# Flatness-Robust Critical Bandwidth*

Scott Kostyshak†

August 29, 2019

**Abstract**

Critical bandwidth (CB) can be used to test the multimodality of densities and regression functions, as well as for clustering methods. This paper proposes a solution to the well-known problem that CB tests are generally inconsistent if the function of interest is constant ("flat") over an interval. The solution, flatness-robust CB (FRCB), exploits the fact that the problem manifests only from regions consistent with the null hypothesis, and thus identifying and excluding them does not alter the null or alternative sets. I provide sufficient conditions for consistency of FRCB, and simulations of a test of regression monotonicity demonstrate the finite-sample properties of FRCB compared with CB for various regression functions. I illustrate the usefulness of FRCB with an empirical analysis of the monotonicity of the conditional mean function of radiocarbon age with respect to calendar age.

Keywords: Critical bandwidth, non-parametric regression, multimodality testing, bootstrap, regression monotonicity

*I thank Douglas Turner for excellent comments.

†University of Florida, Department of Economics. Email: skostyshak@ufl.edu.

# 1 Introduction

Critical bandwidth (CB), introduced by Silverman (1981), is used to test the multimodality of densities and regression functions, to detect mixture distributions, and as a component of clustering methods. CB is discussed in monographs on the bootstrap as an innovative way to carry out hypothesis tests of multimodality,[1] and has applications across many fields. Examples of the diverse applications of CB include identifying the number of growth spurts in height (Harezlak and Heckman, 2001), exploring the multimodality of labor productivity (Henderson et al., 2008), identifying determinants of the U-shape of life satisfaction over the life cycle (Kostyshak, 2017), and testing the ecological niche separation of species with similar dietary requirements (Cumming et al., 2017).

Despite the variety of null hypotheses that CB tests can be used for, the robustness of results based on CB tests has been limited because they are generally inconsistent if the true function of interest is constant ("flat") over an interval. With finite sample sizes, this flatness problem can arise in additional situations, even when the true regression function does not contain a perfectly flat region: If the derivative is close to zero in absolute value, the CB test can suffer from low power and incorrect size, as shown in the simulations in Section 4. Although flatness exclusion is a reasonable assumption in some situations, many real-world relationships have regions over which the regression function is constant or has a small derivative. For example, the height of humans is increasing in the early ages, flattens out, and then decreases.

Flatness exclusion was assumed in the original proof of CB consistency (Silverman, 1983),[2] and has been noted by Mammen et al. (1992), Cheng and Hall (1998), and Hall and Heckman (2000).[3] Further, new techniques that propose improvements to standard CB also assume flatness exclusion (e.g., Ameijeiras-Alonso et al. (2017)). With no solution available, the flatness problem has motivated the development of non-CB methods that do not suffer from the same inconsistency (e.g., Cheng and Hall (1998), Hall and Heckman (2000), and Gijbels et al. (2000)). Such methods are solutions for specific situations in which CB has been used (e.g., regression monotonicity), but no solution has been proposed that applies to all situations for which the flexible CB framework can be used.

This paper proposes a flatness-robust critical bandwidth (FRCB) test that is valid without the assumption of flatness exclusion. Since flat regions do not contradict the null hypothesis about the number of peaks of the function of interest, a consistent test can be achieved by identifying and excluding such regions.[4] Formally, consider a parameter space of functions $\Theta$. Let $\hat{f}_f$ be a semi-parametric estimator of $f \in \Theta$, where the subscript emphasizes that the distribution of the estimator depends on the true parameter $f$. Suppose that $\hat{f}_f$ is consistent for all $f \in \Theta$. That is, suppose that for a relevant $\mathcal{X}$, $sup_{x \in \mathcal{X}} \left| \hat{f}_f(x) - f(x) \right| \xrightarrow{p} 0$.[5] Let the pair $(\Theta^N, \Theta^A)$ partition $\Theta$ into the sets corresponding to the null and alternative hypotheses. Consider a subspace $\Theta^F \subset \Theta$ of functions with flat regions that intersects both $\Theta^N$ and $\Theta^A$.[6] Standard CB techniques exclude $\Theta^F$ from the parameter space. That is, even if for all $f \in \Theta^F$, $\hat{f}_f$ converges in probability to $f$, standard CB tests are still not consistent for this parameter subspace. In this paper, I allow $\Theta^F$ to be part of the parameter space and provide a transformation $\mathbf{T}$ such that for all $f \in \Theta^N$, $\mathbf{T}\hat{f}_f \xrightarrow{p} \tilde{f}_f$ for some $\tilde{f}_f \in \Theta^N \setminus \Theta^F$; and for all $f \in \Theta^A$, $\mathbf{T}\hat{f}_f \xrightarrow{p} \tilde{f}_f$ for some $\tilde{f}_f \in \Theta^A \setminus \Theta^F$. In other words, for any element in $\Theta$, after the transformation standard CB tests are asymptotically valid, since the probability limit of the estimator is not in $\Theta^F$.

[1] See, e.g., Section 16.5, "Testing multimodality of a population," of Efron and Tibshirani (1993); Hall (1992, p. 153); and Davison and Hinkley (1997, p. 189).

[2] Silverman (1983) assumes bounded support of the density, and that the derivative has "no multiple zeros."

[3] The widespread recognition of the inconsistency is reflected by the multiple terms used to refer to it: the "spurious mode problem," the "flatness problem," and the "boundary problem."

[4] In this paper, when discussing the number of modes (equivalently, "peaks"), I am referring to the weak concept: Whenever discussing monotonicity I am referring to weak monotonicity; and I refer to a function $f$ as unimodal if it is weakly unimodal, i.e., there exists a value $m$ for which $f$ is monotonically increasing for $x \leq m$ and monotonically decreasing for $x \geq m$.

[5] Measurability is assumed throughout this paper.

[6] For examples of elements in the null and alternative sets that have flat regions, see Figure 1.

The rest of the paper is organized as follows. Section 2 reviews the standard CB test and the flatness problem, which causes inconsistency, low power, and incorrect size. Section 3 introduces the FRCB test, which forms the core of the paper, and provides sufficient conditions for consistency of the test. Section 4 compares the performance of FRCB to CB in simulations of a test of regression monotonicity. Section 5 illustrates the usefulness of FRCB with an empirical analysis of the monotonicity of the conditional mean function of radiocarbon age with respect to calendar age. Section 6 concludes. Proofs of the theorems are in the Appendix.

# 2 Critical Bandwidth

In this section, I define the class of CB test statistics and give specific examples that fit in the framework. Let $D$ be a random matrix of data with support $\mathscr{D}$, $\mathcal{F}$ the parameter space, $f \in \mathcal{F}$ the function of interest, $G \subseteq \mathbb{R}$ the grid over which $f$ is estimated,[7] and $\mathsf{H}$ the space of smoothing parameters. Let $\hat{f} : \mathscr{D} \times \mathsf{H} \times \mathbb{R} \to \mathcal{F}$ be the semi-parametric estimator that maps data and smoothing parameters to the parameter space, evaluated on a grid. We are interested in null hypotheses of the form $f \in \mathcal{H}_0$ for some $\mathcal{H}_0 \subseteq \mathcal{F}$, such that for $D \subseteq \mathscr{D}$, $\{h, h'\} \subseteq \mathsf{H}$, the following property holds:[8]

$$\hat{f}(D, h, G) \in \mathcal{H}_0,\ h' > h \implies \hat{f}(D, h', G) \in \mathcal{H}_0.$$

For example, if $\mathcal{H}_0$ is the class of monotone functions, satisfying this property requires that if a smoothing parameter value yields an estimate that is monotone, increasing the smoothing parameter from that value also yields an estimate that is monotone. The CB test statistic is then defined as

$$h^{CB}_{stat}(D, G) = \min \left\{ h \in \mathsf{H} \mid \hat{f}(D, h, G) \in \mathcal{H}_0 \right\}.$$

The null hypothesis is rejected if $h^{CB}_{stat}(D, G)$ is too large, which occurs when only a large smoothing parameter can force the estimator into being consistent with the null hypothesis. Critical values are determined from a bootstrap.[9]

**Examples** The seminal CB test of Silverman (1981) corresponds to $\mathcal{F}$ as the class of densities, $\mathcal{H}_0$ as the class of densities with less than a specified number of modes, $\hat{f}$ as a kernel density estimator, and $\mathsf{H}$ as the set of bandwidths. The test of Bowman et al. (1998) corresponds to $\mathcal{F}$ as the class of regression functions, $\mathcal{H}_0$ as the collection of monotone regression functions, and $\hat{f}$ as a non-parametric regression estimator. Harezlak and Heckman (2001) extends $\mathcal{H}_0$ to regression functions of an arbitrary number of modes, and $\hat{f}$ to any estimator with a smoothing parameter. Kostyshak (2017) extends $\mathcal{H}_0$ to quasi-convex and quasi-concave regression functions, and $\hat{f}$ to estimators of generalized additive models, in order to test multivariate hypotheses.

## 2.1 The flatness problem

To gain intuition for why a flat region causes problems for CB tests, consider a simple example in the context of regression. Suppose that

$$y = f(x) + \epsilon,$$

[7] Every CB implementation uses a grid to check that a non-parametric estimate is of a certain shape. The grid can be as dense as desired. In the simulations in Section 4, using a grid of 100 and a grid of 500 yield the same results.

[8] For cases in which this property is not satisfied exactly, simulations suggest that contradictions of the property rarely occur. For example, Bowman et al. (1998) find that "out of the total of 90,000 simulations only two cases were discovered where the estimated regression curve was monotonic at one bandwidth and nonmonotonic at a higher one. This behavior is therefore extremely rare and its effect on the test procedure will be negligible."

[9] To create the bootstrapped data sets, a sample (with replacement) is taken from the residuals and added to $\hat{f}(D, h^{CB}_{stat}, G)$. For more details, see Bowman et al. (1998) and Kostyshak (2017).

where $x$ is independent of $\epsilon$, and $E(\epsilon) = 0$. The CB test is inherently flawed if there exist $f_1 \in \mathcal{H}_0, f_2 \in \mathcal{H}_0^C$ that yield indistinguishable test statistics, even in arbitrarily large samples. For an example of how this situation can occur, suppose that $\mathcal{H}_0$ is the class of monotone functions and that $f_1$ is monotone and contains an interval, say $[a, b]$, over which it is constant. Suppose that $f_2$ is the same as $f_1$, except that it has a dip (violation of monotonicity) in some interval, say $[c, d]$. Since the derivative of $f_1$ is zero over $[a, b]$, for small $h$, $\hat{f}_1 \in \mathcal{H}_0^C$ with high probability because even a moderate amount of variance of $\hat{f}_1$ leads to estimates of the derivative on both sides of zero. As $h$ increases, the variance of $\hat{f}_1$ decreases, and eventually $\hat{f}_1 \in \mathcal{H}_0$ with probability increasing toward 1.[10] Define $h^1_{stat}$ to be the CB test statistic associated with $\hat{f}_1$. The identification issue described above is binding if $\hat{f}_2(D, h^1_{stat}, G)$ is monotone over $[c, d]$: In this case, the violation of monotonicity of $f_2$ over $[c, d]$ was smoothed away as a result of the flat interval over $[a, b]$.

A similar manifestation of the flatness problem can distort the size of the CB test and is not driven inherently by the value of the test statistic, as above, but rather by the bias of the bootstrap. The validity of the CB bootstrap depends on $\hat{f}(D, h^{CB}_{stat}, G)$ being close to $f$ if the null is true, because the bootstrapped data sets are constructed based on $\hat{f}(D, h^{CB}_{stat}, G)$. However, in the presence of flat regions, $\hat{f}(D, h^{CB}_{stat}, G)$ is smoother than $f$. The function $\hat{f}(D, h^{CB}_{stat}, G)$ has a non-zero derivative over the flat interval, and thus the bootstrapped test statistics are smaller, on average, than the original test statistic, yielding a low $p$-value. Even in regions outside the flat region, the over-smoothed $\hat{f}(D, h^{CB}_{stat}, G)$ could be substantially different from $f$, and thus the bootstrapped test statistic can be further biased.

If $f'(x)$ does not equal exactly zero but is small in absolute value over an interval, although the CB test is consistent, it can have poor finite-sample properties. Simulations in Section 4 include such a function under the alternative (see $m_4$ in Figure 1), as well as functions that suffer from the flatness problem under the null (see $m_1$ and $flat_1$). The next section introduces a solution to the flatness problem.

# 3 Flatness-robust Critical Bandwidth

In this section I introduce flatness-robust critical bandwidth (FRCB), which transforms standard CB into a test that is robust to the function of interest having flat regions. We now specify sufficient conditions for consistency of FRCB.

**Definition 1** (**flatness exclusion**)**.** There is no interval $[a, b]$ such that for $x \in [a, b]$, $f'(x) = 0$.

**Assumption 1.** *CB is consistent under flatness exclusion.*

Sufficient conditions for a consistent CB test depend on the function of interest (e.g., density or regression function), as well as the specific estimator used. Consistency under flatness exclusion is assumed in this paper, in order to keep the theorems and proofs general enough to be valid for any specific CB test. For specific proofs of consistency of CB under flatness exclusion, see Silverman (1983) for the case of densities and Kostyshak (2017) for the case of regression functions. Theorems in this section demonstrate that FRCB extends consistency to hold without flatness exclusion.

**Assumption 2.** *$f$ is continuously differentiable.*

**Assumption 3.** *For any $a \in (0, 1)$, there exists a simultaneous confidence band for $f'$, denoted $(U_{f'}, L_{f'})$ , such that $P[L_{f'} \leq f' \leq U_{f'}] = a + o(1)$ and $\|U_{f'} - L_{f'}\|_\infty = o_p(1)$.*

The FRCB test at level $\alpha$ is implemented by the following algorithm:

**Algorithm 1.**

[10] For most semi-parametric estimators, as $h$ tends to infinity, $\hat{f}_1$ becomes linear.

1. *Calculate* $L_{f'}(x)$ *and* $U_{f'}(x)$.[11]

2. *Construct* $\hat{G}^{NF} = \{x \in G \mid 0 \notin [L_{f'}(x), U_{f'}(x)]\}$.

3. *Perform a standard CB test on* $\hat{G}^{NF}$ *at level* $\alpha$.

**Definition 2** (FRCB)**.** The FRCB test statistic is defined as

$$h_{stat}^{FRCB} = h_{stat}(D, \hat{G}^{NF}),$$

where $h_{stat}$ is the standard CB statistic, $D$ is the data, and $\hat{G}^{NF}$ is the grid as constructed in step 2 of Algorithm 1.

**Theorem 1.** *FRCB has asymptotic level* $\alpha$.

*Proof.* Proved in Appendix A.2. □

**Theorem 2.** *FRCB has asymptotic power 1.*

*Proof.* Proved in Appendix A.3. □

**Theorem 3.** *Under flatness exclusion,* $h_{stat}^{FRCB}$ *has the same asymptotic distribution as* $h_{stat}^{CB}$.

*Proof.* Proved in Appendix A.4. □

**Software implementation** FRCB is implemented in the R package `qmutest` (Kostyshak, 2018), which makes it easy for practitioners to apply FRCB to the testing of monotonicity, quasi-convexity, and quasi-concavity of regression functions. The bootstraps used in standard CB and in step 2 of Algorithm 1 are easily parallelized. Compared with other semi-parametric methods, the algorithms are fast and the test is practical when working with millions of observations.

# 4 Simulations

This section discusses the results of simulations of a test of regression monotonicity, and compares rejection rates of the standard CB test with those of the FRCB test for the regression functions shown in Figure 1. The regressors are all uniformly distributed and the additive error term is independent and normally distributed with standard deviation 0.25. The tests were performed using the R package `qmutest` to carry out Algorithm 1 with local polynomial regression as the estimator and thin plate regression splines for simultaneous confidence bands of the derivative.[12] 2,000 bootstrap replications were used for the core CB algorithm and for step 2 of Algorithm 1, with an $\alpha_n^{flat}$ sequence of $n^{-1/2}$. The same estimation method is used for the application in the next section.

Table 1 shows rejection rates at the 0.05 level of CB and FRCB tests for the regression functions graphed in Figure 1. Rows above the solid horizontal line are cases in which the null hypothesis is true. In the case of a flat line ($m_1$), the CB test rejects the null hypothesis of monotonicity more than 10% of the time (twice the nominal level) across all sample sizes. A worse scenario for CB is $flat_1$, for which the test rejects in almost half of the samples with a sample size of 100, and the CB rejection proportion tends to 1 in probability as the sample size increases. For both $m_1$ and $flat_1$, the FRCB test rejects no more than 6% of the time across all sample sizes.

CB tests are known to be conservative, and FRCB tests inherit this property. Even with the correction to conservativeness of CB tests proposed by Hall and York (2001), which is used in the

[11] These are constructed at level $\alpha_n^{flat}$. For details on the selection of $\alpha_n^{flat}$, see Appendix A.

[12] This combination of estimators was chosen because local polynomial regression is well-known and thin plate regression splines have nice properties for estimating the derivative and are the default in R's `mgcv` package Wood (2019). For more information, see Wood (2017). The smoothing parameter for the thin plate regression is chosen by generalized cross-validation.

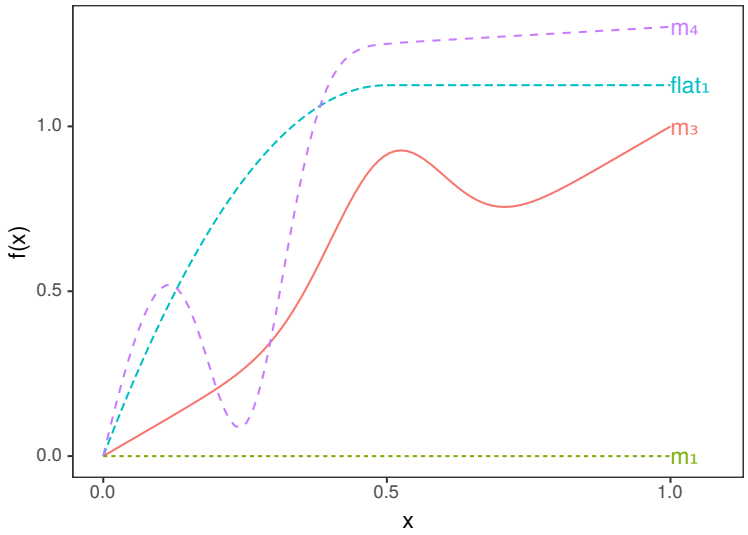



Figure 1: True Regression Functions

See Appendix B for analytical definitions and origins of functions in this figure.

Table 1: Rejection Rates of CB and FRCB

| | sample size | | | | | |
|---|---|---|---|---|---|---|
| $H_0$ | 50 | 100 | 250 | 500 | 1,000 | 2,000 |
| $M_{CB}(X_{m_1})$ | 0.14 | 0.12 | 0.13 | 0.14 | 0.17 | 0.12 |
| $M_{FRCB}(X_{m_1})$ | 0.01 | 0.01 | 0.01 | 0.00 | 0.00 | 0.00 |
| $M_{CB}(X_{flat1})$ | 0.45 | 0.49 | 0.65 | 0.73 | 0.83 | 0.94 |
| $M_{FRCB}(X_{flat1})$ | 0.06 | 0.06 | 0.04 | 0.03 | 0.04 | 0.04 |
| $M_{CB}(X_{m_3})$ | 0.16 | 0.17 | 0.36 | 0.67 | 0.94 | 1.00 |
| $M_{FRCB}(X_{m_3})$ | 0.07 | 0.15 | 0.42 | 0.79 | 0.98 | 1.00 |
| $M_{CB}(X_{m_4})$ | 0.37 | 0.37 | 0.47 | 0.52 | 0.60 | 0.64 |
| $M_{FRCB}(X_{m_4})$ | 0.17 | 0.49 | 0.49 | 0.66 | 0.88 | 1.00 |

This table shows the proportion of times each test rejects the null hypothesis of regression monotonicity at the 0.05 level. In the $H_0$ column, the notation "$M_T(Z)$" means a null hypothesis of regression monotonicity in $Z$, tested using the test $T$ (CB or FRCB). Rows above the solid horizontal line are cases in which the null hypothesis is true. For the shapes of the regression functions, see Figure 1. Proportions are based on 10,000 simulations.

simulations, the test is still conservative: Although the test was run at the nominal 5% level, the rejection rate of the FRCB test is closer to 0% than 5% for $m_1$. More complex calibrations of the level could be explored but are beyond the purview of this paper.

The rows in the table below the solid line are cases in which the null hypothesis is false. Both CB and FRCB asymptotically reject, but for the case of $m_4$, the power of the CB test converges more slowly to 1 than the FRCB test because of the flatness problem that is triggered by the small derivative of the second half of the function. For $m_3$, in contrast to $m_4$, there is no large region with a derivative close to zero. The rejection rates of CB and FRCB thus become similar for large sample sizes for this regression function. The intuition for this similarity is captured in Theorem 3.

Two competing effects determine whether the power of FRCB is larger than CB for small sample sizes. Lack of precision when identifying flat regions can cause non-flat regions—which might contain evidence against the null—to be excluded from the $\hat{G}^{NF}$ grid (see step 2 of Algorithm 1) that is passed to the CB test. This effect explains why CB performs better than FRCB in the case of $m_3$ for sample sizes 50 and 100: The region just after 0.5 is not estimated precisely enough to determine whether it is non-flat, and thus the grid on which $h^{CB}_{stat}$ is calculated might not contain a strong violation of monotonicity. On the other hand, a derivative that is relatively small in absolute value results in low power for the same reason that flatness results in asymptotic inconsistency. This second effect explains why, for the case of $m_4$, FRCB has higher power than CB for sample sizes 100 and larger. Which of the two effects dominates depends on the shape of the regression function and the precision of the non-parametric estimator, which is influenced by factors such as the sample size and the distribution of the error term.

The results of the simulations show how the flatness problem can arise under both the null and the alternative, and when there are no flat parts but there are parts with a small first derivative. It is not surprising that FRCB has lower power than CB when the sample size is 50: Typically in statistics, making fewer assumptions has a cost of lower power. That FRCB has higher power for most sample sizes in the simulation is a bonus feature, and highlights its usefulness in situations in which the derivative is not exactly zero but is small enough to trigger symptoms of the flatness problem.

# 5 Application

We examine a subset of radiocarbon data from Irish oak trees, published by Pearson and Qua (1993) and previously analyzed in the context of regression monotonicity using CB by Bowman et al. (1998). The dataset consists of the radiocarbon age and the calendar age.[13] Calibration of this relationship is important, as the radiocarbon age is often used to predict the calendar age when the calendar age is unknown. The raw data are shown in Figure 2, along with a non-parametric fit and corresponding simultaneous confidence band at the 95% level. The relationship is known to be non-monotone,[14] and thus is an interesting case study for whether the non-monotonicity can be picked up in a small sample. A visual inspection of the simultaneous confidence band does not suggest strong evidence against monotonicity. However, careful visual inspection of the data points themselves does hint at non-monotone fluctuations in the underlying regression function.

A standard CB test, using the same non-parametric estimation method as the simulations, results in a $p$-value of 0.012. Fitted values corresponding to three selected smoothing parameters, including the test statistic value, are shown in Figure 3a. Consider the line corresponding to a smoothing parameter of 37.8, which is considerably smaller than the test statistic value of 56.3. The 37.8 line is monotone, except for the section with calendar ages less than 5,250. This shows that the flatness problem could be binding, i.e., both monotone and non-monotone regression functions exist that would yield the same distribution of the test statistic. Another indication that flatness could be a problem for this

[13] A measure of precision of the radiocarbon dates is also included in the dataset, but for simplicity the variable is ignored in this paper since Bowman et al. (1998) found that incorporating the measure did not substantively change their results.

[14] In the radiocarbon dating literature, these fluctuations are sometimes referred to as "wiggles" or "de Vries effects." For more information, see Taylor and Bar-Yosef (2014, pp. 53–54).

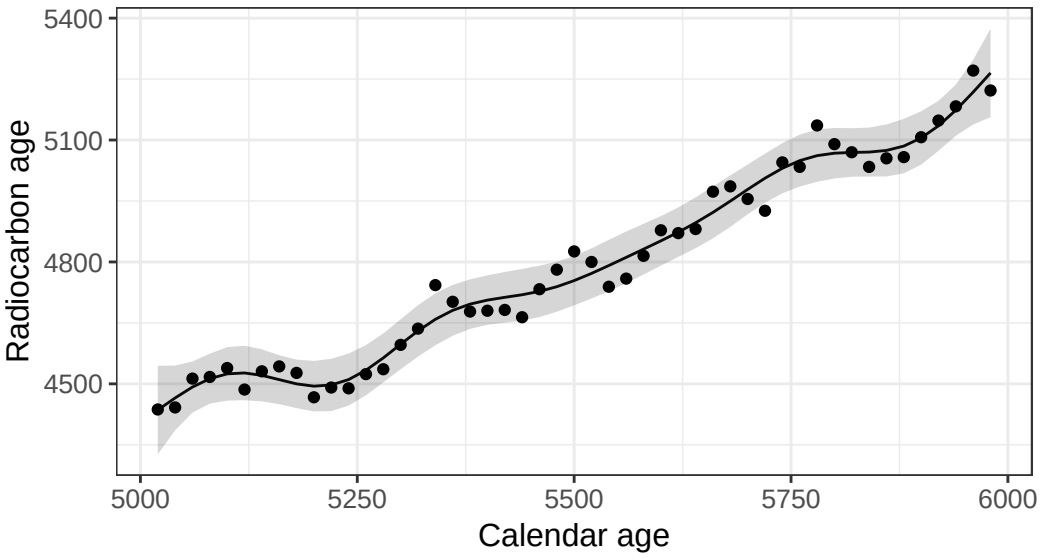



Figure 2: Raw data with non-parametric point and interval estimates of the conditional mean function

This figure shows a non-parametric fit and corresponding simultaneous confidence band at the 95% level. The estimates are from thin plate regression splines with the smoothing parameter chosen by generalized cross-validation. The simultaneous confidence band is constructed from an empirical bootstrap with 2,000 replications; the critical value is the bootstrapped 0.95-quantile of the maximal t-statistic.

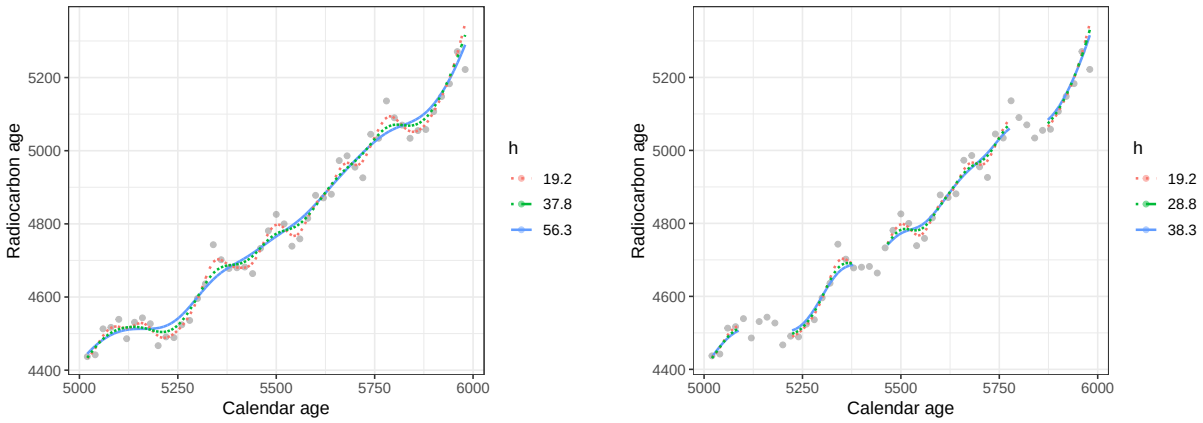
application is that the three fitted value curves have different signs of the derivative over considerable portions of the segment before 5,250. In summary, the small CB $p$-value could be driven by the 12 data points with values less than 5,250, and it is not clear whether those data points do indeed carry enough information to suggest such a strong rejection of monotonicity.

Given the concerns regarding the flatness problem described above, it is interesting to consider using FRCB to examine whether there exists strong evidence of non-monotonicity when regions with a small derivative (e.g., the region before 5,250) are detected and excluded. Since the sample size is small, a large value for $\alpha_n^{flat}$ must be used. The FRCB test statistic is 38.3, which is smaller than the CB statistic (56.3). The FRCB test statistic is always smaller than the CB test statistic, because the set of FRCB inequality constraints is a subset of the CB inequality constraints. In this particular application the difference is large, which is not surprising given the flatness concerns. The fitted value curves corresponding to smoothing parameters from the FRCB test are shown in Figure 3. The gaps between the line segments are a result of the filtering of the CB grid from step 2 of Algorithm 1. The FRCB $p$-value is 0.048, and thus the non-monotone fluctuations in the sample provide some evidence of non-monotonicity in the population.

# 6 Conclusion

This paper provides an asymptotic solution to the fundamental problem that in some situations, even with a large sample size, CB does not provide correct inference. Further, simulations show that FRCB performs better than CB in some finite-sample situations, even when no regions are exactly flat. The application to radiocarbon data provides insight on the flatness problem in an applied situation, and shows how FRCB can be used to provide inference that is robust to flat regions.

By removing an assumption that in practice might be violated, FRCB offers a robust improvement, and, in turn, opportunities to further explore and apply the class of CB tests.




(a) Standard CB

(b) FRCB

Figure 3: Fits for various smoothing parameters, CB and FRCB

# References


Ameijeiras-Alonso, Jose, Rosa M Crujeiras, and Alberto Rodríguez-Casal (2017). "Mode Testing, Critical Bandwidth and Excess Mass." In: *arXiv preprint arXiv:1609.05188v2*.

Bowman, AW, MC Jones, and Irène Gijbels (1998). "Testing Monotonicity of Regression." In: *Journal of Computational and Graphical Statistics* 7.4, pp. 489–500.

Cheng, M-Y and Peter Hall (1998). "Calibrating the Excess Mass and Dip Tests of Modality." In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 60.3, pp. 579–589.

Cumming, Graeme S, Dominic AW Henry, and Chevonne Reynolds (2017). "A Framework for Testing Assumptions About Foraging Scales, Body Mass, and Niche Separation Using Telemetry Data." In: *Ecology and Evolution*.

Davison, Anthony Christopher and David Victor Hinkley (1997). *Bootstrap Methods and their Application*. Cambridge University Press.

Efron, Bradley and Robert J Tibshirani (1993). *An Introduction to the Bootstrap*. Vol. 57. Monographs on Statistics and Applied Probability. Chapman & Hall/CRC.

Ghosal, Subhashis, Arusharka Sen, and Aad W Van Der Vaart (2000). "Testing Monotonicity of Regression." In: *Annals of Statistics*, pp. 1054–1082.

Gijbels, Irène et al. (2000). "Tests for Monotonicity of a Regression Mean with Guaranteed Level." In: *Biometrika* 87.3, pp. 663–673.

Hall, Peter (1992). "The Bootstrap and Edgeworth Expansion." In:

Hall, Peter and Nancy E Heckman (2000). "Testing for Monotonicity of a Regression Mean by Calibrating for Linear Functions." In: *Annals of Statistics*, pp. 20–39.

Hall, Peter and Matthew York (2001). "On the Calibration of Silverman's Test for Multimodality." In: *Statistica Sinica*, pp. 515–536.

Harezlak, Jaroslaw and Nancy E Heckman (2001). "CriSP: A Tool for Bump Hunting." In: *Journal of Computational and Graphical Statistics* 10.4, pp. 713–729.

Henderson, Daniel J, Christopher F Parmeter, and R Robert Russell (2008). "Modes, Weighted Modes, and Calibrated Modes: Evidence of Clustering using Modality Tests." In: *Journal of Applied Econometrics* 23.5, pp. 607–638.

Kostyshak, Scott (June 2017). "Non-Parametric Testing of U-Shaped Relationships." In: *SSRN*. URL: `https://ssrn.com/abstract=2905833`.

— (2018). *qmutest*. R Package Version 0.2. URL: `https://github.com/scottkosty/qmutest`.

Mammen, Enno, James S Marron, and Nick I Fisher (1992). "Some Asymptotics for Multimodality Tests Based on Kernel Density Estimates." In: *Probability Theory and Related Fields* 91.1, pp. 115–132.

Pearson, Gordon W and Florence Qua (1993). "High-Precision 14 C Measurement of Irish Oaks to Show the Natural 14 C Variations from AD 1840–5000 BC: A Correction." In: *Radiocarbon* 35.1, pp. 105–123.

R Core Team (2019). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. Vienna, Austria. URL: `https://www.R-project.org/`.

Silverman, Bernard W (1981). "Using Kernel Density Estimates to Investigate Multimodality." In: *Journal of the Royal Statistical Society*. B 43, pp. 97–99.

— (1983). "Some Properties of a Test for Multimodality Based on Kernel Density Estimates." In: *Probability, Statistics and Analysis* 79, pp. 248–259.

Simonsohn, Uri (June 2017). "Two-Lines: A Valid Alternative to the Invalid Testing of U-Shaped Relationships with Quadratic Regressions." In: *SSRN*. URL: `https://ssrn.com/abstract=3021690`.

Taylor, RE and O Bar-Yosef (2014). *Radiocarbon Dating. 2nd edn. Walnut Creek*. California: Left Coast Press.

Van der Vaart, Aad W (1998). *Asymptotic Statistics*. Vol. 3. Cambridge University Press.

Wood, S.N (2017). *Generalized Additive Models: An Introduction with R*. 2nd ed. Chapman and Hall/CRC.

— (2019). *mgcv*. R Package Version 1.8.28. URL: `https://CRAN.R-project.org/package=mgcv`.

# A Proofs

## A.1 Preliminaries

The following definitions are used to simplify the proofs below. $L_{f'}$ and $U_{f'}$ are as defined in assumption (3), and $\hat{G}^{NF}$ is as constructed in Algorithm 1. Define $R^{CB}$ to equal 1 if the standard CB test rejects, and 0 otherwise. Similarly, define $R^{FRCB}$ to equal 1 if the FRCB test rejects, and 0 otherwise. Let $G^F$ and $G^{NF}$ partition $G$ such that $G^F = \{x \in G \mid f'(x) = 0\}$, $G^{NF} = \{x \in G \mid f'(x) \neq 0\}$. Construct $\alpha_n^{flat}$ to be any sequence that satisfies both $\alpha_n^{flat} \to 0$ and assumption (3); that is, if we define $d_n(\alpha_n^{flat}) = \|U_{f'} - L_{f'}\|_\infty$ to be the sequence of stochastic sup-norm lengths of the confidence band, we require that both $\alpha_n^{flat} \to 0$ and $d_n(\alpha_n^{flat}) \xrightarrow{p} 0$.[15]

## A.2 Proof of Theorem 1

*Proof.* Under the null hypothesis,

$$\begin{aligned} P\left[R^{FRCB} = 1\right] = &\; P\left[R^{CB} = 1 \mid \hat{G}^{NF} \cap G^F = \emptyset\right] P\left[\hat{G}^{NF} \cap G^F = \emptyset\right] + \\ &+ P\left[R^{CB} = 1 \mid \hat{G}^{NF} \cap G^F \neq \emptyset\right] P\left[\hat{G}^{NF} \cap G^F \neq \emptyset\right] \\ \leq &\; \alpha\left(1 - \alpha_n^{flat}\right) + P\left[R^{CB} = 1 \mid \hat{G}^{NF} \cap G^F \neq \emptyset\right] \alpha_n^{flat} \\ \to &\; \alpha, \end{aligned}$$

because $\alpha_n^{flat} \to 0$. □

[15] Intuitively, we require $\alpha_n^{flat}$ to converge to 0, but at a rate that is sufficiently slow that enough precision is still gained. The rate will depend on the particular CB test.

### A.3 Proof of Theorem 2

*Proof.* Define $A$ to be the event $\left\{G^{NF} = \hat{G}^{NF}\right\} = \left\{\hat{G}^{NF} \cap G^F = \emptyset,\, G^{NF} \subseteq \hat{G}^{NF}\right\}$. Then,

$$\begin{aligned} P\left[A^C\right] \leq\ & P\left[\hat{G}^{NF} \cap G^F \neq \emptyset\right] + P\left[G^{NF} \not\subseteq \hat{G}^{NF}\right] \\ =\ & \left(\alpha_n^{flat} + o(1)\right) + o(1) \\ \to\ & 0, \end{aligned}$$

where the $o(1)$ terms come from assumption (3). Then, under the alternative,

$$\begin{aligned} P\left[R^{FRCB} = 1\right] =\ & P\left[R^{CB} = 1 \mid A\right] P\left[A\right] + P\left[R^{CB} = 1 \mid A^C\right] P\left[A^C\right] \\ \to\ & 1, \end{aligned}$$

because $P\left[R^{CB} = 1 \mid A\right] \to 1$ by assumption (1), and $P\left[A^C\right] \to 0$ as shown above. □

### A.4 Proof of Theorem 3

*Proof.* It is sufficient to show that $\left|h_{stat}^{FRCB} - h_{stat}^{CB}\right| \xrightarrow{p} 0$.[16] Define

$$\begin{aligned} \mathcal{C}(x) &= \left\{y \in \mathbb{R} \mid L_{f'}(x) \leq y \leq U_{f'}(x)\right\}, \\ I &= \bigcup_{g \in G} \mathcal{C}(g). \end{aligned}$$

Consider that under flatness exclusion,

$$P\left[\hat{G}^{NF} = G\right] = P\left[0 \notin I\right].$$

Define $B$ to be the event $\{f'(g) \in \mathcal{C}(g) \text{ for all } g \in G\}$. $P[B] = 1 - \left(\alpha_n^{flat} + o(1)\right)$ by step (1) of Algorithm 1 and assumption (3). Define $a = \min_{g \in G} |f'(g)|$. Under flatness exclusion, $a > 0$. Define $Q(x) = |U_{f'}(x) - f'(x)| \vee |L_{f'}(x) - f'(x)|$. Then,

$$P\left[0 \in I | B\right] \leq P\left[\max_{g \in G} Q(g) \geq a\right],$$

which converges to 0 by assumption (3). It follows that

$$\begin{aligned} P\left[0 \in I\right] &= P\left[0 \in I | B\right] P\left[B\right] + P\left[0 \in I | B^C\right] \left(1 - P\left[B\right]\right) \\ &= P\left[0 \in I | B\right] \left[1 - \left(\alpha_n^{flat} + o(1)\right)\right] + P\left[0 \in I | B^C\right] \left(\alpha_n^{flat} + o(1)\right), \end{aligned}$$

where $P\left[0 \in I | B\right] \to 0$ by above, and $\alpha_n^{flat} \to 0$ by construction. Then, $P\left[0 \in I\right] \to 0$, so $P\left(\hat{G}^{NF} = G\right) \to 1$, and thus

$$P\left[h_{stat}^{CB}\left(D, \hat{G}^{NF}\right) = h_{stat}^{CB}\left(D, G\right)\right] \to 1.$$ □

## B Simulation Functions

This section provides the analytical definitions and origins of the functions graphed in Figure 1. The definitions are as follows:

[16] See, e.g., Theorem 2.7 of Van der Vaart (1998).

$$
\begin{aligned}
m_1(x) &= 0 \\
m_3(x) &= x + 0.415e^{-50(x-0.5)^2} \\
m_4(x) &= \begin{cases} 10(x-0.5)^3 - e^{-100(x-0.25)^2} & \text{if } x < 0.5, \\ 0.1(x-0.5) - e^{-100(x-0.25)^2} & \text{otherwise} \end{cases} \\
flat_1(x) &= \begin{cases} 0.5x(1-x) \cdot 9 & \text{if } x < 0.5, \\ 0.5^3 \cdot 9 & \text{otherwise} \end{cases}
\end{aligned}
$$

The $m_n$ functions were explored by Ghosal et al. (2000), and some were also studied in other papers that tested for monotonicity of regression functions: At least one of $m_3$ and $m_4$ was used by Bowman et al. (1998), Hall and Heckman (2000), and Kostyshak (2017). Functions similar to $flat_1$ were explored by Simonsohn (2017).