Sequential Specification Tests to Choose a Model

Adam Sales

January 2, 2017

1 Introduction

2 The Setup

Say, in specifying a model, a research must choose from a discrete, ordered, set of specifications, such as variables to include in a linear model, bandwidths in a regression discontinuity design, or calipers in a propensity-score matching design. Denote the set of possible choices $d = 1, 2, \dots, D$.

The researcher wishes to choose the best model that satisfies a particular testable assumption $\mathcal{A}$. Assume that either $\mathcal{A}$ is false for all d, or that for some $1 \leq d^* \leq D$, $\mathcal{A}$ is true for $d \leq d^*$ and false for all $d > d^*$. Further assume that if d^* exists, it is the optimal choice—for instance, the smallest model, or the biggest dataset, that satisfies $\mathcal{A}$. Finally, assume the researcher has chosen a valid, unbiased test of $\mathcal{A}$ and caluculated p-values for each d: $\boldsymbol{p}_D = p_1, \dots, p_d, \dots, p_D$. The procedure here is to use $\boldsymbol{p}_D$ to choose a specification $\hat{d}$ that is as large as possible without violating $\mathcal{A}$.

A common choice for d in this scenario relies on the logic of null hypothesis testing: for a pre-specified $\alpha \in (0, 1)$, let

$$\hat{d}_\alpha \equiv max\{d : p_d > \alpha\}.$$

That is, $\hat{d}_\alpha$ is the largest value of d for which the null hypothesis that $\mathcal{A}$ is true for $d \le \hat{d}_\alpha$ cannot be rejected at level α. Although it may seem as though the multiplicity of tests invovled in this procedure invalidates the null hypothesis framework, it turns out that this is not the case: the "stepwise intersection-union pricipal" **???** insures that the family-wise error rate is maintained. That is, the probability of falsely rejecting the null—choosing $\hat{d}_\alpha < d^*$, is bounded by α. $\hat{d}_\alpha$ is the specification that would result from testing null hypotheses backwards: for $d' = D, D-1, \dots, d, \dots, 1$, test $H_{0d'}$: $\mathcal{A}$ is true for $d \le d'$. Then, stop testing at $d' = \hat{d}_\alpha - 1$—the first d' for which $p_{d'} \ge \alpha$; reject all null hypotheses $H_{0d'}$ for which $d' \ge \hat{d}_\alpha$, and fail to reject the rest. This protects the family-wise error rate of α since rejecting *any* true null implies rejecting the first true null—a probability α event.

A tempting alternative choice for $\hat{d}$, say $\hat{d}_{\tilde{\alpha}}$, does not have this property. The choice $\hat{d}_{\tilde{\alpha}} \equiv min\{d : p_d < \alpha\} - 1$ selects $\hat{d}$ to be the largest value of d before the first significant p-value. This is equavalent to the opposite procedure: start with the $d' = 1$ and test sequentially for larger values of d' until the first rejection, at $\hat{d}_{\tilde{\alpha}}$, then stop; reject all null hypotheses $H_{0d'}$ for $d' \ge \hat{d}_{\tilde{\alpha}}$ and fail to reject the rest. This procedure does not control family-wise error rates—it is likely to reject more than $100\alpha\%$ valid specifications.

This paper will focus on two data scenarios for SSTs. In the first, the SSTs help determine which data are included in the analysis. For instance, choosing the bandwidth of a regression discontinuity design, or choosing the parameters of a matching design. In this scenario, each choice d corresponds to rows in the dataset that could be included in the analysis. Formally, let $\mathcal{I} = \{1, \dots, N\}$, indices for N candidate cases to be fit in a model. Then let $\mathcal{I} = \rangle_1 \cup \rangle_2 \cup \dots \cup \rangle_d \cup \dots union \rangle_D$. The choice of $\hat{d}$ means fitting the model to the dataset including each of these subsets, $\mathcal{I}_d = \rangle_1 \cup \dots \rangle_d$. Note that the sets denoted with lower-case $\rangle_d$ are disjoint, $\rangle_d \cap \rangle_{d'} = \emptyset$, those denoted with upper-case $\mathcal{I}_d$ are nested—$d > d'$ implies $\mathcal{I}_{d'} \subset \mathcal{I}_d$, and the full set of indices, noted without a subscript, $\mathcal{I} = \mathcal{I}_D$. Finally, let $n_d = |\rangle_d|$, where $|\cdot|$ denotes cardinality, and $\bar{n_d} = |\mathcal{I}_d|$.

A second scenario applies when the dataset is fixed, but the model is not. Here, d indexes a *pre-specified* sequence of models. For instance, if SSTs are used to determine which variables should be included in a model, with the goal being smallest model that satisfies $\mathcal{A}$. Then let $\boldsymbol{X}$ denote the full set of variables, $\boldsymbol{x}_d$ denote the set of variables that would be *subtracted* in the d^{th} step of the sequence, and $\boldsymbol{X}_d = \boldsymbol{x}_{d+1} \cup \cdots \cup \boldsymbol{x}_D$ denote the set of variables that would be included in the analysis, were the analyst to choose d. Note here that bigger values of d correspond to smaller models. In this scenario, the sample size is fixed at N.

2.1 Model Selection and the Logic of Null Hypothesis Testing

In order to avoid certain methodological mistakes, it may be helpful to clarify some of the conceptual distinctions between SSTs and conventional null hypothesis tests (NHTs). The logic of NHTs is familiar to anyone who has taken (and understood) even the most basic college statistics course; nonetheless we restate it here to distinguish it from the logic of SSTs. Typically, researchers use NHTs to reject a null hypothesis that they consider uninteresting—most of the time, that a model parameter is equal to zero—and interpret rejection as evidence in favor of an interesting alternative hypothesis. NHTs cap the probability of a type-I error—falsely rejecting a true null hypothesis—and, given that constraint, seek to minimize the probability of a type-II error, failing to reject a false null hypothesis.

SSTs reverse some of these elements; most importantly, the goal of SSTs is to identify specifications in which an assumption $\mathcal{A}$ is plausible, rather than to identify true alternative hypothesis. In the same vein, type-II errors are typically of more concern for SSTs than for typical NHTs, and type-I errors are less problematic. In fact, a type-II error from a specification test could lead a researcher to fit a misspecified model, which in turn may inflate the probability of a type-I error in her final outcome analysis. For that reason, some methodologists recommend setting α substantially higher for specification tests than for

NHTs in outcome analyses. Still, the hypothesis testing framework, in the case of pont null hypotheses, does not allow a researcher to fix the type-II error rate at a pre-specified value, and then optimize the type-I error rate, though that might be ideal for specificaton tests.

In fact, in continuous data models with continuous parameter spaces, no hypothesis test can provide any evidence in favor of a point null hypothesis. For instance, take the common $H_0 : \theta = 0$, for some parameter $\theta \in \mathbb{R}$. In finite samples, for any type-I or type-II error rate, there will always be some plausible alternative hypothesis $H_a \theta = \epsilon \neq 0$. Further, in these situations, finite sample estimates $\hat{\theta}$ will almost surely be non-zero. This is important to state to avoid misinterpretations of SST procedures as providing evidence, or showing, that an assumption $\mathcal{A}$ is true for certain specificatons d. A common Bayesian argument (e.g. **?**, p. 439; **?**) states that, theoretically, nearly all null hypotheses are false anyway—so testing them makes little sense. In the case of specification tests, that means that an assumption $\mathcal{A}$ can be assumed to be false for all d without even conducting a test; in other words, "all models are wrong" (**?**, p. 2).

"But some are useful." In practice there is much to be gained by considering assumptions such as $\mathcal{A}$. In this framework, it may indeed make sense to identify a set of specifications d for which $\mathcal{A}$ is plausible, or approximately true, and SSTs can be useful in this regard—as long as they are understood correctly, and not as providing evidence *for* $\mathcal{A}$.

In many scenarios the choice of d involves a bias-variance tradeoff: if $d > d^*$, then $\mathcal{A}$ is false and the resulting analysis will be biased. On the other hand, a sub-optimal choice for d often means a high-variance estimate. For instance, in the RDD bandwidth case, choosing $d > d^*$ might mean fitting a misspecified model to Y and R, but choosing $d << d^*$ means discarding data that can boost precision. Rather than choosing a criterion, such as mean-squared-error, that balances bias and variance, the SST approach may be seen as an attempt to hold bias at approximately zero, and minimize variance under that constraint. Granted, this is obviously an overly-optimistic take on model fitting; still, SSTs hope to contrain bias to be approximately zero, and from there minimize variance.

2.2 More Reservations with Null Hypothesis Testing for Model Selection

Applying a strict hypothesis-testing framework to SSTs for model selection has some additional drawbacks. First, it requires researchers to choose a test-level α. While using tuning parameters to mediate the bias-variance tradeoff is not uncommon in statistics, the level α is a particularly hard parameter to choose.

? poses an additional problems with the use of hypothesis tests to choose a model: the need to specify a null hypothesis. In their words (p. 179),

> Whenever a hypothesis test is used to choose between two models, one model must be selected as a null hypothesis. In most instances, this is usually the more parsimonious model and typically a nested test is applied. Often it is difficult to distinguish between the two models because of data quality (multicollinearity, near-identification, or the models being very similar such as in testing for integration). In such cases, the model chosen to be the null hypothesis is unfairly favored.

In other words, because of the structure of null hypothesis tests, which constrain the type-I error rate, the null model is unfairly favored. In our terminology, $\hat{d}$ is likely to be too small, perhaps $\mathbb{E}\hat{d} < d^*$. However, such a bias (if it indeed exists) needn't doom SSTs—an underestimated $\hat{d}$ is merely sub-optimal. In our setup, choosing $\hat{d}$ to be too low will yield and inefficient, but still valid, model. Would that every statistical model were valid yet suboptimal!

More broadly, perhaps, one might argue that null hypothesis tests are design to rule out hypotheses that are inconsistant with the data, not to estimate parameters. However, as **?**

showed, these aims are not contradictory—tests that rule out implausible hypothesis may also point researchers towards the correct answer.

Moving from rejecting implausible specifications to estimating optimal specifications requires a theory, or at least a reasonable heuristic. The following section will suggest one.

3 Finding the Change-Point

In the context of change point estimation, **?** suggests such a heuristic. They discuss a random variable x_t, whose distribution is a function of a continuous covariate t. For $t < d^*$, $\mathbb{E}x_t = \tau_0$, a constant; for $t > d^*$, $\mathbb{E}x_t > \tau_0$. They propose an estimate of d_0 based on p-values p_t testing the hypotheses $H_{0t} : \mathbb{E}x_t = \tau_0$. They note that for $t < d^*$, the null hypotheses are true, so $p_t \sim U(0,1)$, and $\mathbb{E}p_t = 1/2$; when $t > d^*$, the null hypotheses are false, and the p-values converge in probability to zero. That fact leads them to the following least-squares estimator for d^*:

$$\hat{d}_M \equiv argmin_d \sum_{t \le d} (p_t - 1/2)^2 + \sum_{t>d} p_t^2.$$

In other words, the estimate $\hat{d}_M$ is the point at which the p-values cease behaving as p-values testing a true null, with mean 1/2, and instead are drawn from a distribution with a lower mean. It turns out that an equivalent expression for $\hat{d}_M$ is:

$$\hat{d}_M = argmax_d \sum_{t \le d} (p_t - 1/4). \tag{1}$$

? shows that as n_t, the number of data points at each value t, and the number of sampled values of t increase, $\hat{d}_M$ converges in probability to d^*.

The same broad logic applies to any set of p-values from sequential tests: $\hat{d}_M = argmax_d \sum_{t \le d}(p_t - 1/4)$ may be considered an estimate of d^*. In the case of SSTs, for $d \le d^*$, p-values p_d are draws from a $U(0,1)$ distribution, and hence have mean 1/2, and, as n_d or N increase,

$p_d \to_p 0$ for $d > d^*$. Some differences in the details, though, lead to differences in $\hat{d}_M$'s behavior. For instance:

Lemma 1. *If indeed $p_d \to_p 0$ for $d > d^*$, as n_d or N increase, then $\hat{d}_M$ is assymptotically conservative: $Pr(\hat{d}_M > d^*) \to 0$.*

Proof. For each d, $Pr(p_d - 1/4 > 0) \to 0$, implying that for all d', $Pr(\sum_{d^*<t\leq d'}(p_t - 1/4) > 0) \to 0$. Therefore, for $d^* < d \leq D$, $Pr(\sum_{t\leq d}(p_t - 1/4) > \sum_{t\leq d^*}(p_t - 1/4)) \to 0$. □

That is, as sample size increases, the probability that $\hat{d}_M$ suggests a model that violates assumption $\mathcal{A}$ decreases to zero. The same property holds for $\hat{d}_\alpha$, with $\alpha > 0$ fixed, for the same reason.

On the other hand, even with an infinite sample $\hat{d}_M$ may choose a sub-optimal model, $\hat{d}_M < d^*$. As sample size grows, the distribution of p_d, $d \leq d^*$ remains stable at $U(0, 1)$. When $p^*_d - 1/4 < 0$, $\hat{d}_M \neq d^*$, since $\sum_{d\leq d^*-1}(p_d - 1/4) > \sum_{d\leq d^*}(p_d - 1/4)$. Since $Pr(p^*_d - 1/4 < 0) = 1/4$ regardless of sample size, $\hat{d}_M$ will be conservative in large samples. The difference between the SST case discussed here and the change-point case in **?** is that the latter case relies on a continuous covariate that may be sampled from any point on the unit interval, whereas in the SST case the choice set $d = 1, 2, \ldots, D$ is discrete and held fixed in the assymptotics.

In a way, $\hat{d}_M$ is similar to $\hat{d}_{0.25}$, the largest d for which $p_d > \alpha = 0.25$, since both penalize p-values lower than 0.25. However, for a given set of p-values, $\hat{d}_M \leq \hat{d}_{0.25}$. To see this, note that for all $d > \hat{d}_{0.25}$, $p_d < 0.25$, so every summand $p_d - 1/4$ after $\hat{d}_{0.25}$ is negative. Therefore, the maximum of $\sum_{i\leq d}(p_i - 1/4)$ must occur with $d \leq \hat{d}_{0.25}$. While $\hat{d}_{0.25}$ and $\hat{d}_M$ may often coincide, there are also cases in which $\hat{d}_M < \hat{d}_{0.25}$. This will happen when the maximum value of the random walk in (**??**), occurs prior to $\hat{d}_{0.25} - 1$. Then, $\hat{d}_M$ will only equal $\hat{d}_{0.25}$ if $p_{\hat{d}_{0.25}} - 1/4 > max_d\{\sum_{i\leq d}(p_i - 1/4)\} - \sum_{i\leq \hat{d}_{0.25}-1}(p_i - 1/4)$.

In general, the difference between $\hat{d}_\alpha$ and $\hat{d}_M$ will be most pronounced when the distributions of p-values for $d > d^*$ are not monotonically decreasing in probability—in such a scenario, it is most probable that an errent p-value for $d >> d^*$ will be greater than α; one p-value determines $\hat{d}_\alpha$, but $\hat{d}_M$ relies on the entire set of p-values.

4 Edge Testing

Hypothesis tests will only be able to reject specifications that violate assumption $\mathcal{A}$ if they have sufficient power. One simple way to boost power in some scenarios is to focus hypothesis tests on the parts of the data and model that is most likely to violate the assumption—the difference between the analysis under d and under $d-1$.

In contrast to specification tests that researchers use to check fully-specified models, and are designed to check the model as a whole, SSTs are an explicit and planned part of the model selection process. That being the case, their focus should be on differences between potential specifications, rather than on overall suitability. We refer to the former as "edge testing," since it focuses hypothesis tests on edge cases, and the latter "total testing."

When decisions d determine which data are included in the analysis, as in RDD bandwidth selection, the choice between edge and total testing is a choice between null hypotheses to test. The edge null is:

$$H_{0d}^{edge} : \mathcal{A}\text{is true for } i \in \rangle_d \tag{2}$$

whereas the total null is

$$H_{0d}^{tot} : \mathcal{A}\text{is true for } i \in \mathcal{I}_d \tag{3}$$

where, as above, $\mathcal{I}_d = \rangle_1 \cup \dots \rangle_d$, all data included in specification d.

For instance, in selecting a bandwidth for an RDD, as in **?**, researchers test for, say, equality of means of a covariate x between treated subjects, with running variable values

R at one side of the cutoff, and control subjects with R on the other side. Here d indexes candidate bandwidths, $max|R-c| = bw_d$. Then $\rangle_d = \{i : |R_i - c| = bw_d\}$ and $\mathcal{I} = \{i : |R_i - c| \leq bw_d\}$. Therefore, $H_{0d}^{tot} : \mathbb{E}[x|0 < R-c \leq bw_d] = \mathbb{E}[x| - bw_d \leq R-c < 0]$ and $H_{0d}^{edge} : \mathbb{E}[x|R-c = bw_d] = \mathbb{E}[x|R-c = -bw_d]$.[1] For the sake of demonstration, say $var(x) = \sigma^2$. For $d \leq d^*$, $\mathbb{E}[x||R-c| = bw_d] = 0$ but for $d > d^*$ $\mathbb{E}[x|R-c = bw_d] = \tau$ and $\mathbb{E}[x|R-c = -bw_d] = -\tau$. Further, say there are $n_d = n_0$ at each possible bandwidth bw_d. For $d = d^* + 1$, testing H_{0d}^{tot} means comparing the means of two samples of size $(d^*+1)n_0$, each with standard deviation $\sqrt{\sigma^2 + \tau^2(1 - 1/dstar)}$ and with means $\pm\tau/dstar$. On the other hand, a test of H_{0d}^{edge} compares the means of two smaller samples, each of size n_0, with standard deviation σ and means $\pm\tau$. As long as $d^* > 1$, the power of a t-test for H_{0d}^{edge} will be greater than the power for H_{0d}^{tot},[2] better allowing a the SST procedure to distinguish between d^* and d.

The SST case in which a researcher uses SSTs to choose a model specification is analogous. For instance, **?** provide several examples of sequential tests to determine the order of an autoregressive process, and both edge and total tests are represented. The "general to specific" liklihood ratio test compares the determinants of estimated innovation covariance matrices of models assuming order d and $d-1$, respectively. It tests the null hypothesis H_{0d}^{edge} that the order of the process is d against the alternative that the order is $d-1$. They also discuss a "Portmanteau" Lagrange Multiplier test that tests H_{0d}^{tot}, that there is no serial correlation in the residuals of a VAR(d) model. In this case, the difference between edge and total tests lies in the alternative—an edge test compares model d to model $d+1$, whereas a total test compares all candidate model to the same alternative.

[1] These are simplifications of Assumption 4 in **?**, which treats x as fixed, not random.

[2] The non-centrality parameter in the H_{0d}^{tot} test is $\frac{2\tau/d^*}{\sqrt{\sigma^2+\tau^2(1-1/d^*)}} \frac{\sqrt{(d^*+1)n_0}}{\sqrt{2}}$ and the non-centrality parameter in the H_{0d}^{edge} test is $\frac{2\tau}{\sigma}\frac{\sqrt{n_0}}{\sqrt{2}}$

5 Two Data Examples

To illustrate these ideas—edge testing, and the change point and hypothesis testing approaches to selecting d—we will briefly illustrate them with two data examples. The two examples correspond to the two broad categories of specification we have discussed: selecting data to analyze and selecting a model specification.

5.1 SSTs in Regression Discontinuity Bandwidth Selection: Estimating the Effect of Academic Probation on College GPAs

At many universities, students who fail to achieve a minmum GPA are put on academic probation (AP) (See, e.g. **?**). This provides them access to a set of resources designed to address personal issues that may be hindering their performance. Perhaps more importanty, AP is a threat—students on AP who do not improve are subject to disciplinary measures such as suspension. **?** recognized that AP can form a regression discontinuity design (RDD), in which treatment is a function of a "running variable" with a pre-determined cutoff. Specifically the treatment Z, students' AP status, is (almost) a deterministic function of a "running variable" R, students' grade-point-averages (GPAs). Students with a GPA below a pre-determined cutoff, $R < c$, are put on AP. That being the case, students with GPAs just below c may be comparable to students with GPAs just above c—comparing these two sets of students allows researchers to estimate the effect of AP on outcomes Y. The challange becomes defining "just above" and "just below"; SSTs may be able to play a role here.

For example, **?** (CFT) suggests directly comparing the outcomes of subjects with R very close to c, say with $R \in [c - bw, c + bw]$ for some bandwidth $bw > 0$ To choose bw, CFT uses pre-treatment covariates $\boldsymbol{X}$, and covariate balance tests range of candidate bandwidths. For each possible bw, they test the hypothesis that the covariates are balanced:

$$\boldsymbol{X} \perp\!\!\!\perp Z | R \in [c - bw, c + bw] \tag{4}$$

and choose the largest bandwidth in which (**??**) cannot be rejected.

Bandwidth selection for RDDs, and the role of covariate balance tests, encompasses a growing literature. As its name suggests, regression discontinuity typically relies on regression modeling: the goal is to model Y as a function of R on either side of c to estimate the average treatment effect for subjects with R in an infinitesimally-small interval around the cutoff c (See **?**). In contrast, CFT dispenses with regression alltogether. One popular way to ensure robustness to model misspecification is to fit the regression models to a subset of the data with R in a window around c. A number of methods exist to choose an optimal bandwidth bw—the width of the window—that is both large enough to allow for precise effect estimation but small enough to ensure robustness. **?** suggest using non-parametric estimates of the curvature of the regression function of Y on R, combined with local linear regression, to choose a bw that minimizes mean-squared-error. However, other authors have suggested choosing bw (or an analogous quantity) based on SSTs, including **?**, which presents a Bayesian approach analogous to CFT's, **?**, which discusses the use of robust regression models, and **?**, which proposes a method to estimate effects for subjects with R farther from c. In the latter paper, SSTs do not test covariate balance, but the irrelevence of R conditional on covariates X, for subjects in a given bandwidth.

This section will illustrate several approaches to SSTs in the context of estimating the effect of AP for first year college students on subsequent GPAs. For the sake of simplicty, the discussion will be limited to CFT's general approach to regression discontinuity designs; however, many of the SST methods can be extended to other RDD analyses. In their analysis, CFT considered a set of seven covariates: students' high-school GPA (expressed

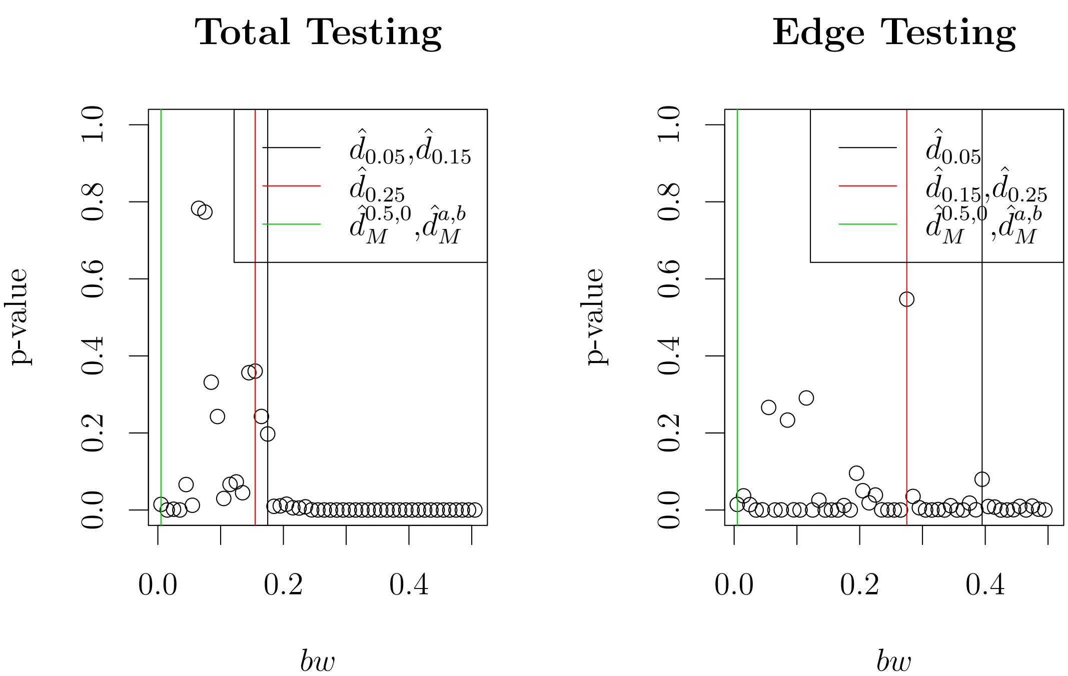

Figure 1: P-values from total and edge testing for balance in all seven covariates from the LSO analysis. Vertical lines denote bandwidth choices using different criteria.

in percentiles), age at college metriculation, number of attempted credits, gender, native language (English or other), birth place (North America or other) and university campus (the university consisted of three campuses). A version of Hotellings T^2 test that models treatment assginment Z, and not X, as random (**?**) is used to test balance. The resulting p-values, Total Testing—testing hypotheses (4)—and Edge Testing, testing $H_{0bw} : X \perp\!\!\!\perp Z||R| = bw$, are plotted in Figure **??**.

In this case, total and edge testing paint similar pictures: covariates are imbalanced for most bandwidths. Both $\hat{d}_M^{0.5,0}$ and $\hat{d}_M^{a,b}$, marked with a green vertical lines in Figure **??**, select the lowest possible bandwidth. Since the first p-value 0.015 is below any α considered, a strategy that chooses the last non-rejected bandwidth (denoted above as $\hat{d}_{\tilde{\alpha}}$) rejects every bandwidth. According to these methods, the CFT method is unsuitible for this dataset.

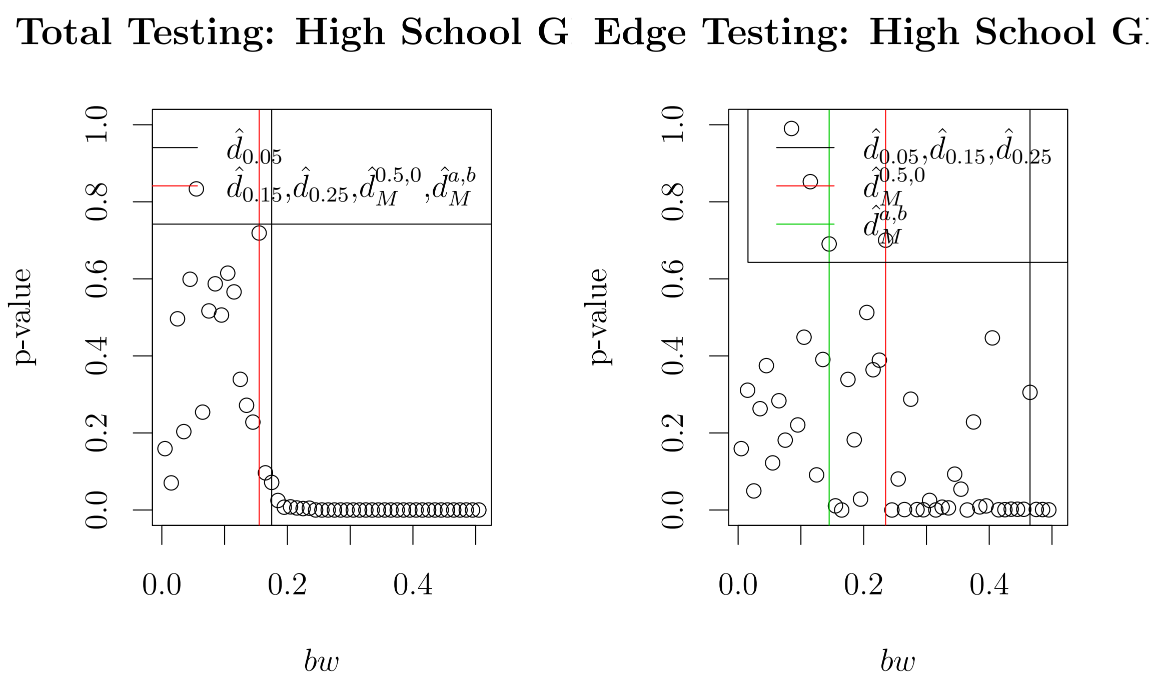

Figure 2: P-values from total and edge testing for balance in all seven covariates from the LSO analysis. Vertical lines denote bandwidth choices using different criteria.

However, the scattered large p-values at some bandwidths lead versions of $\hat{d}_\alpha$ to select larger bandwidthds. This fact illustrates a weakness of $\hat{d}_\alpha$: strong evidence against a model specification will be discarded in the presence of one favorable p-value.

To better illustrate differences between the window selection strategies, we consider the covariate high school GPA alone. Since the outcme of interest is itself a GPA, prior measures of GPA are arguably the most relevant and important to control. P-values from total and edge tests of balance in high school GPA are displayed in Figure 2. Fortunately for the illustration here, high school GPA may be balanced for small bws. P-values from total and edge testing in 2 are markedly different. Edge testing shows some bandwiths, considered in isolation, appear to plausibly satisfy covariate balance, while others do not. On the other hand, the p-values from total testing are more nearly monotonic: imbalance in high school

GPA at lower bandwidths causes balance tests to reject at higher bandwidths as well. The opposite also occurs: for instance, at bandwidth 0.155, total testing gives 0.719, whereas edge testing gives 0.011. The explanation is that high school GPA at that bandwidth is indeed imbalanced, but in the opposite direction of other imbalances: the AP students at bw =0.155 had *higher* high school GPAs than those who were not on AP, to an extent that counteracted imbalances at smaller bandwidths.

That said, edge testing suggested larger bandwidths for almost all procedures.

This example suggests that edge testing might only be suitible for change-point-based window selectors. Futher, the more flexible $\hat{d}_M^{a,b}$ outperformed the assymptotic estimator $\hat{d}_M^{0.5,1}$ by choosing b =NA, so that even moderately small p-values suggested departures from covariate balance.

6 A Simulation Study

7 Discussion