A Comparison of Different Data-Driven Procedures to Determine the Bunching Window

Vincent Dekker * Karsten Schweikert †

29th September 2020

Preliminary Draft

Do not cite or circulate without permission

In this paper, we compare three data-driven procedures to determine the bunching window in a Monte Carlo simulation of taxable income. Following the standard approach in the empirical bunching literature, we fit a flexible polynomial model to a simulated income distribution, excluding data in a range around a pre-specified kink. First, we propose to implement methods for the estimation of structural breaks to determine a bunching regime around the kink. A second procedure is based on Cook's distances aiming to identify outlier observations. Finally, we apply the iterative procedure proposed by Bosch et al. (2020) which evaluates polynomial counterfactual models for all possible bunching windows. While our simulation results show that all three procedures are fairly accurate, the Bosch et al. (2020) procedure is the preferred method to detect the bunching window when no prior information about the true size of the bunching window is available.

JEL Classification: C15, H00, J01

Keywords: Bunching Window, Cook's Distances, Monte Carlo Simulation, Structural Breaks, Taxable Income

*University of Hohenheim, Chair of Public Economics & Computational Science Lab, Schloss Hohenheim 1D, 70599 Stuttgart, Germany. Email: vincent.dekker@uni-hohenheim.de

†University of Hohenheim, Core Facility Hohenheim & Institute of Economics, Schloss Hohenheim 1C, 70599 Stuttgart, Germany. Email: karsten.schweikert@uni-hohenheim.de

1 Motivation

The bunching approach as introduced by Saez (2010) has been applied to a range of economic problems in recent years (see, among others, Chetty et al. (2011), le Maire and Schjerning (2013), Bastani and Selin (2014) and Seim (2017)).[1] The methodology exploits the bunching behaviour of subjects at kink points of a policy system. For example, kink points in the marginal tax rate or, as in the original paper by Saez (2010), the kink points of the Earned Income Tax Credit (EITC) system are utilised to elicit the elasticity of taxable income (ETI).[2] Due to the convex change in the budget set at the threshold, it is no longer optimal for some individuals – known as the bunching mass – to locate above the threshold. These individuals maximise their utility by moving to the threshold and hence, an increased mass will be observable at the threshold. The difference between this mass at the threshold and a counterfactual mass derived under the assumption of a smooth budget set is known as the excess mass. The excess mass can then be used to derive, for example, the ETI.

Although the bunching procedure has been acknowledged and replicated in many studies, some authors have found reason to criticise the approach. Einav et al. (2017) show that the bunching pattern can be matched by several alternative economic models, each with a different implication for the counterfactual model. Specifically, the authors find that the frictionless model as assumed in Saez (2010) could lead to elasticities that are up to five times lower than when utilising a richer dynamic model. This could partially explain the substantial difference between elasticities estimated via the bunching approach and methods that rely on instruments in the spirit of Gruber and Saez (2002), with the latter approach finding larger elasticities. A more fundamental critique with respect

[1] For a recent overview of the bunching literature, see Kleven (2016)

[2] An alternative model for notches in a tax system was developed by Kleven and Waseem (2013), but due to the existence of a strictly dominated region in their setting as opposed to the case of a kink, the identification technique for notches deviates in some respect and our paper only covers kink points.

to the identification of the ETI is developed separately in Bertanha et al. (2019) and Blomquist and Newey (2017), who argue that the structural parameter identified when utilising the bunching approach can be consistent with any ETI, when the distribution of heterogeneity in preferences is unrestricted. They also argue that the counterfactual distribution utilised in previous studies might not integrate to one, thus preventing it from actually being a distribution. The authors propose to either restrict the distribution of heterogeneity or utilise different budget sets to gain information on the distribution of preferences. An alternative estimator is suggested by Aronsson et al. (2018) to address the same issue. While most studies work within a model framework that avoids making assumptions about the unobservable income component and potential optimisation frictions, Aronsson et al. (2018) propose to estimate the ETI with a maximum likelihood estimator based on a log linear labor supply model with log-normally distributed unobserved income component. Although these assumptions can be restrictive, they can in principle be relaxed to capture more general settings allowing the authors to identify the ETI in those cases as well.

The above-mentioned papers criticise the translation of the bunching reaction into the structural parameter ETI and this discussion is beyond the scope of this paper. Here, we focus on a simplified setting and restrict our analysis to smooth counterfactual distributions. However, additional problems exist within the standard bunching framework itself, i.e., before the estimate of the excess mass is translated into the ETI. For example, Blomquist and Newey (2017) criticise the dependence on functional form assumptions and Bosch et al. (2020) argue that relying on eyeballing when determining the counterfactual model should be replaced by a data-driven procedure. We focus on improving the conventional bunching method as such and investigate the performance of different procedures to select the bunching window. We aim to provide recommendations for practitioners how these procedures can be used to make the specification of

the counterfactual model less reliant on researchers' subjective decisions.

Although theoretically all individuals should bunch exactly at the kink point, they face optimisation frictions such as adjustment costs and attention costs that prevent them from optimal adjustment. Yearly income is particularly difficult to predict perfectly for some individuals and this difficulty increases with more complicated tax codes. Therefore, empirically, a bunching window of some size around the threshold is observable (Chetty et al., 2011). The bulk of studies implementing the bunching approach relies on eyeballing to determine the bunching window and hedges against any concerns that this might drive results by altering the bunching window in robustness checks. We argue, in the spirit of Bosch et al. (2020), that the choice of the (only) appropriate bunching window is crucial to the unbiasedness of the bunching estimator and therefore requires a sound data-driven technique to identify it. The bunching window is directly linked to the estimation of the counterfactual model which helps to identify the relative excess mass (b). A bunching window that is too large encompasses individuals that do not bunch and leads to an overestimation of b, irrespective of the location of the kink point in the income distribution. Likewise, a bunching window that is too small leaves out bunching individuals and therefore underestimates the relative excess mass. The imprecision in the measurement of b translates directly into biased estimates of the elasticity.

To circumvent this problem, we discuss two new data-driven procedures and compare them to the data-driven procedure derived in Bosch et al. (2020). First, we propose to implement methods for the estimation of structural breaks following Bai and Perron (1998, 2003). When moving along the income distribution, the data behave in a certain way, noisy or smooth. At the start of the bunching window however, non-random deviations from this behaviour can be detected and these form the first structural break in the data. The second structural break is at that point, where the process reverts from the bunching window back to the unaffected region, i.e., a smooth income distribution.

Thus, two structural break points have to be estimated. A second procedure is based on Cook's distances that tries to identify influential observations and outliers (Cook, 1977). We identify the Cook's distances - scaled deviations from the least squares projected value - and say that the bunching window is made up of adjacent data points around the threshold that have a significantly positive distance. Third, we apply the iterative counterfactual procedure proposed by Bosch et al. (2020) which tests all possible combinations of excluded regions within a pre-determined interval. The procedure returns a bunching window for each excluded region, from which the mode is taken to determine the bunching window that is to be used for the estimation of the relative excess mass.

We compare these three data-driven procedures to determine the bunching window in a Monte-Carlo simulation of taxable income. In contrast to Aronsson et al. (2017) who use a behavioral model to draw data, we use a statistical model to draw simulated datasets which most closely mimic large administrative datasets. Our judgement on the methods is based on two separate questions: First, we compare the three methods with respect to the number of false negatives and false positives they produce. Second, we compare the estimates to a frictionless scenario (i.e. a known bunching "window" located exactly at the kink), where we study the behaviour of all three approaches with respect to optimisation frictions.

For the baseline specification, we find that the method based on estimation of structural breaks is able to identify most of the bunching individuals. In comparison, the other two procedures perform nearly identically and fail to detect nearly twice the amount of bunching individuals than the procedure by Bai and Perron. We can attribute the superiority of structural break testing methods in this respect to a very conservative estimation of the bunching window. Consequently, the percentage of individuals wrongly attributed as bunching is nearly twice as high as compared to implementing Cook's distances or the iterative counterfactual procedure. Variations in key parameters reveal a strong sensit-

ivity of the methods for the estimation of structural breaks. Especially when the true bunching window is narrow or very spread out, these methods fail to identify the correct bunching window. The other two procedures are overall more robust and only show sensitivity in the most extreme specifications. Cook's distances have difficulties determining the true bunching window especially with small sample sizes, large standard deviations and when the bunching mass is not spread symmetrically around the threshold. Therefore, we find that the procedure by Bosch et al. (2020) performs best for determining the size of the bunching window, although Cook's distances run it a close second, especially when considering its ease of implementation. Secondary analyses, relating the elasticity estimates to a baseline scenario without optimisation frictions, support the finding that the method proposed by Bosch et al. (2020) is superior in extreme cases.

In Section 2, we start by discussing the bunching methodology and highlighting the nature of the problem, before we present our three data-driven procedures that are evaluated in the Monte-Carlo simulations. The simulation design is described in Section 3, where we also discuss the outcome measures on which we base our judgement. Section 4 discusses our simulation results and provides an empirical illustration. Section 5 concludes.

2 Methodology

Our exposition of the bunching approach closely follows Kleven (2016) to illustrate the main ideas behind the bunching estimator. Although the bunching methodology has been applied in many fields of economics in recent years, its roots lie in the estimation of the ETI and therefore, we will focus our discussion on this application. However, our results are also relevant for other applications of the bunching approach, such as those discussed in Kleven (2016).

In the ETI case, we consider individuals with preferences defined over after-tax income

(value of consumption) and before-tax income (cost of effort). The utility function for those individuals is defined by $u(z - T(z), z/n)$, where z is earnings, $T(z)$ is a tax function, and n is ability (heterogenous over individuals). The baseline tax system is assumed to be linear $T(z) = \tau z$. We now consider the existence of a convex kink at the earnings threshold z^* at which the marginal tax rate increases from the lower tax rate τ_1 to the upper tax rate τ_2. The methodology exploits that individuals bunch at kink points to avoid paying the higher marginal tax rate. Following the seminal papers of Saez (2010) and Chetty et al. (2011), the mass of these taxpayers is used to identify the elasticity of taxable income. It can be shown that the excess mass (B) is approximately equal to the integral over the smooth counterfactual density ($h_0(z)$) between the kink point z^* and the earnings of the marginal buncher ($z^* + \Delta z^*$),

$$B = \int_{z^*}^{z^*+\Delta z^*} h_0(z)\,\mathrm{d}z \simeq h_0(z^*)\Delta z^*, \tag{1}$$

if the counterfactual density is constant on the bunching segment $(z^*, z^* + \Delta z^*)$. While the constant (or uniform) density assumption seems restrictive, it has to be emphasized that the assumption only has to hold over the usually small bunching segment and not over the full earnings distribution. It can further be relaxed by using a trapezoidal approximation (Saez, 2010) or other higher order approximations for the integral. However, note that in the presence of large kinks, we have to impose restrictions on the shape of the counterfactual density to identify the elasticity. For example, Bertanha et al. (2019) demonstrate that the elasticity is partially identified if one is willing to restrict the slope magnitude of the density. More specifically, they restrict the class of functions to those that are Lipschitz continuous with known constant $M \in (0, \infty)$.

Defining the relative excess mass as $b = \frac{B}{h_0(z^*)}$, the ETI at the earnings threshold z^*

can be written as

$$e(z^*) \simeq \frac{b}{z^* \cdot \ln(\frac{1-\tau_1}{1-\tau_2})} \tag{2}$$

where $\ln(\frac{1-\tau_1}{1-\tau_2})$ is the change in the net-of-tax rate at the earnings threshold. The denominator of (2) encompasses known policy parameters, so that the only parameter that needs to be estimated is b. However, this is made difficult by the fact that b is a composite of the excess mass B and the counterfactual density evaluated at the kink $h_0(z)$, and the excess mass can only then be estimated if the counterfactual distribution is known. While the constant density assumption works well in certain situation, namely at small kinks with little optimisation frictions, we specifically address the more general case with optimisation frictions. Consequently, we follow Chetty et al. (2011)'s approach and fit a higher-order polynomial to the observed income density excluding the bunching window. When the data are binned, B can be defined as the observed number of individuals within the bunching window less the number of individuals that would have been in that region in the absence of the kink point. Therefore, it is necessary to precisely estimate the counterfactual distribution h_0 over the bunching window.[3]

More specifically, the counterfactual income distribution around the kink point is estimated employing an auxiliary polynomial regression,

$$N_i = \sum_{j=0}^{q} \beta_j X_i^j + \sum_{w=l}^{u} \gamma_w \cdot I[X_i = w] + \epsilon_i, \qquad i = 1, \ldots, n, \tag{3}$$

where N_i denotes the number of individuals in each income bin, X_i is the midpoint of each income bin, n is the number of income bins and l and u define the lower and upper bound of the bunching window, respectively. After the order of the polynomial (q) is determined by the Bayesian Information Criterion (BIC), the predicted values for each income bin

[3]While the region to the left and to the right of the bunching window are used to estimate the counterfactual density, we primarily need a precise estimate of it over the bunching window, i.e., an out-of-sample prediction, to obtain b.

are determined using the polynomial coefficients β_j and approximate the counterfactual income distribution. Since polynomial regressions do not produce counterfactuals which integrate to one, this approach only provides a local approximation.

The outlined strategy is widely applied in the bunching literature (Bastani and Selin, 2014; Devereux et al., 2014; Best and Kleven, 2018), but faces some important issues: (i) the true shape of the counterfactual distribution in the bunching segment is unknown and we have to restrict the class of possible distributions to identify the bunching mass, (ii) the identification via polynomial regressions does not ensure that the density predicted by the polynomial regression integrates to one, and (iii) the bunching window is difficult to determine when optimisation frictions are present. One the one hand, Bertanha et al. (2019) discuss alternative estimation strategies to obtain unbiased estimates of the structural parameter, thereby focussing on the first two issues. On the other hand, Aronsson et al. (2018) choose a parametric approach with explicit distributional assumptions and estimate the ETI with maximum likelihood. Instead, we accept the shortcomings of the polynomial approach and aim to improve the estimation of our polynomial coefficients β_j, which in turn would lead to more precise estimates for the ETI if the necessary conditions hold (quasi-linear isoelastic utility function and smooth counterfactual distribution around the kink point). For this matter, it is crucial to know the exact bounds of the bunching window, i.e., which observations to exclude from the determination of the counterfactual model. Additionally, our approach can help us to identify cases in which the bunching window is too wide. In those cases, it might be difficult to uphold any strong local assumptions about the counterfactual density which are needed to ensure the unbiasedness of the bunching estimator. In the following, we discuss three data-driven procedures to identify the bounds of the bunching window.

2.1 Structural breaks

One way to identify the bunching window is to employ methods from the literature on stability in linear regression models. We assume a smooth population income distribution in the absence of bunching and model the bunching window as a separate regime in which the structure of the distribution changes. We determine the lower bound of the bunching window as a first structural break, or put differently, as the starting point of the second regime. Correspondingly, we determine the upper bound of the bunching window as a second structural break which marks the starting point of the third regime. For this matter, we sequentially add breakpoints to our polynomial regression model using dynamic programming to select the partition of the sample which optimizes the model fit. As these methods were described in several studies by Bai and Perron, we shall henceforth refer to this procedure as BP.

We follow Bai and Perron (1998) and consider a model with a maximum of two breaks and three regimes in which the coefficients remain constant. The regression equation for each regime is then given as

$$N_i = \sum_{j=0}^{q} \beta_j^k X_i^j + \epsilon_i, \qquad (i = i_{k-1} + 1, \ldots, i_k, \quad k = 1, 2, 3). \tag{4}$$

The coefficient estimates are obtained by minimizing the sum of squared residuals for each partition of the income bins with respect to β_j^k,

$$S_n(n_1, n_2) = \sum_{k=1}^{3} \sum_{i=n_{k-1}+1}^{n_k} \sum_{j=0}^{q} (N_i - \beta_j^k X_i^j)^2, \tag{5}$$

under the assumption that $n_0 = 0$ and $n_3 = n$. We denote the estimated coefficients for each partition as $\hat{\beta}_j(n_{k-1}, n_k)$. The estimated breakpoints $(\hat{n}_1, \hat{n}_2)$ are then obtained by substituting the coefficient estimates $\hat{\beta}_j(n_{k-1}, n_k)$ into the objective function and

minimizing over all partitions,

$$(\hat{n}_1, \hat{n}_2) = \underset{(n_1,n_2)}{\operatorname{argmin}} \; S_n(n_1, n_2). \tag{6}$$

The optimisation is restricted by the requirement that each regime contains sufficient observations so that the coefficients can be identified. The minimum number of observations in each regime is a function of the degree of the polynomial in our application. Bai and Perron (2003) discuss an efficient algorithm to obtain global minimisers of the sum of squared residuals in (5). For one threshold in the tax schedule, we specify the maximum number of breaks to be two and expect to find exactly two breakpoints. If the algorithm determines less than two breakpoints, we conclude that the bunching window is not correctly estimated or, put differently, that the algorithm does not detect a substantial bunching mass.

2.2 Outlier detection

Polynomial regressions are generally estimated by least squares. The regression coefficients are chosen so that the sum of squared errors (SSE) is minimised. Least squares estimation thus tries to avoid large distances between the predicted values and the observed values. Squaring the residuals gives proportionally more weight to extreme points in a sample. Hence, extreme points have a substantial influence on the placement of the polynomial regression curve. The sensitivity of the estimation procedure to extreme observations can be used to detect data points which do not follow the specified model structure, so-called outliers. In the following, we describe how outliers close to the threshold can be identified using Cook's distances.

Cook (1977) investigates the contribution of data points to the determination of the least squares estimate and proposes to evaluate the Cook's D statistics

$$D_i = \frac{(\hat{\boldsymbol{\beta}}_{(-i)} - \hat{\boldsymbol{\beta}})' \boldsymbol{X}'_K \boldsymbol{X}_K (\hat{\boldsymbol{\beta}}_{(-i)} - \hat{\boldsymbol{\beta}})}{MSE(K+1)}, \qquad i = 1, \ldots, n, \tag{7}$$

to indicate potentially critical observations. $\hat{\boldsymbol{\beta}}_{(-i)}$ denotes the least squares estimate of the $(K+1)\times 1$ coefficient vector $\boldsymbol{\beta}$ with the ith data point omitted, $\boldsymbol{X}_K$ is the $n \times (K+1)$ matrix containing all regressors including the constant, K indicates the degree of the polynomial and MSE denotes the mean square error of the regression. Alternatively, the statistic can be expressed in a more intuitive way (Kim et al., 2001) as

$$D_i = \frac{\hat{\epsilon}_i^2}{MSE(K+1)} \left(\frac{h_i}{(1-h_i)^2} \right), \tag{8}$$

where $\hat{\epsilon}_i^2$ is the squared residual i and h_i is the leverage of observation i. The leverage of the i-th observation can be computed as the i-th diagonal element of the projection matrix $H = \boldsymbol{X}_K (\boldsymbol{X}'_K \boldsymbol{X}_K)^{-1} \boldsymbol{X}'_K$. Cook's D statistic provides a measure to evaluate the influence of each observation on the placement of the regression curve as a function of the leverage and the size of the residual attached to this observation.

In practice, one needs to specify a cut-off value to decide whether observation i is considered an outlier observation. Since $D_i \sim F(K+1, n-K-1)$ under the normality assumption (Cook, 1977), we can use quantiles of the central F-distribution to construct cut-off values depending on how sensitive the outlier detection should be. For the main results in our simulation study, we use the following rule-of-thumb cut-off value[4]

$$D_i > \frac{4}{n-K-1}, \tag{9}$$

[4]Cook's D statistics are conceptually similar to the DFFITS statistics proposed in Belsley et al. (1980). Their relationship can be expressed by $D_i = \frac{1}{p} \frac{\hat{\sigma}^2_{(i)}}{\hat{\sigma}^2} DFFITS_i^2$ (Cook and Weisberg, 1982). This allows us to transform the cut-off value for DFFITS suggested in Belsley et al. (1980) and Bollen and Jackman (1990) and use it for Cook's D.

which corresponds roughly to the 1% confidence region of the D_i statistic. Naturally, the procedure is sensitive to the choice of cut-off value so that we additionally run simulations with 5% and 10% confidence regions to check our results for robustness. The Cook's D is computed for each observation, i.e. for each income bin. If we find a set of adjacent observations which are determined to be outliers and if this set includes the threshold, we take those income bins as an estimate for the bunching window.

2.3 Iterative counterfactual procedure

The iterative counterfactual procedure proposed by Bosch et al. (2020), henceforth referred to as BDS, tests all different combinations of excluded regions within a predetermined interval. For each excluded region, a polynomial counterfactual model is estimated, which minimises the BIC whilst omitting the data points within the excluded region. Based on this counterfactual, those subsequent observations around the threshold that lie outside the 95% confidence interval determine the actual bunching window. The procedure returns a bunching window for each excluded region. From this set of bunching windows, the mode is taken to determine the bunching window that is to be used for the estimation of the excess mass (Bosch et al., 2020).

Formally, the bunching window in Bosch et al. (2020) is derived as follows: Let $x_- \in \{-X, (-X+1), \ldots, 0\}$ and $x_+ \in \{0, 1, \ldots, X\}$ be the respective lower and upper bound of the excluded region, where X represents the midpoint of each income bin. The number of bins to be used as excluded region is set to 20 in our application, which already gives a number of combinations in excess of 400 that are used as excluded regions.[5] For every tuple (x_-, x_+), we fit a polynomial of order q to the data, very much in the spirit of (3):

$$\tilde{N}_j^{BW} = \sum_{i=0}^{q} \beta_i X_j^i + \varepsilon_j \qquad \forall \;\; j \notin [x_-, x_+], \tag{10}$$

[5] As shown in Bosch et al. (2020), the results are not sensitive to the number of excluded regions tested.

where q is chosen as to minimise the BIC. Further, we predict the counterfactual values $\hat{N}_j^{BW}$ and the associated standard error of the forecast:

$$\hat{N}_j^{BW} = \sum_{i=0}^{q} \hat{\beta}_i X_j^i \qquad \forall \ j. \tag{11}$$

We then calculate the upper value of the confidence interval CI_j^+ for a t-value of 1.96. To determine whether there are more individuals than predicted in an income bin j, we subtract the CI_j^+ from the observed number of taxpayers for each j:

$$E_j = N_j - CI_j^+. \tag{12}$$

A positive E_j means that the number of individuals in income bin j exceeds the upper bound of the confidence interval of the predicted number of individuals, as estimated by the polynomial regression. This indicates that more individuals are in the bin than should be in the absence of a kink point. The lower bound of the bunching window, conditional on the respective excluded region, is given by:

$$l(x_-, x_+) = j_l^* + 1, \qquad \text{where} \quad j_l^* = \max\{j \in \mathbb{Z}_- : E_j < 0\}, \tag{13}$$

which is the smallest adjacent income bin j, starting from the threshold at 0, that still satisfies the condition $E_j > 0$. Similarly, the upper bound, conditional on the respective excluded region, is given by:

$$u(x_-, x_+) = j_u^* - 1, \qquad \text{where} \quad j_u^* = \min\{j \in \mathbb{Z}_+ : E_j < 0\}, \tag{14}$$

which is the largest adjacent income bin j, starting from the threshold at 0, that still satisfies the condition $E_j > 0$. By following this procedure, different values are obtained for the lower and upper bounds of the bunching window. We follow the suggestion by

Bosch et al. (2020) and use the mode to determine the lower (l) and upper (u) bound of the bunching window when estimating (3). By doing so, we choose those boundary values which appear most often and hence are suitable over a wide spectrum of counterfactual specifications. However, other measures of centrality like the mean and median do not lead to substantially different results, although the mean is sensitive to outliers in the bunching window estimation.

3 Simulation Design

To evaluate the performance of the three procedures, we set up a framework similar to the well-known classification tables, for example used for binary regressions. As outlined before, the bunching window need not be too small nor too large. Hence, a successful procedure needs to minimise misclassifications. On the one hand, our *false negatives* are made up of those individuals that truly bunch, but lie outside the bunching window. On the other hand, our *false positives* encompass those individuals inside the bunching window that do not belong there, because they did not alter their taxable income. Finally, we compute the elasticity parameter according to (2) and investigate how a wrongly specified bunching window affects the bunching estimator. Since the restrictive assumptions of the polynomial strategy are violated in some specifications of our data-generating process, we expect the bunching estimator to be biased in those cases and additionally study the relative bias compared to situations without optimisation frictions to distinguish between both sources of bias.

In each iteration, we draw from a log-normal distribution of potential incomes[6], $z_0 = 10,000 \exp(1 + 0.5X)$, where X is standard-normally distributed. The distribution is visualized in Figure 1. From these, the after-tax incomes subject to the elasticity e

[6]Potential incomes in our setting are incomes that a person can reach at maximum, given a standard-normally distributed ability (e.g. IQ) in the population.

are derived for the lower ($(1-\tau_1)^e z_0$) and upper ($(1-\tau_2)^e z_0$) tax rate. Whenever $(1-\tau_2)^e z_0 \leq z^*$, i.e., the individual would be better off by bunching at the kink, we declare this individual to be a buncher and assign him/her the value of z^*. To account for optimisation frictions, we add an error term ε to the before-tax income of the bunchers, which creates the bunching window.[7] This excercise yields the realised or before-tax income distribution (z_1). Then, the expected after-tax incomes are

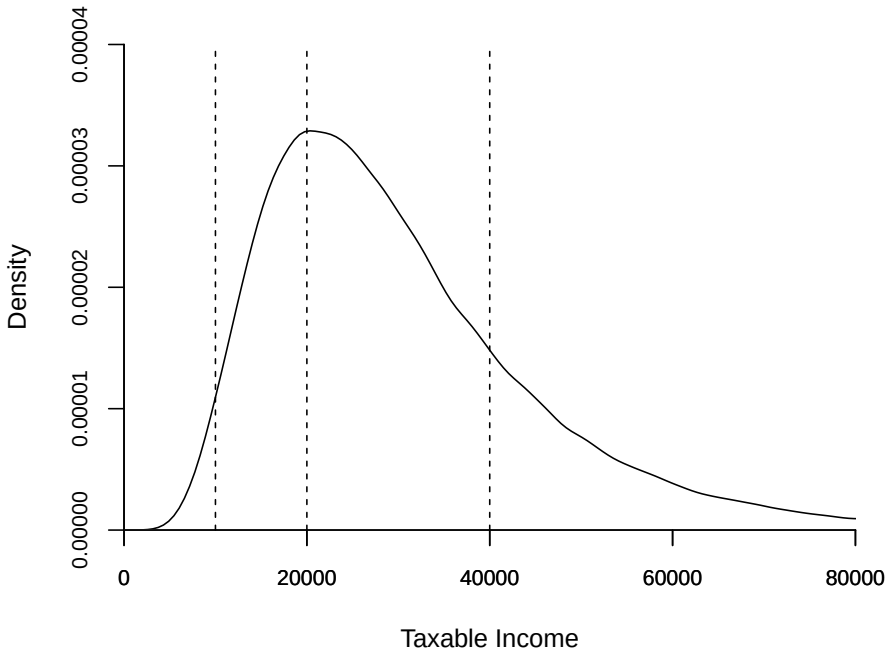
$$z = \begin{cases} (1-\tau_1)^e z_1, & \text{if} \quad (1-\tau_1)^e z_0 \leq z^* \\ z^* + \varepsilon, & \text{if} \quad (1-\tau_1)^e z_0 > z^* \geq (1-\tau_2)^e z_0 \\ (1-\tau_2)^e z_1, & \text{if} \quad (1-\tau_2)^e z_0 > z^*, \end{cases}$$

where z^* is the threshold at which point the marginal tax rate increases from τ_1 to τ_2 and $z_1 = z_0$ for all non-bunchers, i.e., they realise their full potential in our setting. For each specification, we run 1,000 replications. As a baseline scenario, we choose a sample size of $N = 200,000$, a binwidth of 100, a kink point at $z^* = 40,000$, a tax difference of $\Delta\tau_{1,2} = 0.1$, and an elasticity of $e = 0.1$. The error is normally distributed with standard deviation $\sigma = 100$, i.e., $\varepsilon \sim N(0, 10,000)$.

Because the analysed procedures are data-driven, the sample size should play a role in determining the appropriate bunching window. Likewise, the choice of binwidth and the location of the kink point determines the shape of the section from the income distribution under consideration, which could influence the ability of the data-driven procedures to determine the correct bunching window. Since the majority of papers (e.g. Chetty et al., 2011; Bastani and Selin, 2014; Seim, 2017) focus on top incomes,

[7] Note that we focus exclusively on optimisation frictions for true bunchers, adding the error only to those taxpayers trying to avoid paying the higher marginal tax rate. If we instead also considered income shocks outside of the control of taxpayers, we could add the error ε to each before-tax income as is done in Aronsson et al. (2018). Additional simulations with the alternative error specification yielded almost identical results (see Figure 7 in the Appendix). This implies that our preferred methods can also identify bunchers precisely in noisier settings.

Figure 1: Log-normal distribution of before-tax incomes with kink point locations (vertical dashed lines)



we chose $z^* = 40,000$ in our baseline, which lies in the descending part of the income distribution. We consider two additional thresholds in the ascending part $z^* = 12,000$ and at the mode of the distribution $z^* = 20,000$, respectively. Patel et al. (2017) show that the bias of the bunching estimator is a function of the slope of the counterfactual distribution over the bunching window. Approximating the counterfactual distribution should be most successful in case of the second threshold where the slope of the density curve is flat. This is made increasingly difficult at the third threshold with a moderately negatively sloped segment of the density curve and most difficult for the first threshold with a steeply positive sloped segment of the density curve. Hence, our specification of the thresholds allows us to investigate all relevant types of kink points. Larger changes in the tax rate should lead to more individuals with bunching behaviour and we thus vary the size of the tax change to analyse the appropriateness of the three procedures when the

tax changes are small or large. We use a base tax rate of $\tau_1 = 0.2$ and adjust τ_2 to increase the tax difference. Again, our baseline specification of the tax difference, $\Delta\tau_{1,2} = 0.1$, is motivated by the literature (le Maire and Schjerning, 2013; Bastani and Selin, 2014; Bosch et al., 2020). A similar argument can be made for a higher elasticity and therefore, we impose a range of different elasticities of the data. Our baseline elasticity of $e = 0.1$ is chosen to resemble the elasticity found in the literature for self-employed individuals (e.g. le Maire and Schjerning, 2013; Bastani and Selin, 2014) or UK firms at the upper kink in the tax schedule (Devereux et al., 2014). Larger elasticities are reported, for example, in Patel et al. (2017) considering the bunching behaviour of private firms and in Devereux et al. (2014) for the lower kink in the UK corporate tax schedule. For different approaches to estimate the ETI besides the bunching methodology, Neisser (2017) finds in a meta study that most elasticities are between 0 and 1. The error term used to simulate optimisation frictions, controls the width of the bunching window and provides the toughest test for our three procedures. Especially when the mass of bunching individuals is spread out over a larger part of the income distribution, eyeballing can become difficult. Variations in the standard deviation, as well as the distribution of the error term are used to determine whether the data-driven procedures are able to determine what eludes the naked eye.

4 Results

4.1 Monte Carlo Simulation

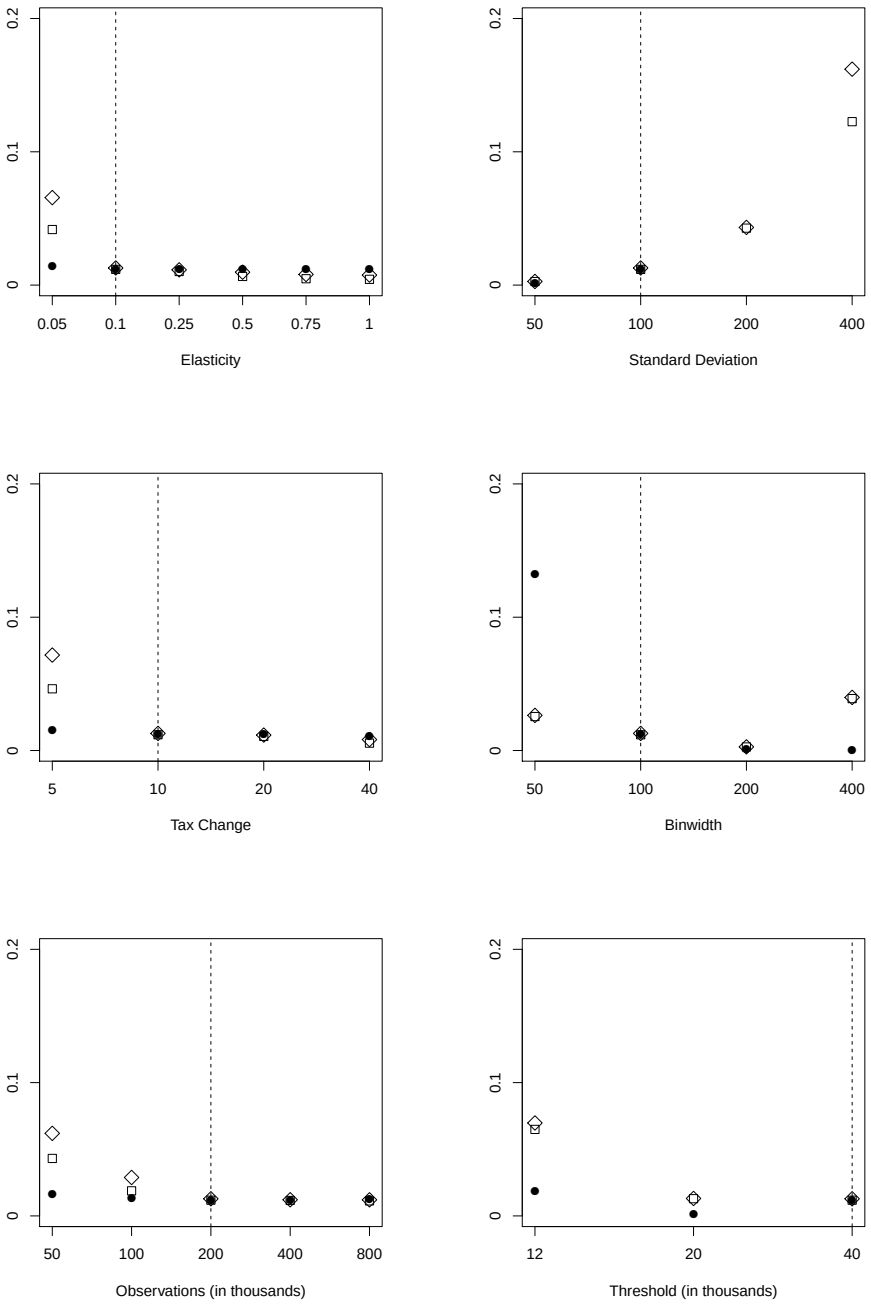
In the first set of specifications, the error term follows a normal distribution which is symmetric around the mean. We consider different configurations of all key parameters of the baseline specification. To account for the possibility of asymmetric bunching windows, we alter the structure of the error term in such a way that we obtain a left-

skewed and a right-skewed distribution in later specifications. The results for the *false negatives*-simulations for the normally distributed error term are depicted in Figure 2. We report the average percentage of *false negatives* over 1,000 replications.

The results show that across all specifications, BP (•) produces the smallest number of *false negatives* with the exception of specifications with the smallest binwidth and largest standard deviations. Both a small binwidth and a large standard deviation result in a widespread distribution of bunching individuals around the threshold. This in turn leads to the problem that the data no longer show structural deviations from their projected values and BP fails to detect a structural break in these settings. Both Cook's D (⋄) and BDS (□) show difficulties when the expected excess mass is very small, as indicated by the higher rate of *false negatives* for the smallest elasticity and smallest tax change, but once a critical excess mass is reached, a further increase in the elasticity or the tax difference has no significant impact on the rate of *false negatives*. Only BDS reaches a sufficiently low number of false positives when the standard deviation is increased to its maximum value of 400, i.e., the widest bunching window we consider.

Perhaps the most striking observation from Figure 2 is the considerable increase in *false negatives* for Cook's D and BDS when considering a threshold in the ascending part of the income distribution. This finding mostly hinges on the small number of observations that are to the left of the threshold, when utilising a binwidth of 100. Following Chetty et al. (2011), we use a maximum of 50 bins on either side of the threshold and given the nature of our income distribution, we observe relatively few individuals with incomes lower than 12,000. The lower boundary of incomes has a mean of 2,000 and the number of observations is near zero. In this sense, our findings suggest that Cook's D and BDS perform badly if sample sizes are small. A similar picture emerges when we consider the baseline threshold for top incomes and different sample sizes. As a consequence, we conclude that both Cook's D and BDS need a sufficiently

Figure 2: False negatives



Notes: The figure depicts the average percentage of *false negatives* classified using three different approaches for variations in key parameters. We draw 1,000 replications for each specification. BP allowing for two structural breaks is depicted as circles (•), the procedure implementing Cooks distances is shown as diamonds (⋄) and the BDS procedure as squares (□). The vertical dashed line highlights the baseline specification. Values outside the range 0 to 0.2 are not shown (BP for $\sigma > 100$).

large number of observations, as well as a significant tax change and elasticity to minimise the number of *false negatives*. As long as the specification is not too extreme, BP always selects the largest bunching window, minimising the number of *false negatives*, and should be the method of choice, if *false negatives* are the main concern.

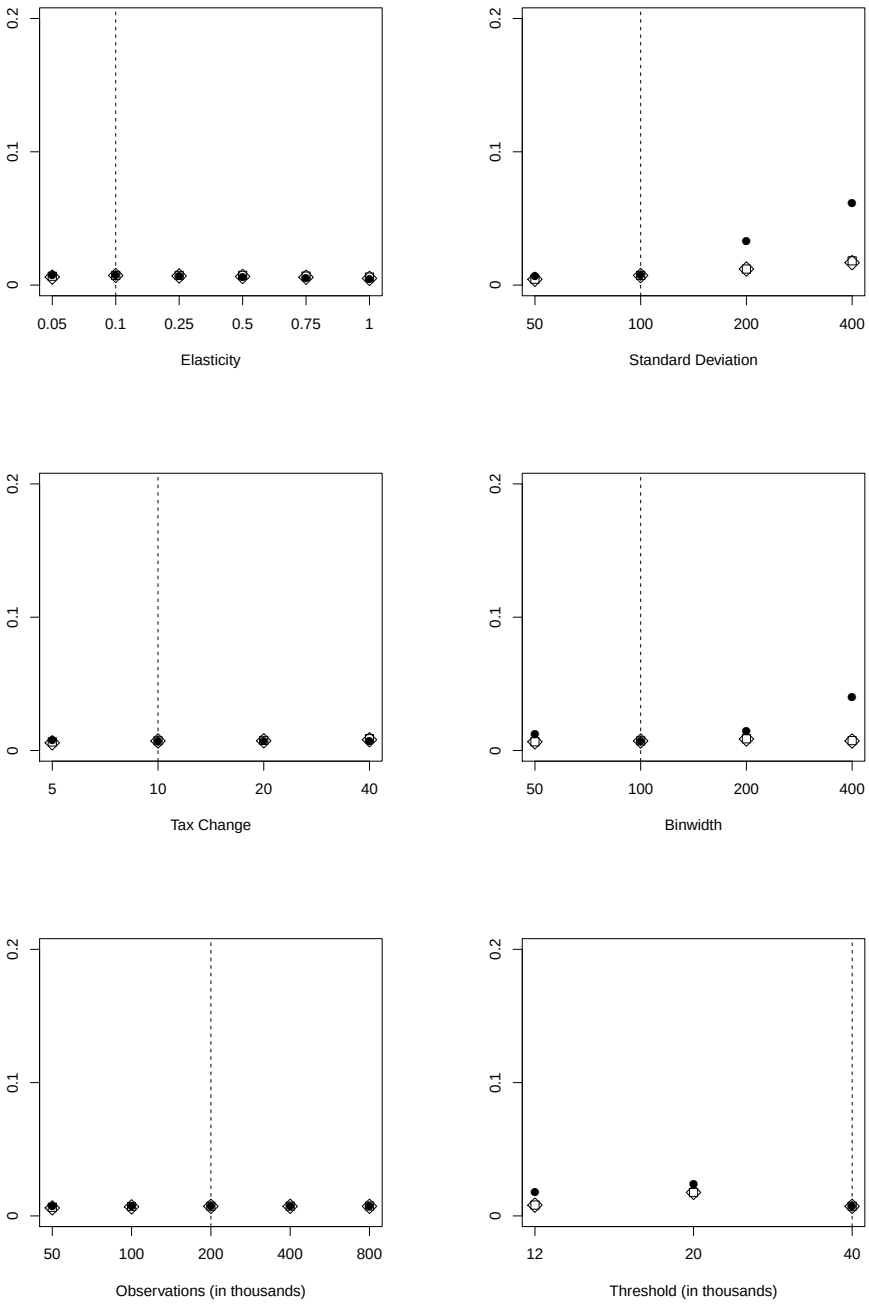
The conservative nature of determining a bunching window using BP, however, comes at the cost of an increased number of *false positives*. Because of the minimum required distance between two structural breaks, BP delivers a larger bunching window on average and therefore includes more *false positives* than either Cook's D or BDS. This finding is robust across all specifications, as shown in Figure 3. In line with the results from Figure 2, Cook's D and BDS perform very similarly across all tested specifications.

Since no procedure simultaneously minimises the number of *false negatives* and *false positives*, we now consider the average bias of the elasticity estimator in Figure 4. These results help us to show how wrongly classifying individuals as bunchers or non-bunchers might affect the results of empirical studies. As we have shown in Figure 2 and Figure 3, BP minimises the number of *false negatives* but the corresponding number of *false positives* is larger than for the other two procedures. We always report the average bias of the bunching estimator for a similar configuration without optimisation frictions to have a suitable benchmark in situations where the bunching estimator is naturally biased. Additionally, we computed the variance and the mean squared error of the elasticity estimates. Variances are almost identical for all procedures which can be attributed to the large sample sizes generated (motivated by the availability of large administrative datasets for most empirical studies). Hence, we do not report these results.[8]

Our results show that the average bias of the benchmark without optimisation frictions is surprisingly small if the elasticity is moderate. As expected, we observe the largest bias for the first threshold with the steepest slope and almost no bias at the second threshold located at the mode of our before-tax income distribution. On the basis of the

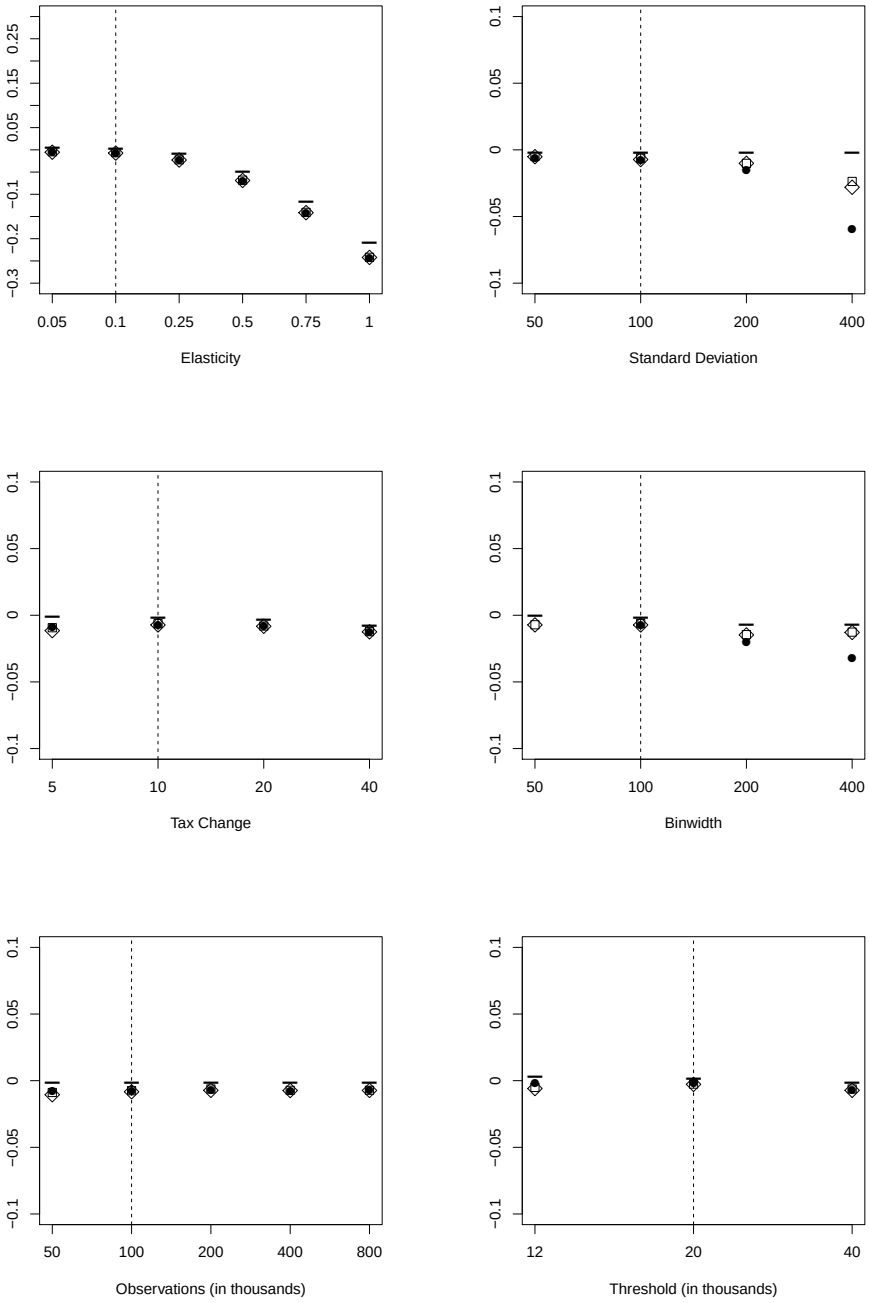
[8]Results are available from the authors upon request.

Figure 3: False positives



Notes: The figure depicts the average percentage of *false positives* classified using three different approaches for variations in key parameters. We draw 1,000 replications for each specification. BP allowing for two structural breaks is depicted as circles (•), the procedure implementing Cooks distances is shown as diamonds (⋄) and the BDS procedure as squares (□). The vertical dashed line highlights the baseline specification.

Figure 4: Average bias



Notes: The figure depicts the average bias of the elasticity estimator using three different approaches for variations in key parameters. We draw 1,000 replications for each specification. BP allowing for two structural breaks is depicted as circles (•), the procedure implementing Cooks distances is shown as diamonds (⋄) and the BDS procedure as squares (□). Benchmark results without optimisation frictions are indicated with a dash (−). The vertical dashed line highlights the baseline specification. Values outside the range −0.1 to 0.1 are not shown for binwidth.

average bias, we have to reject BP as an appropriate method to determine the bunching window because of its sensitivity to more extreme, but realistic specifications. Over most specification, BP has the strongest negative bias and in one specification (binwidth of 50) does not detect two breakpoints and hence does no produce a sensible estimate for the bunching window. Given the complexity of certain modern day tax systems, optimisation frictions could play a major role and here, BP performs worst of the three procedures. Cook's D and BDS again perform similarly, with the exception of Cook's D performing worse when the threshold is located at 12,000 (with our baseline elasticity of 0.1) or when the elasticity is large. These two findings are of varying importance for empirical practice. The bunching literature in general finds small elasticities (Chetty et al., 2011; Bastani and Selin, 2014; Kleven, 2016) and if the absolute bias reported in Figure 4 is perceived relative to the increasing values of the true elasticity, the differences between Cook's D and BDS are still small. However, the large average bias of the benchmark without optimisation frictions for elasticities above 0.5 suggests that the polynomial approach should not be used in these cases. Since Aronsson et al. (2017) also consider large elasticities and find that the bunching estimator has low relative bias, we investigate to what degree the performance of the bunching estimator is related to the slope of the income distribution. Figure 8 shows the corresponding results for both alternative thresholds at 12,000 and 20,000. The estimator is positively biased at the first threshold with a positive slope, slightly negatively biased at the mode of the income distribution, and negatively biased at the third threshold with a negative slope. It appears that larger elasticities are especially problematic for the bunching estimator if thresholds are located at the descending part of the income distribution. Consequently, the bunching estimator should be applied with caution if similarly located kinks are analyzed empirically. Taking into account that BDS performs more robust than Cook's D and that the performance of Cook's D depends on the choice of the cut-off value[9],

[9]Additional simulations with cut-off values corresponding to the 5% (10%) confidence region generated

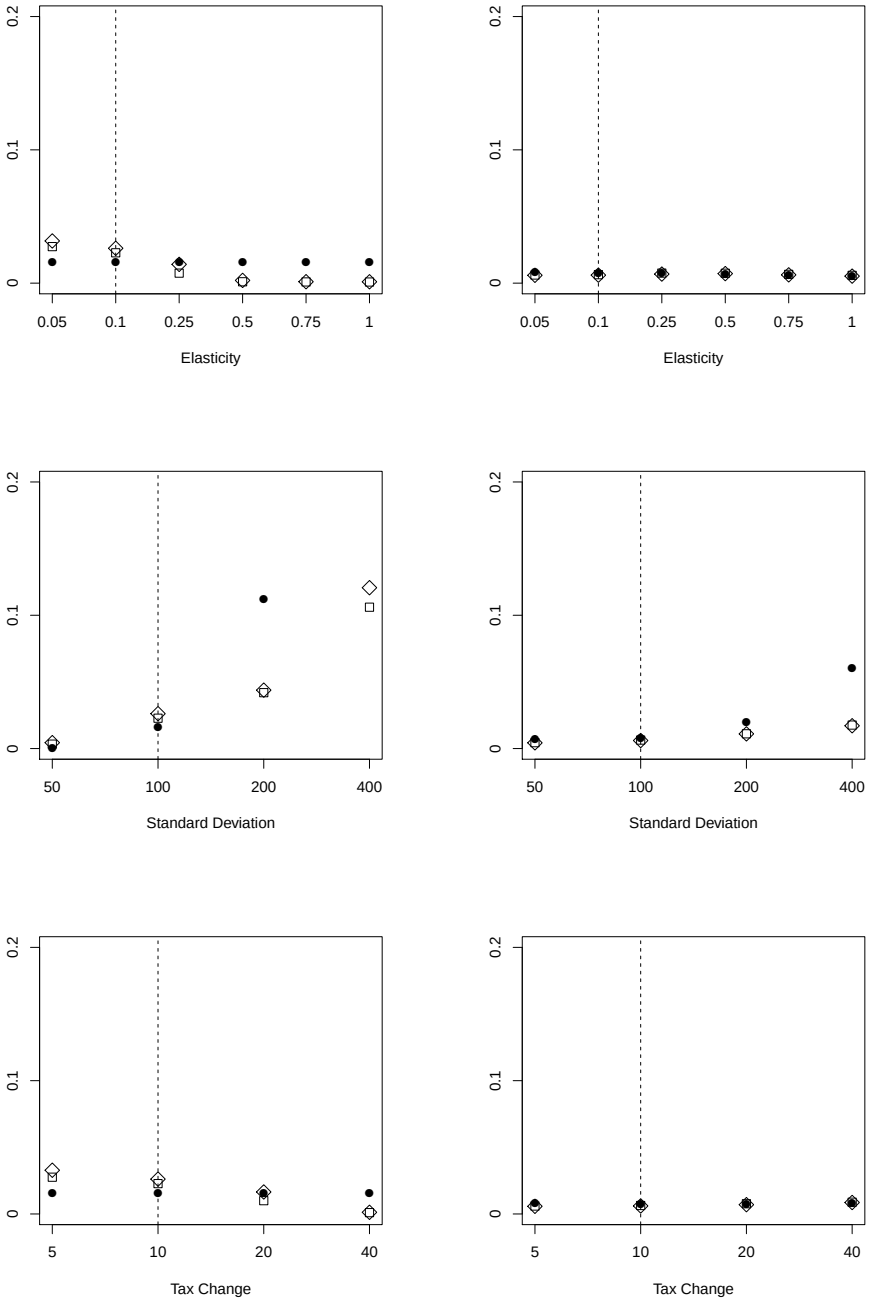
should make BDS preferable over Cook's D.

We have thus far shown results for estimations, where the bunching window was symmetric around the threshold. But as pointed out by Bosch et al. (2020), there are good reasons to believe that the bunching window is asymmetric in some cases. We therefore vary the structure of the optimisation frictions in such a way that we have more mass to the left (right) of the threshold. For this purpose, we draw ε from a skew-normal distribution (Azzalini, 1985). The probability density function (pdf) of the skew-normal distribution is given as $f(x) = 2\phi(x)\Phi(\alpha x)$, where $\phi(\cdot)$ and $\Phi(\cdot)$ denote the pdf and the cumulative distribution function of the standard normal distribution, respectively. The skewness of the distribution is controlled by the parameter α. The distribution is left-skewed if $\alpha < 0$ and right-skewed if $\alpha > 0$. We use $\alpha = -0.6$ and $\alpha = 0.6$ for our simulations. The average percentage of false negatives and false positives for three selected key parameters and asymmetrically distributed optimisation frictions are shown in Figure 5 (Figure 6).

BP overall shows little sensitivity to either the left- or right-side asymmetric specifications, with similar limitations when regarding extreme specifications as in the normally distributed error term case of our baseline specification. Strikingly, an asymmetric bunching window reduces the sensitivity of BP to extreme standard deviations. This finding might be explained by the location of our baseline threshold at a downward sloping segment of the income distribution. Shifting probability mass to the left/right of the kink-point creates a more distinctive break which can be detected more easily by the BP procedure. The consistency of Cook's D, however, seems to be vulnerable when the bunching window is asymmetric around the threshold. Especially with small elasticities and small tax changes, Cook's D has problems correctly identifying bunchers. The results of the BDS procedure remain consistent with the findings for the symmetric

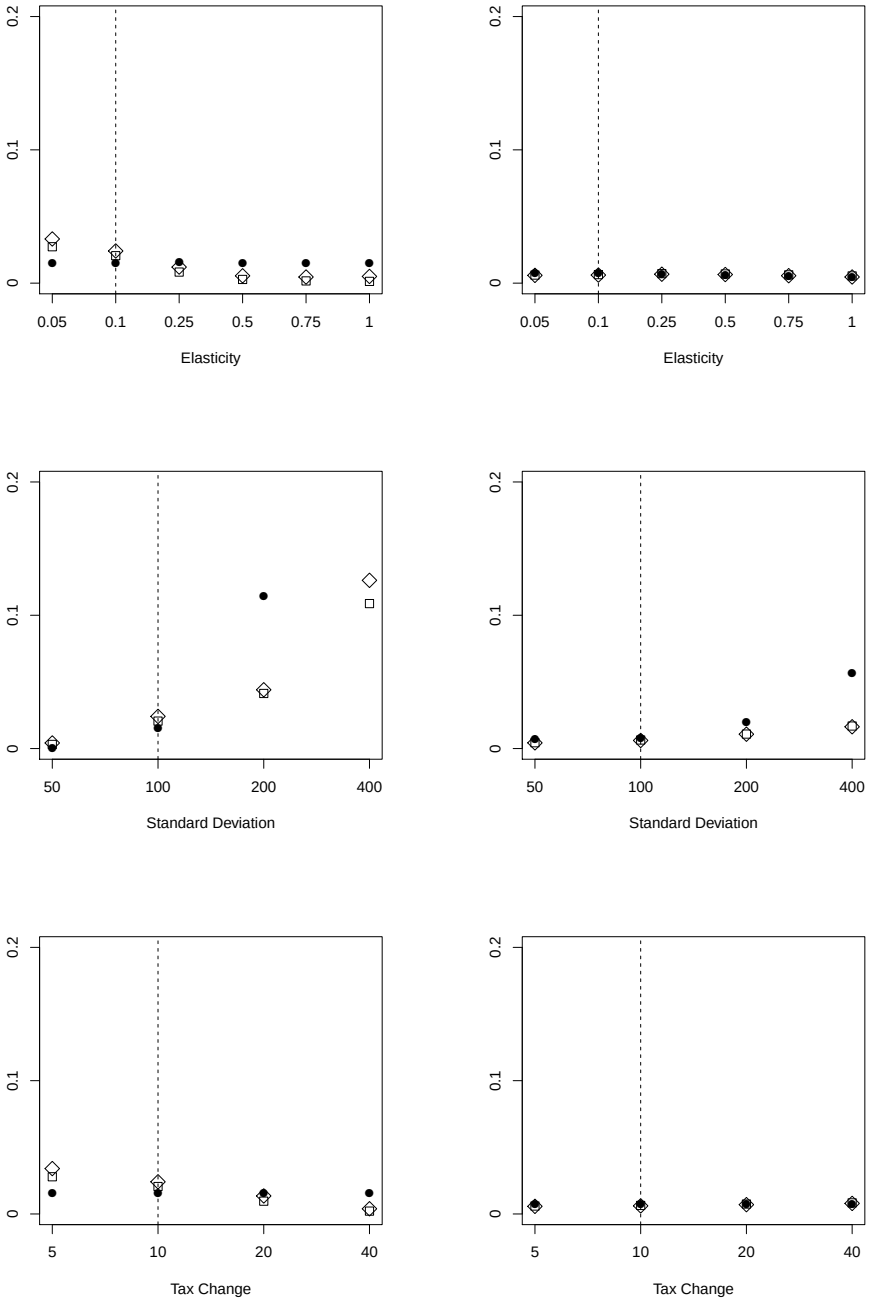
more *false negatives* and a more strongly biased bunching estimator. Particularly, if the bunching window is more spread out, Cook's D fails to identify the boundaries of the bunching window and selects a bunching window which is too wide.

Figure 5: False negatives and false positives for left-skewed distribution



Notes: The figure depicts the average percentage of *false negatives* (*false positives*) classified using three different approaches for variations in selected key parameters. The left graphs show *false negatives*, the right graphs show *false positives*. We draw 1,000 replications for each specification. BP allowing for two structural breaks is depicted as circles (•), the procedure implementing Cooks distances is shown as diamonds (⋄) and the BDS procedure as squares (□). The vertical dashed line highlights the baseline specification. Values outside the range 0 to 0.2 are not shown (BP for $\sigma = 400$).

Figure 6: False negatives and false positives for right-skewed distribution



Notes: The figure depicts the average percentage of *false negatives* (*false positives*) classified using three different approaches for variations in selected key parameters. The left graphs show *false negatives*, the right graphs show *false positives*. We draw 1,000 replications for each specification. BP allowing for two structural breaks is depicted as circles (•), the procedure implementing Cooks distances is shown as diamonds (⋄) and the BDS procedure as squares (□). The vertical dashed line highlights the baseline specification. Values outside the range 0 to 0.2 are not shown (BP for $\sigma = 400$).

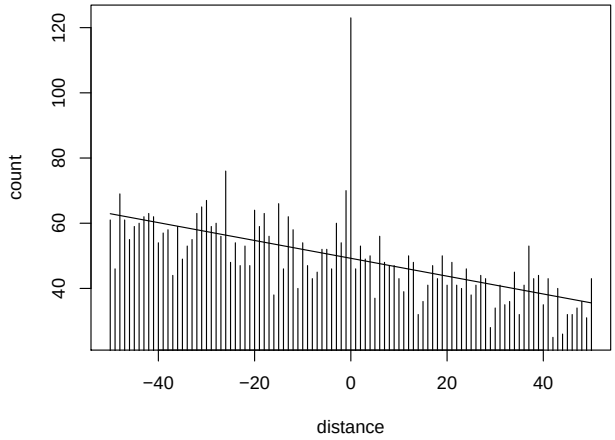
bunching window. In summery, BP delivers the most consistent results when the anticipated optimisation frictions are not severe. If optimisation frictions are expected to play a significant role, BDS simultaneously minimises the number of false negatives and false positives. It should therefore be the procedure of choice in empirical applications.

Finally, we follow Patel et al. (2017) and evaluate the performance of our data-driven methods at several placebo kinks with different slopes. We generate replications without bunching behaviour and compute the ETI for thresholds located at 12,000, 16,000, 20,000, 30,000, 40,000, and 50,000 to capture different slopes in the ascending and descending part of the income distribution. Our results are depicted in Figure 9. Although the BP procedure requires a minimum number of observations in the middle segment and therefore miss-classifies some non-bunchers, the average bias for all methods is quite small. It seems to be more important to examine large empirical ETI estimates more closely which are often based on wider bunching windows that tend to create problems for the (local) bunching estimator.

4.2 Empirical Application

To further highlight the significant influence that the choice of the bunching window has on estimating the ETI, we analyse a subset of the data utilised in Bosch et al. (2020). The data come from the public use IPO file from the Dutch central bureau for statistics (Centraal Bureau voor de Statistiek, 2001), which is a yearly panel of administrative data that contains exact taxable income. We use the publicly available file for the year 2007 (Centraal Bureau voor de Statistiek, 2001) to show the differences in estimated elasticities across the three data-driven procedures. In line with Bosch et al. (2020), we clean the data from individuals that receive benefits of some sort, are not in the active workforce or have no income. This is important to abstract from mass points that are generated by other sources than the tax schedule. For example

Table 1: ETI upper threshold in the Netherlands (2007)



	BP	Cook's D	BDS
Bunching Window	$[-3, 1]$	$[-1, 0]$	$[-1, 0]$
ETI	0.022	0.020	0.020

Notes: The graph shows the distribution of taxable income around the upper threshold of the Dutch tax system. The binwidth is 100. BIC determined the counterfactual model to be linear. $z^* = 53,064$, $\tau_1 = 0.42$, $\tau_2 = 0.52$. Data provided by Centraal Bureau voor de Statistiek (2001).

close to the first threshold, many individuals are located that receive the same amount of (unemployment) benefits.[10] Our sample thus includes 105,112 individuals that are either employed or self-employed.

We start our analysis by looking at the top tax threshold, where the change in the marginal tax rate is biggest and thus, the incentive to bunch is largest. Table 1 first shows the distribution of taxable income around the upper threshold of the Dutch tax system, where the marginal tax rate changes by 10 percentage points. Then, the results from the analysis using the three data-driven procedures discussed in this study are shown. As became evident through the course of this section, Cook's D and BDS perform similarly and also in this real world application, they both identify the bunching window to be from -1 to 0, which translates into an ETI of 0.020. BP on the other hand determines the bunching window to go from -3 to $+1$, which translates into an ETI of 0.022.

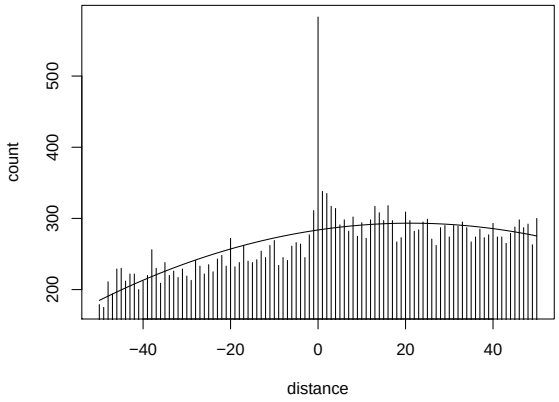
Table 2 shows the results for the first (Panel A) and second (Panel B) threshold of the Dutch tax system in the year 2007. Similar to the findings regarding the top tax threshold, we find the same bunching window using Cook's D or the BDS procedure, but a different and larger window using the BP method. This translates into a smaller elasticity that is identified at the first threshold using the BP method in comparison to the other two and a larger elasticity at the second threshold.

The empirical findings are in line with our simulation results and highlight the sensitivity of the estimator to the specification of the bunching window. Whilst BP estimates an elasticity that is 10% larger at the top tax threshold, it estimates an elasticity that is 51% smaller at the first and 11.8% larger at the second threshold, when compared to Cook's D and the BDS procedure. Furthermore, our simulation approach hinted at sensitivities of Cook's D regarding asymmetric bunching windows. Although none of the bunching windows identified in our empirical application are perfectly symmetric

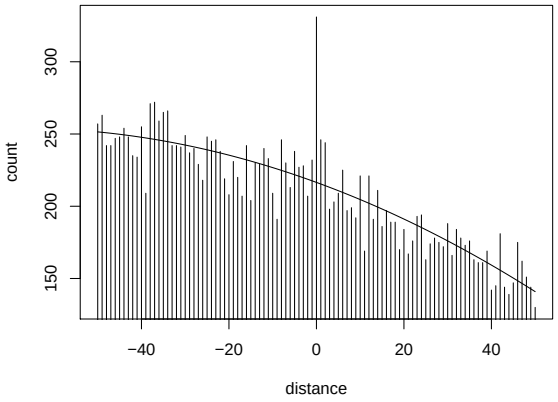
[10] Note that this artificial mass point would especially affect the BP procedure, whilst the other two procedures are better able to cope with this, provided that the mass point is not too close to the tax induced mass point.

Table 2: Distribution around first and second tax thresholds in the Netherlands

Panel A: First Threshold



Panel B: Second Threshold



	$z^* = 17,319$, $\tau_1 = 0.3365$, $\tau_2 = 0.414$			$z^* = 31,122$, $\tau_1 = 0.414$, $\tau_2 = 0.42$		
	BP	Cook's D	BDS	BP	Cook's D	BDS
BW	$[-4, 1]$	$[-1, 4]$	$[-1, 4]$	$[-1, 7]$	$[0, 2]$	$[0, 2]$
ETI	0.062	0.094	0.094	0.304	0.272	0.272

Notes: Distribution of taxable income around the first and second threshold of the Dutch tax system. The binwidth is 100. BIC determined the counterfactual model to be quadratic at both thresholds. BW = Bunching Window, ETI = Elasticity of Taxable Income.

around the threshold, they are not very far off (at least when identified by Cook's D and BDS). We conclude that the asymmetry of the bunching window needs to be greater than in the Dutch tax system case for Cook's D to deliver different results than BDS. Because the bunching mass stays quite precisely around the threshold, BP encounters a difficult situation in this application. As we have argued, BP needs, by construction, a certain number of bins inside one regime. This seems to be the reason that BP always finds a larger bunching window than the other two procedures.

5 Conclusion

We have compared three different data-driven procedures to determine the bunching window in order to elicit, which method is able to identify bunching best. The bunching window should encompass those individuals, that change their behaviour and only those. Thus, we set up a framework similar to a classification table and analyze the number of *false negatives* and *false positives*, where the former measures how many bunchers the respective procedure fails to detect and the latter measures those individuals that are falsely flagged as bunching. A successful data-driven procedure should ideally minimise both types of errors.

When the bunching window is normally distributed around the threshold, Cook's D and BDS performed best with slight advantages of BDS over Cook's D if standard deviations were large or sample sizes were small. Cook's D stands representative for the general class of outlier detection methods. Further research is necessary to determine if other outlier detection techniques can be designed to outperform Cook's D in terms of computational costs and accuracy. BP performed best when the tax change was small or the optimisation frictions were not too large. BDS seemed to be the most robust method and showed a balanced performance over all specifications. In a real world application, we showed that the size of the bunching window has a significant impact on the bunching

estimator and subsequently its translation into the ETI. Overall, we conclude that BDS should be the preferred method to detect the bunching window when nothing is known about the true shape of the bunching window, although Cook's D, especially for its ease of implementation, comes a close second.

There are several avenues for future research. Since the standard bunching estimator is only unbiased under restrictive assumptions, robust bunching estimators might be employed which are not based on polynomial regressions. Available estimators in the literature, for example those discussed in Bertanha et al. (2019), abstract from empirically relevant optimisation frictions and assume that the bunching window is known. Alternative parametric estimators, proposed by Aronsson et al. (2018), explicitly model optimisation frictions but require strong distributional assumptions. Further, Cattaneo et al. (2019) show that the standard polynomial approach does not produce reliable predictions at boundary points, which can be another source of bias of the bunching estimator, and propose to use local polynomial density estimators for predicting the counterfactual distribution. It remains an open question how these estimators can be modified to ensure a data-driven determination of the bunching window.

References

Aronsson, T., K. Jenderny, and G. Lanot (2017): "The Quality of the Estimators of the ETI," *Umea School of Business and Economics*, 1–59.

——— (2018): "Alternative parametric bunching estimators of the ETI," *Umea School of Business and Economics*, 1–39.

Azzalini, A. (1985): "A Class of Distributions Which Includes the Normal Ones," *Scandinavian Journal of Statistics*, 12, 171–178.

Bai, J. and P. Perron (1998): "Estimating and Testing Linear Models with Multiple Structural Changes," *Econometrica*, 66, 47–78.

——— (2003): "Computation and analysis of multiple structural change models," *Journal of Applied Econometrics*, 18, 1–22.

Bastani, S. and H. Selin (2014): "Bunching and non-bunching at kink points of the Swedish tax schedule," *Journal of Public Economics*, 109, 36–49.

Belsley, D. A., E. Kuh, and R. E. Welsh (1980): *Regression Diagnostics: Identifying Influential Data and Sources of Collinearity*, New York: Wiley, 1 ed.

Bertanha, M., A. H. McCallum, and N. Seegert (2019): "Better Bunching, Nicer Notching," *SSRN Electronic Journal*.

Best, M. C. and H. J. Kleven (2018): "Housing market responses to transaction Taxes: Evidence from notches and stimulus in the U.K," *Review of Economic Studies*, 85, 157–193.

Blomquist, S. and W. K. Newey (2017): "The bunching estimator cannot identify the taxable income elasticity," .

Bollen, K. A. and R. W. Jackman (1990): "Regression Diagnostics: An Expository Treatment of Outliers and Influential Cases," in *Modern Methods of Data Analysis*, ed. by J. Fox and J. S. Long, Sage Publications.

Bosch, N., V. Dekker, and K. Strohmaier (2020): "A Data-Driven Procedure to Determine the Bunching Window-An Application to the Netherlands," *International Tax and Public Finance*, forthcoming.

Cattaneo, M. D., M. Jansson, and X. Ma (2019): "Simple Local Polynomial Density Estimators," *Journal of the American Statistical Association*, 0, 1–7.

Centraal Bureau voor de Statistiek (2001): "Inkomenspanelonderzoek - IPO - 2001-2002, 2005, 2006, 2007," DANS. https://doi.org/10.17026/dans-xnb-2yw7.

Chetty, R., J. N. Friedman, T. Olsen, and L. Pistaferri (2011): "Adjustment Costs, Firm Responses, and Micro vs. Macro Labor Supply Elasticities: Evidence from Danish Tax Records." *The Quarterly Journal of Economics*, 126, 749–804.

Cook, R. D. (1977): "Detection of Influential Observation in Linear Regression," *Technometrics*, 19, 15–18.

Cook, R. D. and S. Weisberg (1982): *Residuals and Influence in Regression*, New York: Chapman & Hall, 1 ed.

Devereux, M. P., L. Liu, and S. Loretz (2014): "The elasticity of corporate taxable income: New evidence from UK tax records," *American Economic Journal: Economic Policy*, 6, 19–53.

Einav, L., A. Finkelstein, and P. Schrimpf (2017): "Bunching at the kink: implications for spending responses to health insurance contracts," *Journal of Public Economics*, 146, 27–40.

Gruber, J. and E. Saez (2002): "The elasticity of taxable income: Evidence and implications," *Journal of Public Economics*, 84, 1–32.

Kim, C., Y. Lee, and B. U. Park (2001): "Cook's distance in local polynomial regression," *Statistics and Probability Letters*, 54, 33–40.

Kleven, H. J. (2016): "Bunching," *Annual Review of Economics*, 8, 435–464.

Kleven, H. J. and M. Waseem (2013): "Using Notches to Uncover Optimization Frictions and Structural Elasticities: Theory and Evidence from Pakistan," *The Quarterly Journal of Economics*, 128, 669–723.

le Maire, D. and B. Schjerning (2013): "Tax bunching, income shifting and self-employment," *Journal of Public Economics*, 107, 1–18.

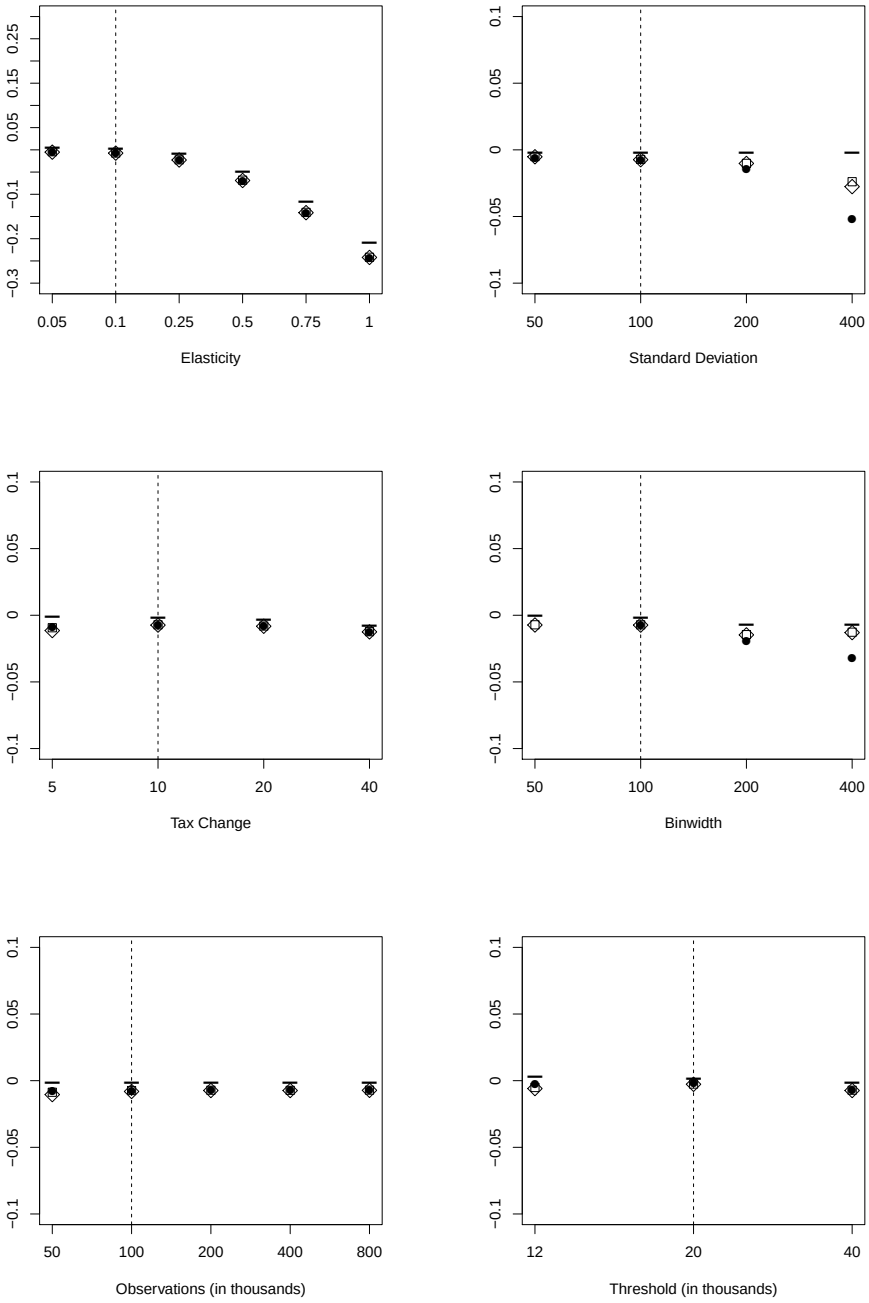
Neisser, C. (2017): "The elasticity of taxable income: a meta-regression analysis," *ZEW Discussion Papers*, 17.

Patel, E., N. Seegert, and M. G. Smith (2017): "At a Loss: The Real and Reporting Elasticity of Taxable Income," *SSRN Electronic Journal*, 1–50.

Saez, E. (2010): "Do Taxpayers Bunch at Kink Points?" *American Economic Journal: Economic Policy*, 2, 180–212.

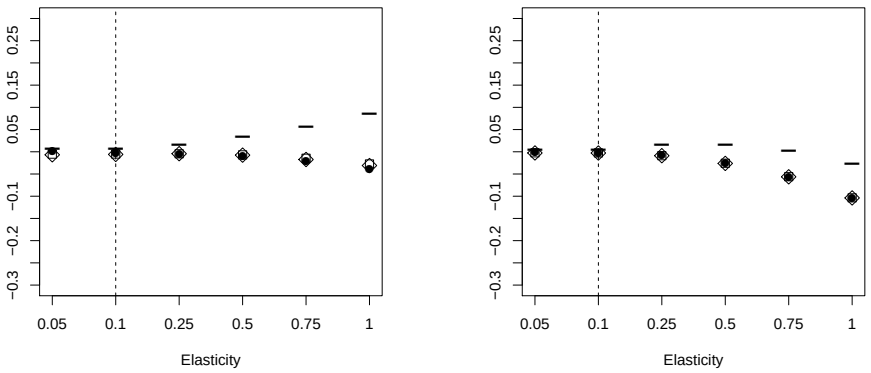
Seim, D. (2017): "Behavioral Responses to Wealth Taxes: Evidence from Sweden," *American Economic Journal: Economic Policy*, 9, 395–421.

Figure 7: Average bias (alternative error specification)



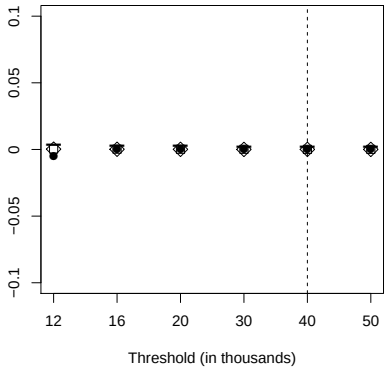
Notes: The figure depicts the average bias of the elasticity estimator using three different approaches for variations in key parameters. We draw 1,000 replications for each specification. BP allowing for two structural breaks is depicted as circles (•), the procedure implementing Cooks distances is shown as diamonds (⋄) and the BDS procedure as squares (□). Benchmark results without optimisation frictions are indicated with a dash (−). The vertical dashed line highlights the baseline specification. Values outside the range −0.1 to 0.1 are not shown for binwidth.

Figure 8: Average bias (alternative thresholds)



Notes: The figure depicts the average bias of the elasticity estimator using three different data-driven approaches for thresholds located at 12,000 (left) and 20,000 (right). We draw 1,000 replications for each specification. BP allowing for two structural breaks is depicted as circles (•), the procedure implementing Cooks distances is shown as diamonds (⋄) and the BDS procedure as squares (□). Benchmark results without optimisation frictions are indicated with a dash (−). The vertical dashed line highlights the baseline specification.

Figure 9: Average bias for different placebo kinks



Notes: The figure depicts the average bias of the elasticity estimator using three different approaches for different placebo kinks. We draw 1,000 replications for each specification. BP allowing for two structural breaks is depicted as circles (•), the procedure implementing Cooks distances is shown as diamonds (⋄) and the BDS procedure as squares (□). Benchmark results without optimisation frictions are indicated with a dash (−). The vertical dashed line highlights the baseline specification.