes its target estimand.

For the specific context considered in this article, CJM develop valid asymptotic expansions of several MSE criterion functions. These results can be briefly summarized as follows:

$$\text{MSE}[\hat{\theta}(h)] \approx \text{AMSE}[\hat{\theta}(h)],$$

where

$$\text{AMSE}[\hat{\theta}(h)] = h^{p+1}B_p^2(\theta) + h^{p+2}B_{p+1}^2(\theta) + \frac{1}{nh}V_p(\theta),$$

with, for the unrestricted model,

$$\hat{\theta}(h) \text{ representing one of } \{\hat{f}_{-,p}(h); \hat{f}_{+,p}(h); \hat{f}_{+,p}(h) - \hat{f}_{-,p}(h); \hat{f}_{+,p}(h) + \hat{f}_{-,p}(h)\}$$

while, for the restricted model,

$$\hat{\theta}(h) \text{ representing one of } \{\hat{f}_{+,p}^R(h) - \hat{f}_{-,p}^R(h); \hat{f}_{+,p}^R(h) + \hat{f}_{-,p}^R(h)\},$$

and, of course,

$$\theta \text{ representing one of } \{f_-; f_+; f_+ - f_-; f_+ + f_-\},$$

as a appropriate according to the choice of  $\hat{\theta}(h)$ . Crucially, for each combination of estimator  $\hat{\theta}(h)$  and estimand  $\theta$ , the corresponding bias constants ( $B_p(\theta)$ ,  $B_{p+1}(\theta)$ ) and variance constant  $V_p(\theta)$  are different. In all cases, as it is usually the case in nonparametric problems, these constants involve features of both the data generating process as well as the nonparametric estimator.

Given a choice of estimator and estimand, and provided preliminary estimates of the leading asymptotic constants in the associated MSE expansion are available, it is straightforward to construct a plug-in bandwidth selector. In our implementations, we consider the five alternative plug-in rules for bandwidth selection mentioned above, which employ simple rule-of-thumbs to approximate the leading constants in the MSE

expansions. Technical and methodological details underlying these choices are given in CJM, and not reproduced here to conserve space.

Specifically, `rdbwdensity` allows for the following alternative bandwidth selectors. We do not introduce additional notation to reflect the estimation of the leading AMSE constants only to ease the exposition, but our implementations are automatic as they rely on preliminary rule-of-thumb estimators to approximate those constants.

- *Unrestricted model with different bandwidths*: when no restrictions are imposed on the model and  $(h_1, h_x)$  are allowed to be different, the bandwidths are chosen to minimize the AMSE of the density estimators separately, that is,

$$\hat{h}_{1,p} = \operatorname{argmin}_{h>0} \operatorname{AMSE}[\hat{f}_{-,p}(h)] \quad \text{and} \quad \hat{h}_{x,p} = \operatorname{argmin}_{h>0} \operatorname{AMSE}[\hat{f}_{+,p}(h)].$$

- *Unrestricted model with equal bandwidth*: when no restrictions are imposed on the model but  $(h_1, h_x)$  are forced to be equal, the common bandwidth may be chosen in two distinct ways.

(1) Difference of densities:

$$\hat{h}_{\text{diff},p} = \operatorname{argmin}_{h>0} \operatorname{AMSE}[\hat{f}_{+,p}(h) - \hat{f}_{-,p}(h)].$$

(2) Sum of densities:

$$\hat{h}_{\text{sum},p} = \operatorname{argmin}_{h>0} \operatorname{AMSE}[\hat{f}_{+,p}(h) + \hat{f}_{-,p}(h)].$$

- *Restricted model*: when the restrictions are imposed on the model, then  $h = h_1 = h_x$  by construction. In this case, the common bandwidth may be chosen in two distinct ways as well.

(1) Difference of densities:

$$\hat{h}_{\text{diff},p}^{\text{R}} = \operatorname{argmin}_{h>0} \operatorname{AMSE}[\hat{f}_{+,p}^{\text{R}}(h) - \hat{f}_{-,p}^{\text{R}}(h)].$$

(2) Sum of densities:

$$\hat{h}_{\text{sum},p}^{\text{R}} = \operatorname{argmin}_{h>0} \operatorname{AMSE}[\hat{f}_{+,p}^{\text{R}}(h) + \hat{f}_{-,p}^{\text{R}}(h)].$$

All the bandwidth selectors above have close form solutions, and their specific decay rate (as a function of the sample size) depend on the specific choices. This is an important point, because of  $B_p(\theta)$  or  $B_{p+1}(\theta)$  may be zero depending on the choice of  $\theta$  and  $p$ . For example, if  $\theta = f_+ - f_-$  then  $B_p(\theta) \propto f_+^{(p)} - f_-^{(p)}$  when  $p = 2$  but  $B_p(\theta) \propto f_+^{(p)} + f_-^{(p)}$  when  $p = 3$ , which implies that  $B_p(\theta) = 0$  under the plausible assumption that  $f_+^{(p)} = f_-^{(p)}$ . Following the discussion in CJM, we also implement to simple “regularization” approaches that avoid this possible degeneracies:

- *Unrestricted model with different bandwidths:*

$$\hat{h}_{1,\text{comb},p} = \text{median}\{\hat{h}_{1,p}, \hat{h}_{\text{diff},p}, \hat{h}_{\text{sum},p}\}$$

$$\hat{h}_{r,\text{comb},p} = \text{median}\{\hat{h}_{r,p}, \hat{h}_{\text{diff},p}, \hat{h}_{\text{sum},p}\}.$$

- *Unrestricted model with equal bandwidth:*

$$\hat{h}_{\text{comb},p} = \min\{\hat{h}_{\text{diff},p}, \hat{h}_{\text{sum},p}\}.$$

- *Restricted model:*

$$\hat{h}_{\text{comb},p}^{\text{R}} = \min\{\hat{h}_{\text{diff},p}^{\text{R}}, \hat{h}_{\text{sum},p}^{\text{R}}\}.$$

## 2.6 Overview of Methods

We have discussed how the density point estimators and corresponding standard errors are constructed in order to form the test statistic  $T_p(h)$ . As mentioned above, these estimators depend on whether the unrestricted or the restricted model is considered. In addition, the bandwidth  $h = (h_1, h_r)$  may be chosen in different ways, including both cases where the two bandwidths are assumed equal and cases where they are allowed to be different. In this section, we close the discussion of manipulation testing by briefly addressing the issue of critical value (or quantile) choice to form the testing procedure.

As mentioned in passing before, we rely on large sample approximations to justify a result of the form:  $T_p(h) \stackrel{a}{\sim} \mathcal{N}(0, 1)$ , provided an appropriate choice of bandwidth  $h$  and polynomial order  $p$  is used. Specifically, the command `rddensity` implements three distinct methods for inference. Each of these methods take a different approach to handle the potential presence of a first-order bias in the statistic  $T_p(h)$  when a “large” bandwidth  $h$  is used (e.g., when any of the MSE-optimal bandwidth choices discussed above are used).

To give a brief discussion of the three alternatives manipulation tests, we let  $h_{\text{MSE},p}$  denote any of the bandwidth choices given above when  $p$ -th order local polynomial density estimators are used. The following discussion applies to all cases, unrestricted or restricted models with equal or unequal bandwidths chosen by any of the methods mentioned before. Let  $\alpha \in (0, 1)$  and  $\chi_1^2(\alpha)$  denote the  $\alpha$ -th quantile of a chi-squared distribution with 1 degree of freedom. The three methods for inference implemented in `rddensity` are as follows:

- *Robust bias-correction approach.* This approach to inference employs ideas in [Calonico, Cattaneo, and Titiunik \(2014b\)](#) and [Calonico, Cattaneo, and Farrell \(2017a\)](#), which are based on analytic bias-correction coupled with variance adjustments, leading to testing procedures with improved distributional properties in finite samples. In particular, letting  $q \geq p + 1$ , the manipulation test takes the form:

$$\text{An } \alpha \text{ level test rejects } H_0 \quad \text{iff} \quad T_q^2(h_{\text{MSE},p}) > \chi_1^2(1 - \alpha).$$

In words, the construction of  $T_q^2(h_{\text{MSE},p})$  ensures an asymptotically valid distributional approximation because  $q \geq p + 1$ . Thus, in this case, the possible first-order bias of the statistic  $T_p^2(h_{\text{MSE},p})$  (note change from  $q$  to  $p$  in subindex) is removed by employing a higher-order polynomial in the estimation of the densities and, of course, adjusting the standard error formulas accordingly.

This approach to manipulation testing is the default option in `rddensity` because it is always theoretically justified and leads to power improvements asymptotically. For example, see the simulations presented in Section 6.

- *Small bias approach/Undersmoothing approach.* A second approach to inference would be to ignore the impact of a possibly first-order bias in the distributional approximation for the statistic  $T_p^2(h_{\text{MSE},p})$ . In some specific cases, such an approach may be justified (e.g., if  $B_p(\theta)$  happens to be exactly zero for a choice of  $p$  and  $\theta$ ) and the bandwidth was chosen in an ad-hoc manner, but this cannot be justified in general. Alternatively, this approach is valid if the user employs an undersmoothed bandwidth choice relative to the MSE-optimal bandwidth  $h_{\text{MSE},p}$ .

For completeness, the command `rddensity` also implements a conventional Wald test without bias-correction. To describe this procedure, suppose  $h$  is the bandwidth used, then the command implements the conventional testing procedure:

$$\text{Reject } H_0 \quad \text{iff} \quad T_p^2(h) > \chi_1^2(1 - \alpha),$$

which delivers a valid  $\alpha$  level test only when the smoothing bias in approximating  $f(x)$  at the cutoff point  $\bar{x}$  is indeed “small”. In practice, this requires choosing an undersmoothed bandwidth, for example, an ad-hoc choice is  $h = s \cdot h_{\text{MSE},p}$  for some user-chosen scale value  $s \in (0, 1)$ . This testing procedure is also reported when the option `all` is specified.

### 3 rddensity syntax

This section describes the syntax of the command `rddensity`, which implements the manipulation tests for a choice of bandwidth(s).

#### 3.1 Syntax

```
rddensity runvar [if][in][, c(cutoff) p(pvalue) q(qvalue)
  fitselect(fitmethod) kernel(kernelfn) h(hvalue hvalue)
  bwselect(bwmethod) vce(vcemethod) all plot
  plot_range(xmin xmax) plot_n(nl nr) plot_grid(gridmethod)
  genvars(varname) level(a) graph_options(...) ]
```

`runvar` is the running variable (a.k.a. score or forcing variable).

*cutoff* is the RD cutoff. Default is  $\bar{x} = 0$ .

*pvalue* is the order of the local-polynomial used to construct the density point estimator. Default is  $p = 2$  (local-quadratic).

*qvalue* is the order of the local-polynomial used to construct the bias-corrected density point estimator. Default is  $q = p + 1$  (local-cubic for the default  $p = 2$ ).

*fitmethod* specifies the model used for estimation and inference. Options are:

**unrestricted.** For density estimation without any restrictions (two-sample, unrestricted inference). This is the default option.

**restricted.** For density estimation assuming equal c.d.f. and higher-order derivatives.

*kernelfn* is the kernel function  $K(\cdot)$  used to construct the local-polynomial estimator(s). Options are: **triangular**, **uniform** and **epanechnikov**. Default is **triangular**.

*hlvalue* and *hrvalue* are the values of the main bandwidths  $h_1$  and  $h_x$ , respectively. If only one value is specified, then  $h_1 = h_x$  is used. If not specified, they are computed by the companion command `rdbdensity`.

*bwmethod* specifies the bandwidth selection procedure to be used. Options are:

**each.** Bandwidth selection based on MSE of each density separately, that is,  $\hat{h}_{1,p}$  and  $\hat{h}_{x,p}$ . Available only when the unrestricted model is used.

**diff.** Bandwidth selection based on MSE of difference of densities, that is,  $\hat{h}_{\text{diff},p}$ .

**sum.** Bandwidth selection based on MSE of sum of densities, that is,  $\hat{h}_{\text{sum},p}$ .

**comb.** Bandwidth is selected as a combination of the alternatives above, that is, either  $(\hat{h}_{1,\text{comb},p}, \hat{h}_{x,\text{comb},p})$ ,  $\hat{h}_{\text{comb},p}$  or  $\hat{h}_{\text{comb},p}^R$ , depending on the model chosen. This is the default option.

*vcemethod* specifies the procedure used to compute the variance-covariance matrix estimator. Options are:

**plugin.** For asymptotic plug-in standard errors.

**jackknife.** For jackknife standard errors. This is the default option.

*all* reports two different manipulation tests:

- (1) Conventional test statistic (not valid when using MSE-optimal bandwidth choice).
- (2) Robust bias-corrected statistic. This is the default option.

*plot* produces manipulation test plot.

`plot_range(xmin xmax)` specifies lower (*xmin*) and upper (*xmax*) end points for manipulation test plot.

`plot_n(nl nr)` specifies number of evaluation points below (*nl*) and above (*nr*) cutoff to be used for manipulation test plot. Default is 10 on each side.

`plot_grid(gridtype)` specifies location of evaluation points, evenly spaced (*es*) or quantile spaced (*es*), to be used for manipulation test plot. Default is evenly spaced.

`genvars(varname)` when provided, *varname* is used to create variables with the infor-

mation used to construct the manipulation test plot. See help file for details.

`level(a)` specifies the level of confidence intervals for the manipulation test plot. Default is 95.

`graph_options(...)` options passed to the plot command.

## 3.2 Description

`rddensity` provides an implementation of manipulation tests using local polynomial density estimators. The user must specify the running variable. This command permits fully data-driven inference by employing the companion command `rdbwdensity`, which may also be used as a stand-alone command.

## 4 rdbwdensity syntax

This section describes the syntax of the command `rdbwdensity`. This command implements the different bandwidth selection procedures for manipulation tests based on local polynomial density estimators discussed above.

### 4.1 Syntax

```
rdbwdensity runvar [if] [in] [, c(cutoff) p(pvalue)
    fitselect(fitmethod) kernel(kernelfn) vce(vcemethod) ]
```

*runvar* is the running variable (a.k.a. score or forcing variable).

*cutoff* is the RD cutoff. Default is  $\bar{x} = 0$ .

*pvalue* is the order of the local-polynomial used to construct the point-estimator. Default is  $p = 2$  (local-cubic).

*fitmethod* specifies the model used for estimation and inference. Options are:

**unrestricted.** For density estimation without any restrictions (two-sample, unrestricted inference). This is the default option.

**restricted.** For density estimation assuming equal c.d.f. and higher-order derivatives.

*kernelfn* is the kernel function  $K(\cdot)$  used to construct the local-polynomial estimator(s).

Options are: **triangular**, **uniform** and **epanechnikov**. Default is **triangular**.

*vcemethod* specifies the procedure used to compute the variance-covariance matrix estimator. Options are:

**plugin.** For asymptotic plug-in standard errors.

**jackknife.** For jackknife standard errors. This is the default option.

## 4.2 Description

`rdbwdensity` implements several bandwidth selection procedures specifically tailored for manipulation testing based on the local polynomial density estimators implemented in `rddensity`. The user must specify the running variable.

## 5 Illustration of Methods

We illustrate our `Stata` commands employing the same dataset already used in [Calonico, Cattaneo, Farrell, and Titiunik \(2017b\)](#) and [Cattaneo, Titiunik, and Vazquez-Bare \(2016\)](#), where related RD commands are introduced and discussed. This facilitates the comparison across the `Stata` and `R` packages available for analysis and interpretation of RD designs: the package introduced herein discusses a falsification approach to RD designs via Manipulation Testing, while the other packages discuss graphical presentation and falsification, estimation and inference.

The dataset `rddensity_senate.dta` contains the running variable from a larger dataset constructed and studied in [Cattaneo, Frandsen, and Titiunik \(2015\)](#), focusing on party advantages in U.S. Senate elections for the period 1914–2010. Thus, the unit of observation is a state in the U.S. In this section, we focus on the running variable used to analyze the RD effect of the Democratic party winning a U.S. Senate seat on the vote share obtained in the following election for that same seat. This empirical illustration is analogous to the one presented by [McCrary \(2008\)](#) for U.S. House elections. The variable of interest is `margin`, which ranges from  $-100$  to  $100$  and records the Democratic party's margin of victory in the statewide election for a given U.S. Senate seat, defined as the vote share of the Democratic party minus the vote share of its strongest opponent. This corresponds to the index, score or running variable. By construction, the cutoff is  $\bar{x} = 0$ .

First we load the database and present summary statistics:

```
. use rddensity_senate.dta, clear
. sum margin
```

Variable	Obs	Mean	Std. Dev.	Min	Max
margin	1,390	7.171159	34.32488	-100	100

The dataset has a total of 1,390 observations, with an average Democratic party's margin of victory of about 7 percentage points.

We now conduct a Manipulation Test using the command `rddensity` with its default options.

```
. rddensity margin
Computing data-driven bandwidth selectors.
RD Manipulation Test using local polynomial density estimation.
```

	Cutoff c = 0	Left of c	Right of c		
				Number of obs =	1390
				Model =	unrestricted
				BW method =	comb
				Kernel =	triangular
				VCE method =	jackknife
Number of obs		640	750		
Eff. Number of obs		408	460		
Order est. (p)		2	2		
Order bias (q)		3	3		
BW est. (h)		19.841	27.119		

Running variable: margin.

Method	T	P> T
Robust	-0.8753	0.3814

The output of the command `rddensity` contains a variety of useful information. First, the upper-left panel gives basic summary statistics on the data being used, separate for control ( $X_i < \bar{x}$ ) and treatment units ( $X_i \geq \bar{x}$ ). This panel also reports the value of the bandwidth(s) chosen. Second, the upper-right panel includes general information regarding the overall sample size and implementation choices of the manipulation test. Finally, the lower panel reports the results from implementing the manipulation test. In this first execution, the test statistic is constructed using a  $q = 3$  polynomial, with different bandwidths chosen for an unrestricted model with polynomial order  $p = 2$ . Specifically, the bandwidth choices are:  $(\hat{h}_{1,\text{comb},p}, \hat{h}_{x,\text{comb},p}) = (19.841, 27.119)$ , leading to effective sample sizes of  $N_- = 408$  and  $N_+ = 460$  for control and treatment groups, respectively. The final manipulation test is  $T_q(\hat{h}_{1,\text{comb},p}, \hat{h}_{x,\text{comb},p}) = -0.8753$ , with a p-value of 0.3814. Therefore, in this application there is no statistical evidence of systematic manipulation of the running variable.

The command `rddensity` also offers an automatic plot of the manipulation test. This plot is implemented using the package `lpdensity` for local polynomial based density estimation in `Stata` and `R`. If the user does not have this package installed, `rddensity` will issue an error and will request the user to install it. (In `R` the package `lpdensity` is declared as a dependency so no error ever occurs.) The following command installs the package `lpdensity` in `Stata`:

```
. net install lpdensity, from(https://sites.google.com/site/nppackages/lpdensity/stata)
replace
```

To obtain the default manipulation test plot, all it is needed is to add the option `plot` to `rddensity` as follows:

```
. rddensity margin, plot
Computing data-driven bandwidth selectors.
RD Manipulation Test using local polynomial density estimation.
      Cutoff c = 0 | Left of c   Right of c
-----|-----|-----
Number of obs     |       640     750
Eff. Number of obs |       408     460
Order est. (p)    |         2         2
Order bias (q)    |         3         3
EW est. (h)       |    19.841    27.119
Number of obs =   1390
Model =           unrestricted
BW method =       comb
Kernel =          triangular
VCE method =      jackknife

Running variable: margin.
-----|-----|-----
      Method |           T           P>|T|
-----|-----|-----
      Robust |    -0.8753         0.3814

. graph export output/figure1.pdf, replace
(file output/figure1.pdf written in PDF format)
```

The resulting default plot is given in Figure 1.

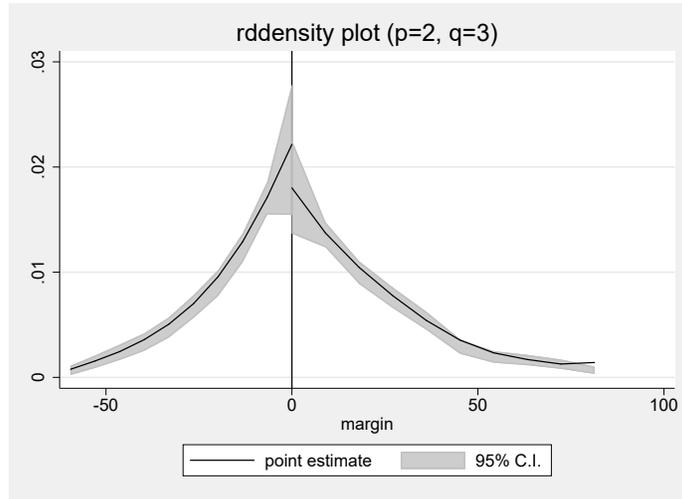


Figure 1: Manipulation Test Plot (default options).

The basic manipulation test plot can be improved using user-specified options. For example, the following command changes its legends and general appearance. The result plot is given in Figure 2.

```
. rddensity margin, plot ///
> graph_options(graphregion(color(white)) ///
> xtitle("Margin of Victory") ytitle("Density") legend(off))
Computing data-driven bandwidth selectors.
RD Manipulation Test using local polynomial density estimation.
Cutoff c = 0 | Left of c | Right of c | Number of obs = 1390
Model = unrestricted
BW method = comb
Kernel = triangular
VCE method = jackknife
Number of obs | 640 | 750
Eff. Number of obs | 408 | 460
Order est. (p) | 2 | 2
Order bias (q) | 3 | 3
BW est. (h) | 19.841 | 27.119
Running variable: margin.
Method | T | P>|T|
Robust | -0.8753 | 0.3814
. graph export output/figure2.pdf, replace
(file output/figure2.pdf written in PDF format)
```

To further illustrate some of the capabilities of our main function `rddensity` we consider a few additional runs. We can obtain two distinct statistics, conventional and bias-corrected, by including the option `all`. This gives the following output:

```
. rddensity margin, all
Computing data-driven bandwidth selectors.
RD Manipulation Test using local polynomial density estimation.
Cutoff c = 0 | Left of c | Right of c | Number of obs = 1390
Model = unrestricted
```

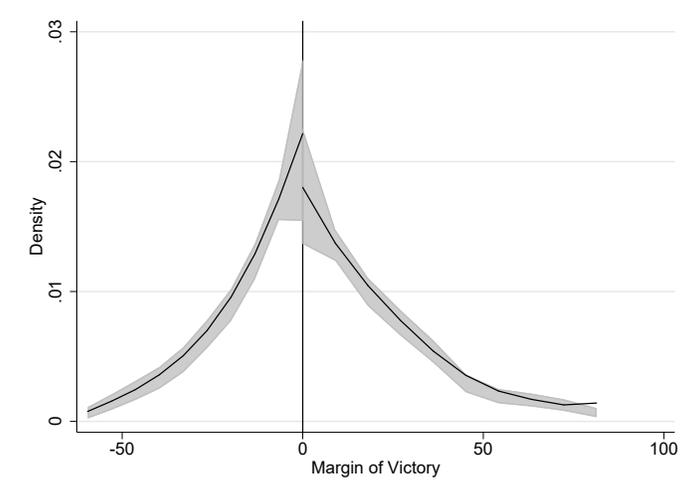


Figure 2: Manipulation Test Plot (with user options).

```

Number of obs      |      640      750      BW method   =      comb
Eff. Number of obs |      408      460      Kernel      =      triangular
Order est. (p)     |          2      2      VCE method  =      jackknife
Order bias (q)     |          3      3
BW est. (h)        |     19.841    27.119
    
```

Running variable: margin.

Method	T	P> T
Conventional	-1.6506	0.0988
Robust	-0.8753	0.3814

This second output still employs all the default options, but not it reports two test statistics. The first statistic, labeled “Conventional”, will exhibit asymptotic bias and hence will over-reject the null hypothesis of no manipulation when the MSE-optimal or other “large” bandwidth is used. This is confirmed in the simulation study reported in Section 6. The second statistic, labeled “Robust”, implements inference based on robust bias-correction and is the default and recommended option for implementing a manipulation test.

The following output showcases some other features of `rddensity`. Here we conduct a manipulation test using the restricted model and plug-in standard errors:

```

. rddensity margin, fitselect(restricted) vce(plugin)
Computing data-driven bandwidth selectors.
RD Manipulation Test using local polynomial density estimation.
Cutoff c = 0 | Left of c   Right of c   Number of obs =      1390
              |             |             | Model         =      restricted
              |             |             | BW method     =      comb
              |             |             | Kernel        =      triangular
              |             |             | VCE method    =      plugin
-----+-----+-----+-----+-----+-----+-----+-----+-----+
Number of obs |      640      750      1390
Eff. Number of obs |      396      362      1390
Order est. (p)   |          2      2      2
Order bias (q)  |          3      3      3
    
```

BW est. (h)   18.753 18.753		
Running variable: margin.		
Method	T	P> T
Robust	-1.4768	0.1397

In this case, because the restricted model is being used, a common bandwidth for both control and treatment units is selected. This value is  $\hat{h}_{\text{comb},p}^R = 18.753$  with the choice  $p = 2$  and employing now the plug-in standard errors estimator (instead of the jackknife method, as used before). This empirical finding shows that we continue to not reject the null hypothesis of no manipulation (p-value is 0.1397), even when the restricted model is used, which provides further empirical evidence in favor of the validity of the RD design in this application.

To close this section, we also report an output from the companion command `rdbdensity`. This command was used all along implicitly by `rddensity`, but here we employ it as a stand-alone command to illustrate some of its features. The default output for the empirical applications is as follows:

```
. rdbdensity margin
Bandwidth selection for manipulation testing.
Cutoff c = 0.000 | Left of c Right of c
Number of obs = 1390
Model = unrestricted
Kernel = triangular
VCE method = jackknife
```

	Left of c	Right of c
Number of obs	640	750
Min Running var.	-100.000	0.011
Max Running var.	-0.079	100.000
Order loc. poly. (p)	2	2

```
Running variable: margin.
```

Target	Bandwidth	Variance	Bias <sup>2</sup>
left density	19.841	0.109	0.000
right density	27.569	0.085	0.000
difference densities	27.119	0.194	0.000
sum densities	19.531	0.194	0.000

The output of `rdbdensity` mimics as close as possible the one from `rddensity`. Notice that the defaults are all the same (e.g.,  $\bar{x} = 0$ ,  $p = 2$ ,  $K(\cdot) = \text{triangular}$ , and unrestricted model). The main results are reported in the lower panel, where now the output includes different estimated bandwidth choices. These choices depend on the MSE criterion function (and the model considered, as discussed previously): the first row (labeled “left density”) reports  $\hat{h}_{l,p}$ , the second row (labeled “right density”) reports  $\hat{h}_{r,p}$ , the third row (labeled “difference densities”) reports  $\hat{h}_{\text{diff},p}$ , and the fourth row (labeled “sum densities”) reports  $\hat{h}_{\text{sum},p}$ . Of course,  $\hat{h}_{\text{comb},p}$  may be easily constructed using the above information.

A similar set of results may also be obtained for the restricted model. For this case the command is `rdbdensity margin, fitselect(restricted)`, but we do not report these results here to conserve space.

Finally, we briefly illustrate how the two commands can be combined:

```
. qui rdbdensity margin
```

```

. mat h = e(h)
. local hr = h[2,1]
. rddensity margin, h(10 `hr`)
RD Manipulation Test using local polynomial density estimation.
      Cutoff c = 0 | Left of c   Right of c
-----+-----+-----
Eff. Number of obs |         640       750
Order est. (p)     |          251       464
Order bias (q)     |           2         2
      BW est. (h)   |        10.000     27.569
Running variable: margin.
-----+-----+-----
      Method |           T           P>|T|
-----+-----+-----
      Robust |       -1.0331         0.3016

```

In this last example, first `rdbwdensity` is used to estimate the bandwidths quietly, but then the left bandwidth is set manually ( $h_l = 10$ ) while the right bandwidth is estimated ( $h_r = 27.569$ ) when executing `rddensity`.

The companion replication file (`rddensity_illustration.do`) includes the syntax of all the examples discussed above, as well as additional examples not included here to conserve space. These examples are:

1. `rddensity margin, kernel(uniform)`  
Manipulation testing using uniform kernel.
2. `rddensity margin, bwselect(diff)`  
Manipulation testing with bandwidth selection based on MSE of difference of densities.
3. `rddensity margin, h(10 15)`  
Manipulation testing using user-chosen bandwidths  $h_l = 10$  and  $h_r = 15$ .
4. `rddensity margin, p(2) q(4)`  
Manipulation testing using  $p = 2$  and  $q = 4$ .
5. `rddensity margin, c(5) all`  
Manipulation testing at cutoff  $\bar{x} = 5$  with all statistics.
6. `rdbwdensity margin, p(3) fitselect(restricted)`  
Bandwidth selection for restricted model with  $p = 3$ .
7. `rdbwdensity margin, kernel(uniform) vce(jackknife)`  
Bandwidth selection with uniform kernel and jackknife standard error.

## 6 Simulation Study

This section reports the numerical findings from a simulation study aimed to illustrate the finite sample performance of our new manipulation test.

We consider several implementations of the manipulation test  $T_p(h)$ , which varies according to the choice of polynomial order ( $p = 2$  or  $p = 3$ ) and bandwidth choice. We analyze the performance of the inference procedure using a grid of bandwidths around the MSE-optimal choice, as well as a data-driven implementation of this optimal bandwidth choice, allowing for both equal and difference bandwidth choices on either side of the threshold. We also investigate the performance of robust bias-correction, as we report both the naïve testing procedure without bias correction and its robust bias-corrected version. For implementation, in all case, we also consider both asymptotic plug-in or jackknife variance estimation.

We consider the data generating process (DGP) given by  $X_i \sim \sqrt{3/5}\mathcal{T}(5)$ , where  $\mathcal{T}(k)$  denotes a Student's  $t$  distribution with  $k$  degrees of freedom, and a cutoff point  $\bar{x} = 0.5$ , which induces an asymmetric distribution. We employ a sample size  $n = 1,000$ , and all simulations were based on 2,000 replications with a triangular kernel to implement our density estimators.

## 6.1 Point Estimation and Empirical Size

In this section, we report simulation results concerning the point estimation properties of underlying density estimator at the boundary point  $\bar{x}$  as well as the average rejection rate (empirical size) of our proposed manipulation test. All the numerical results are given in Table 1.

We first describe the format of Table 1. Starting with the columns, each table reports:

(i) the bandwidths used to construct the density estimators on the left and on the right of the cutoff  $\bar{x}$  (columns under label “Bandwidth”);

(ii) the average bias of the two density estimators and difference thereof, simulation variability, standardized bias of the difference of density estimators and simulation MSE of the difference of density estimators at the cutoff  $\bar{x}$  (columns under label “Density Estimators”);

(iii) the average of the asymptotic plug-in standard error estimator and the empirical size of the associated feasible manipulation test (columns under label “Plug-in SE”); and

(iv) the average of the jackknife standard error estimator and the empirical size of the associated feasible manipulation test (columns under label “Jackknife SE”).

Continuing with the rows, and to better understand the role of the bandwidth choice  $h_n$  on the finite sample performance of the density and standard error estimators and of the manipulation test, Table 1 reports:

(i) a grid of fixed bandwidths constructed around the theoretical infeasible MSE-optimal bandwidth (rows under label “Bandwidth Grid”);

(ii) estimated bandwidths, which are allowed to differ on the two sides of the cutoff,

obtained as  $\hat{h}_{1,\text{comb},p}$  and  $\hat{h}_{\tau,\text{comb},p}$  (rows under label “Est.  $\hat{h}_-, \hat{h}_+$ ”); and

(iii) estimated bandwidths, which are required to be the same on the two sides of the cutoff, obtained as the smaller of  $\hat{h}_{\text{diff},p}$  and  $\hat{h}_{\text{sum},p}$  (rows under label “Est.  $\hat{h}_- = \hat{h}_+$ ”).

The simulation results allows us to explore the finite sample performance of both the different ingredients entering the manipulation test (density, standard error, and bandwidth estimators) and the test itself (and by implication the quality of the Gaussian distributional approximation). Given the large amount of information, we only offer a summary of the main results we observe from the Monte Carlo experiment.

We first look at non-random bandwidths, which helps separate the finite sample performance of the main theoretical results in the paper from the impact of bandwidth estimation. From the grid of bandwidths around the MSE-optimal  $h_{\text{MSE},p}$ , we find that (i) the simulation variability of the difference of density estimators (column labeled “sd”) is approximated very well by the jackknife standard error estimator and reasonably well by the asymptotic plug-in standard error estimator (compare to columns labeled “mean”), and (ii) the manipulation test exhibits some empirical size distortion when using the MSE-optimal bandwidth  $h_{\text{MSE},p}$ , as expected, but excellent empirical size when undersmoothing this bandwidth choice. These numerical findings indicate that our main theoretical results concerning bias, variance and distributional approximations, as well as the consistency of the proposed standard error formulas, are borne out in the Monte Carlo experiment.

We now explore the impact of bandwidth selection on estimation and inference in the context of the manipulation test. We focus on the last six rows of Table 1. When looking at the different bandwidth estimators, we find that (i) our bandwidth estimator  $\hat{h}_p$  tends to deliver smaller values than  $h_{\text{MSE},p}$  on average, a finding that actually helps control the empirical size of the manipulation test, and (ii) the robust bias-correction approach (i.e., inference based on  $T_p(\hat{h}_{p-1})$ ) delivers manipulation tests with very good empirical size properties. Because the robust bias-correction approach is theoretically justified and valid for all sample sizes, we recommend it as the best alternative for applications.

To summarize, based on the simulation evidence we obtained, we recommend for empirical work the manipulation test based on the feasible statistic  $T_p(\hat{h}_{p-1})$  constructed using a  $p$ -th order local polynomial density estimator, a MSE-optimal bandwidth choice coupled with robust bias-correction, and the corresponding jackknife standard error estimator. This testing procedure exhibited close-to-correct empirical size across all designs we considered, and performed as well as (if not better) than all the alternatives we explored. These numerical findings are in agreement with our main theoretical results.

## 6.2 Empirical Power

To complement the numerical results presented above, we also investigate the the power of the manipulation test constructed using the local polynomial density estimator. We

continue to employ the same DGP, but now we scale the data so that the true density satisfies

$$f_+/f_- \in \{0.5, 0.6, 0.7, 0.8, 0.9, 1.0, 1.1, 1.2, 1.3, 1.4, 1.5\}$$

at the cutoff  $\bar{x} = 0.5$ . Recall that the sample size is  $n = 1,000$  and all simulations are based on 2,000 replications.

The results are given in Table 2. We first describe the format of this table. The two main columns report:

(i) the bandwidths used to construct the density estimators on the left and on the right of the cutoff  $\bar{x}$  (columns under label “Bandwidth”); and

(ii) the density discontinuity and the corresponding rejection rates (columns under label “Rejection Rate”). Note that the column  $f_+/f_- = 1$  corresponds to the null hypothesis being true, and the rejection rate under that column is just the empirical size of the test.

Continuing with the rows of Table 2, we examine the empirical power with the infeasible MSE-optimal bandwidth as well as the estimated bandwidths, with either the plug-in or the jackknife standard error employed.

Based on the simulation evidence obtained, we find that using the infeasible MSE-optimal bandwidth will lead to size distortion and the power curve does not attain minimum when the null hypothesis is true (for example in the  $p = 2$  case in Table 2, the minimum rejection rate occurs when  $f_+/f_- = 1.2$ ), which again confirms the finding that without bias correction, the manipulation test will not only be inconsistent but also lose power.

For the different bandwidth estimators, we again find that the robust bias-correction approach delivers manipulation tests with very good empirical size properties, since with either method, the power curve achieves minimum when the null hypothesis is true. Also as expected, bias correction will lead to some power loss compared with the  $T_p(\hat{h}_p)$  case.

## 7 Conclusion

This article discussed the implementation of nonparametric manipulation tests using local polynomial density estimators, which are useful both for falsification of RD designs as well as for empirical research analyzing whether units are self-selection into a particular group or treatment status. We introduced two commands: `rddensity` and `rdbwdensity`. These commands employ ideas from Cattaneo, Jansson, and Ma (2017a). In particular, the first command implements several nonparametric manipulation tests, while the second command provides an array of bandwidth selection methods. Companion R functions are also available from the authors. The latest version of this and related software for RD designs can be found at: <https://sites.google.com/site/rdpackages/>.

## 8 Acknowledgments

We thank Sebastian Calonico, David Drukker, Rocio Titiunik and Gonzalo Vazquez-Bare for useful comments that improved this manuscript as well as our implementations. The first author gratefully acknowledge financial support from the National Science Foundation through grants SES-1357561 and SES-1459931. The second author gratefully acknowledge financial support from the National Science Foundation through grant SES-1459967 and the research support of *CREATES* (funded by the Danish National Research Foundation under grant no. DNR78).

## 9 About the authors

Matias D. Cattaneo is a Professor of Economics and a Professor of Statistics at the University of Michigan.

Michael Jansson is the Edward G. and Nancy S. Jordan Family Professor of Economics at the University of California, Berkeley.

Xinwei Ma is a Ph.D. Candidate in Economics at the University of Michigan.

## 10 References

- Calonico, S., M. D. Cattaneo, and M. H. Farrell. 2017a. On the Effect of Bias Estimation on Coverage Accuracy in Nonparametric Inference. *Journal of the American Statistical Association*, forthcoming .
- Calonico, S., M. D. Cattaneo, M. H. Farrell, and R. Titiunik. 2017b. `rdrobust`: Software for Regression Discontinuity Designs. *Stata Journal* 17(2): 372–404.
- Calonico, S., M. D. Cattaneo, and R. Titiunik. 2014a. Robust Data-Driven Inference in the Regression-Discontinuity Design. *Stata Journal* 14(4): 909–946.
- . 2014b. Robust Nonparametric Confidence Intervals for Regression-Discontinuity Designs. *Econometrica* 82(6): 2295–2326.
- . 2015a. Optimal Data-Driven Regression Discontinuity Plots. *Journal of the American Statistical Association* 110(512): 1753–1769.
- . 2015b. `rdrobust`: An R Package for Robust Nonparametric Inference in Regression-Discontinuity Designs. *R Journal* 7(1): 38–51.
- Cattaneo, M. D., and J. C. Escanciano. 2017. *Regression Discontinuity Designs: Theory and Applications (Advances in Econometrics, volume 38)*. Emerald Group Publishing.
- Cattaneo, M. D., B. Frandsen, and R. Titiunik. 2015. Randomization Inference in the Regression Discontinuity Design: An Application to Party Advantages in the U.S. Senate. *Journal of Causal Inference* 3(1): 1–24.

- Cattaneo, M. D., M. Jansson, and X. Ma. 2017a. Simple Local Polynomial Density Estimators. working paper, University of Michigan .
- . 2017b. `lpdensity`: Local Polynomial Density Estimation and Inference. working paper, University of Michigan .
- Cattaneo, M. D., R. Titiunik, and G. Vazquez-Bare. 2016. Inference in Regression Discontinuity Designs under Local Randomization. *Stata Journal* 16: 331–367.
- . 2017c. Comparing Inference Approaches for RD Designs: A Reexamination of the Effect of Head Start on Child Mortality. *Journal of Policy Analysis and Management* 36(3): 643–681.
- Cheng, M.-Y., J. Fan, and J. S. Marron. 1997. On Automatic Boundary Corrections. *Annals of Statistics* 25(4): 1691–1708.
- Fan, J., and I. Gijbels. 1996. *Local Polynomial Modelling and Its Applications*. New York: Chapman & Hall/CRC.
- Frandsen, B. 2017. Party Bias in Union Representation Elections: Testing for Manipulation in the Regression Discontinuity Design When the Running Variable is Discrete. In *Regression Discontinuity Designs: Theory and Applications (Advances in Econometrics, volume 38)*, ed. M. D. Cattaneo and J. C. Escanciano, 29–72. Emerald Group Publishing.
- McCrary, J. 2008. Manipulation of the running variable in the regression discontinuity design: A density test. *Journal of Econometrics* 142(2): 698–714.
- Otsu, T., K.-L. Xu, and Y. Matsushita. 2014. Estimation and Inference of Discontinuity in Density. *Journal of Business and Economic Statistics* 31(4): 507–524.

Table 1: Density Estimation and Manipulation Test (Empirical Size)

	Bandwidth		Density Estimators						Plug-in SE		Jackknife SE	
	left	right	bias <sub>-</sub>	bias <sub>+</sub>	bias	sd	bias/sd	mse	mean	size	mean	size
<b>Grid <math>h_-, h_+</math></b>												
0.5×	0.538	0.538	0.014	-0.001	-0.015	0.092	0.163	0.816	0.091	0.050	0.091	0.050
0.6×	0.645	0.645	0.021	-0.003	-0.024	0.084	0.292	0.721	0.083	0.056	0.083	0.059
0.7×	0.753	0.753	0.031	-0.005	-0.036	0.078	0.460	0.699	0.078	0.077	0.077	0.078
0.8×	0.861	0.861	0.042	-0.007	-0.049	0.073	0.666	0.735	0.073	0.107	0.072	0.112
0.9×	0.968	0.968	0.054	-0.010	-0.064	0.069	0.921	0.832	0.069	0.143	0.068	0.152
$h_{MSE,p}$	1.076	1.076	0.067	-0.013	-0.080	0.065	1.223	1.000	0.066	0.208	0.064	0.232
1.1×	1.183	1.183	0.081	-0.016	-0.097	0.062	1.567	1.248	0.064	0.312	0.061	0.345
1.2×	1.291	1.291	0.095	-0.019	-0.114	0.059	1.939	1.565	0.061	0.468	0.059	0.497
1.3×	1.399	1.399	0.109	-0.023	-0.132	0.056	2.333	1.940	0.059	0.609	0.056	0.646
1.4×	1.506	1.506	0.121	-0.027	-0.148	0.054	2.737	2.363	0.057	0.747	0.054	0.785
1.5×	1.614	1.614	0.133	-0.031	-0.165	0.052	3.144	2.820	0.056	0.856	0.053	0.876
<b>Est. <math>\hat{h}_-, \hat{h}_+</math></b>												
$T_p(\hat{h}_p)$	0.618	0.710	0.023	-0.004	-0.027	0.089	0.303	0.827	0.083	0.084	0.083	0.091
$T_p(\hat{h}_{p-1})$	0.254	0.218	0.014	-0.001	-0.015	0.142	0.103	1.928	0.141	0.054	0.143	0.050
$T_{p+1}(\hat{h}_p)$	0.618	0.710	0.007	0.005	-0.002	0.130	0.017	1.593	0.130	0.052	0.131	0.046
<b>Est. <math>\hat{h}_- = \hat{h}_+</math></b>												
$T_p(\hat{h}_p)$	0.596	0.596	0.021	-0.002	-0.023	0.094	0.248	0.880	0.088	0.080	0.087	0.085
$T_p(\hat{h}_{p-1})$	0.187	0.187	0.009	0.000	-0.008	0.159	0.053	2.395	0.157	0.059	0.160	0.052
$T_{p+1}(\hat{h}_p)$	0.596	0.596	0.005	0.005	-0.001	0.137	0.005	1.776	0.136	0.056	0.137	0.052

(a)  $p = 2$

	Bandwidth		Density Estimators						Plug-in SE		Jackknife SE	
	left	right	bias <sub>-</sub>	bias <sub>+</sub>	bias	sd	bias/sd	mse	mean	size	mean	size
<b>Grid <math>h_-, h_+</math></b>												
0.5×	0.604	0.604	0.003	0.003	0.000	0.133	0.001	1.948	0.134	0.048	0.135	0.046
0.6×	0.725	0.725	0.001	0.004	0.004	0.121	0.030	1.607	0.122	0.045	0.124	0.047
0.7×	0.845	0.845	-0.004	0.005	0.009	0.112	0.082	1.378	0.113	0.052	0.114	0.048
0.8×	0.966	0.966	-0.007	0.007	0.013	0.105	0.128	1.219	0.106	0.048	0.107	0.044
0.9×	1.087	1.087	-0.007	0.008	0.015	0.099	0.153	1.103	0.100	0.050	0.101	0.046
$h_{MSE,p}$	1.208	1.208	-0.006	0.009	0.015	0.094	0.156	1.000	0.095	0.052	0.095	0.049
1.1×	1.328	1.328	-0.003	0.009	0.012	0.090	0.139	0.903	0.090	0.056	0.091	0.054
1.2×	1.449	1.449	0.002	0.010	0.008	0.086	0.094	0.816	0.087	0.052	0.087	0.048
1.3×	1.570	1.570	0.008	0.010	0.002	0.083	0.020	0.746	0.084	0.048	0.083	0.053
1.4×	1.691	1.691	0.016	0.010	-0.007	0.080	0.082	0.697	0.081	0.050	0.080	0.050
1.5×	1.812	1.812	0.026	0.009	-0.016	0.077	0.211	0.676	0.079	0.048	0.077	0.053
<b>Est. <math>\hat{h}_-, \hat{h}_+</math></b>												
$T_p(\hat{h}_p)$	1.190	1.237	-0.006	0.008	0.014	0.098	0.140	1.077	0.095	0.051	0.097	0.050
$T_p(\hat{h}_{p-1})$	0.610	0.705	0.010	0.005	-0.005	0.129	0.040	1.830	0.131	0.048	0.132	0.046
$T_{p+1}(\hat{h}_p)$	1.190	1.237	0.007	0.006	-0.001	0.129	0.007	1.818	0.134	0.044	0.135	0.040
<b>Est. <math>\hat{h}_- = \hat{h}_+</math></b>												
$T_p(\hat{h}_p)$	1.119	1.119	-0.006	0.008	0.014	0.100	0.136	1.124	0.099	0.048	0.100	0.051
$T_p(\hat{h}_{p-1})$	0.590	0.590	0.009	0.001	-0.007	0.136	0.055	2.043	0.137	0.050	0.138	0.052
$T_{p+1}(\hat{h}_p)$	1.119	1.119	0.008	0.006	-0.002	0.134	0.016	1.960	0.140	0.043	0.140	0.038

(b)  $p = 3$

**Notes:** (i) columns “Bandwidth”: bandwidths for left and right density estimators; (ii) columns “Density Estimators”: bias of left and right density estimators, and bias, standard deviation, standardized bias and mean squared error of difference of density estimators; (iii) columns “Plug-in SE”: average of plug-in standard error (“mean”) and empirical size of corresponding manipulation test  $T_p(h)$  (“size”); (iv) columns “Plug-in SE” report average of jackknife standard error (“mean”) and empirical size of corresponding manipulation test  $T_p(h)$  (“size”); (v)  $\hat{h}_p$  denotes estimated MSE-optimal bandwidth for  $p$ -order density estimator.

Table 2: Manipulation Test (Empirical Power).

	Bandwidth		Rejection Rate					$H_0 : f_+/f_- = 1$ vs. $H_a$ : value below						
	left	right	0.5	0.6	0.7	0.8	0.9	1	1.1	1.2	1.3	1.4	1.5	
<b>Plug-in SE</b>														
$T_p(h_{MSE,p})$	1.076	1.076	1.000	0.990	0.922	0.734	0.454	0.212	0.080	0.036	0.060	0.126	0.233	
<b>Est. <math>\hat{h}_-, \hat{h}_+</math></b>														
$T_p(\hat{h}_p)$	0.615	0.712	0.888	0.715	0.498	0.284	0.154	0.072	0.064	0.078	0.134	0.218	0.330	
$T_p(\hat{h}_{p-1})$	0.254	0.215	0.478	0.316	0.202	0.122	0.074	0.056	0.049	0.060	0.072	0.094	0.130	
$T_{p+1}(\hat{h}_p)$	0.615	0.712	0.474	0.313	0.191	0.112	0.074	0.053	0.050	0.065	0.100	0.143	0.187	
<b>Est. <math>\hat{h}_- = \hat{h}_+</math></b>														
$T_p(\hat{h}_p)$	0.594	0.594	0.865	0.686	0.456	0.266	0.138	0.070	0.060	0.074	0.122	0.190	0.285	
$T_p(\hat{h}_{p-1})$	0.185	0.185	0.353	0.239	0.151	0.099	0.066	0.050	0.047	0.060	0.079	0.100	0.131	
$T_{p+1}(\hat{h}_p)$	0.594	0.594	0.458	0.301	0.188	0.114	0.072	0.057	0.055	0.068	0.088	0.117	0.159	
<b>Jackknife SE</b>														
$T_p(h_{MSE,p})$	1.076	1.076	1.000	0.990	0.926	0.747	0.480	0.232	0.090	0.045	0.071	0.148	0.265	
<b>Est. <math>\hat{h}_-, \hat{h}_+</math></b>														
$T_p(\hat{h}_p)$	0.615	0.712	0.866	0.700	0.490	0.288	0.155	0.076	0.067	0.082	0.136	0.231	0.352	
$T_p(\hat{h}_{p-1})$	0.254	0.215	0.459	0.309	0.196	0.114	0.063	0.047	0.044	0.055	0.070	0.098	0.139	
$T_{p+1}(\hat{h}_p)$	0.615	0.712	0.450	0.305	0.196	0.118	0.068	0.051	0.047	0.062	0.104	0.143	0.200	
<b>Est. <math>\hat{h}_- = \hat{h}_+</math></b>														
$T_p(\hat{h}_p)$	0.594	0.594	0.844	0.670	0.450	0.266	0.142	0.073	0.063	0.076	0.123	0.202	0.310	
$T_p(\hat{h}_{p-1})$	0.185	0.185	0.345	0.230	0.148	0.097	0.057	0.043	0.044	0.054	0.072	0.100	0.132	
$T_{p+1}(\hat{h}_p)$	0.594	0.594	0.434	0.292	0.188	0.114	0.074	0.052	0.050	0.066	0.088	0.126	0.170	

(a)  $p = 2$ 

	Bandwidth		Rejection Rate					$H_0 : f_+/f_- = 1$ vs. $H_a$ : value below						
	left	right	0.5	0.6	0.7	0.8	0.9	1	1.1	1.2	1.3	1.4	1.5	
<b>Plug-in SE</b>														
$T_p(h_{MSE,p})$	1.208	1.208	0.701	0.456	0.236	0.110	0.056	0.057	0.086	0.143	0.218	0.311	0.412	
<b>Est. <math>\hat{h}_-, \hat{h}_+</math></b>														
$T_p(\hat{h}_p)$	1.205	1.259	0.680	0.455	0.264	0.138	0.076	0.057	0.090	0.141	0.210	0.306	0.409	
$T_p(\hat{h}_{p-1})$	0.617	0.712	0.472	0.291	0.168	0.098	0.064	0.050	0.055	0.078	0.108	0.152	0.190	
$T_{p+1}(\hat{h}_p)$	1.205	1.259	0.478	0.280	0.149	0.078	0.051	0.050	0.060	0.076	0.106	0.137	0.179	
<b>Est. <math>\hat{h}_- = \hat{h}_+</math></b>														
$T_p(\hat{h}_p)$	1.134	1.134	0.658	0.442	0.240	0.124	0.066	0.052	0.084	0.122	0.191	0.272	0.368	
$T_p(\hat{h}_{p-1})$	0.594	0.594	0.447	0.284	0.166	0.102	0.068	0.051	0.056	0.070	0.092	0.122	0.162	
$T_{p+1}(\hat{h}_p)$	1.134	1.134	0.450	0.263	0.148	0.075	0.052	0.052	0.056	0.070	0.098	0.129	0.162	
<b>Jackknife SE</b>														
$T_p(h_{MSE,p})$	1.208	1.208	0.646	0.424	0.228	0.107	0.057	0.054	0.083	0.150	0.236	0.336	0.446	
<b>Est. <math>\hat{h}_-, \hat{h}_+</math></b>														
$T_p(\hat{h}_p)$	1.205	1.259	0.624	0.430	0.256	0.134	0.074	0.053	0.086	0.142	0.232	0.336	0.448	
$T_p(\hat{h}_{p-1})$	0.617	0.712	0.444	0.286	0.174	0.102	0.063	0.041	0.054	0.072	0.112	0.152	0.208	
$T_{p+1}(\hat{h}_p)$	1.205	1.259	0.444	0.274	0.152	0.080	0.047	0.040	0.055	0.080	0.110	0.154	0.204	
<b>Est. <math>\hat{h}_- = \hat{h}_+</math></b>														
$T_p(\hat{h}_p)$	1.134	1.134	0.600	0.415	0.236	0.123	0.064	0.049	0.078	0.128	0.210	0.300	0.403	
$T_p(\hat{h}_{p-1})$	0.594	0.594	0.434	0.275	0.170	0.100	0.064	0.045	0.048	0.068	0.092	0.130	0.177	
$T_{p+1}(\hat{h}_p)$	1.134	1.134	0.420	0.267	0.148	0.078	0.047	0.042	0.053	0.072	0.101	0.140	0.184	

(b)  $p = 3$ 

**Notes:** (i) columns “Bandwidth”: bandwidths for left and right density estimators under the null hypothesis  $f_+/f_- = 1$ . For other cases the bandwidth for estimating  $f_+$  is adjusted proportional to  $f_+/f_-$ ; (ii)  $\hat{h}_p$ : estimated MSE-optimal bandwidth for  $p$ -order density estimator.