Comparison of confidence interval methods and I(M)SE-optimal bandwidth estimators for kernel density estimation

Research Module in Econometrics and Statistics Term Paper

in the M.Sc. Economics Programm

at the Rheinische Friedrich-Wilhelms-Universität Bonn

Module organiser: JProf. Dr. Claudia Noack

Written in the Winter Semester 2023/24 by:

Justin Franken, Torben Haferkamp and Niklas Niedermeier

Contents

1 Introduction

Kernel Density Estimation (KDE) is a statistical method for estimating the probability density function (PDF) of a random variable from observed data points. Unlike parametric methods that assume a specific functional form for the underlying distribution, KDE provides a flexible and non-parametric approach to density estimation. KDE works by applying a kernel function to each data point, summing these functions, and then weighting the resulting sum to generate a smooth density estimate. One simple application is data visualization, which aids in understanding the distribution and identifying anomalies or outliers in random variables. Beyond visualization, KDE finds applications in various fields such as classification in machine learning, estimating spatial density distributions of points or events, and performing non-parametric regression.

While the selection of a kernel function is significant, it usually holds less importance compared to choosing the bandwidth. It's important to note that differences between kernel functions are minor in finite sample cases, and tend to be smoothed out asymptotically. However, the bandwidth plays an important role in determining the shape, width, smoothness, and accuracy of the density estimate. Increasing the bandwidth yields smoother but more biased estimators, while decreasing it may cause overfitting, reflecting the classic bias-variance trade-off in non-parametric methods. Thus, careful consideration of the bandwidth is essential for achieving an optimal KDE, with extensive literature available on this topic. In this study, we examine three distinct bandwidth selection approaches introduced by Silverman (1986) (Rule of Thumb), Sheather and Jones (1991a) (Plug-In), and Rudemo (1982), Stone (1984), and Bowman (1984) (Least Squares Cross-Validation). These methods all aim to minimize the integrated (mean) squared error of the KDE.

To evaluate the quality of our estimators, we assess their performance by computing confidence intervals for the estimated density functions. An inherent challenge in kernel density estimation is that the estimator is unbiased only asymptotically, resulting in a bias in finite sample cases. In this work, we build upon literature from Hall (1992) (bias correction undersmoothing) and Calonico, Cattaneo, and Farrell (2018) (robust bias correction) who recognized this bias problem in how they build confidence intervals.

Given that these methods, similar to the bandwidth selection methods discussed previously, still rely on asymptotics, we are interested in their performance in finite sample scenarios. We therefore explore different combinations of confidence interval computation and bandwidth selection methods across four different density functions. Our Monte Carlo Simulation reveals marginal differences among bandwidth selection methods, with the Silverman rule of thumb demonstrating slightly superior efficacy in attaining coverage probability and resultant interval lengths. Furthermore, robust bias correction and undersmoothing emerge as the most effective methods for confidence interval computation, consistently providing valid intervals compared to bias correction.

This paper follows a similar approach to Calonico, Cattaneo, and Farrell (2018), comparing various confidence interval computation methods across four probability densities.

We proceed as follows: Section 2 outlines the kernel density estimator and its statistical properties. Section 3 introduces three bandwidth methods, while Section 4 presents three confidence interval methods. Section 5 details our Monte Carlo Simulation, and in Section 6, we apply our findings to real household income data. Finally, Section 7 presents our conclusions.

2 Basic Theory

This section establishes a theoretical foundation, beginning with the derivation of the Kernel Density Estimator for the univariate case, analyzing its statistical properties, and examining the multivariate case, with a focus on the univariate case in subsequent chapters.

2.1 The kernel density estimator

Consider the random sample $\{X_i\}_{i=1}^n$ originating from the unknown PDF f_X and cumulative distribution function F_X (CDF). Fix some interior point $X \in \mathbb{R}$ of the random sample. Our objective is to determine f_X. To initiate this process, we substitute X with $\tilde{X} = X + h\epsilon$, where h is a positive real number and ϵ is a random variable with a known PDF $k(\cdot)$ and CDF $K(v) = \int_{-\infty}^{v} k(\epsilon)d\epsilon$. Let's begin by examining the CDF of $\tilde{X}$:

$$\begin{aligned} F_{\tilde{X}}(x) &= P\left(X + h\epsilon \leq x\right) \\ &= P\left(\epsilon \leq \frac{x - X}{h}\right) \\ &= E_X\left(P\left(\epsilon \leq \frac{x - X}{h} \mid X\right)\right) \\ &= E_X\left(K\left(\frac{x - X}{h}\right)\right) \\ &= \int K\left(\frac{x - z}{h}\right) f_X(z)\, dz \end{aligned}$$

Since we know that $f_{\tilde{X}}(x) = \frac{dF_{\tilde{X}}(x)}{dx}$, we can write

$$f_{\tilde{X}}(x) = \frac{1}{h} \int k\left(\frac{x - z}{h}\right) f_X(z)\, dz.$$

If we substitute the expected value with respect to X by the sample average we will get:

Definition 1. *The kernel density estimator*

$$\hat{f}(x) = \frac{1}{nh} \sum_{i=1}^{n} k\left(\frac{x - X_i}{h}\right).$$

In chapter 3 we discuss how to determine h optimally. The following (typical) form is assumed for our kernel function:

Assumption 1. *$k(v)$ is a bounded nonnegative kernel function and has the following properties:*

1. $\int k(v)\,dv = 1$ *(i.e. can serve as a probability function)*
2. $k(v) = k(-v)$ *(i.e. symmetric around zero)*[1]
3. $\int v^2 k(v)\,dv > 0$ *(i.e. v has positive variance)*

Examples of kernel functions that fall under our assumption 1 are:

Kernel	$k(v)$
Uniform	$\frac{1}{2}\ \mathbf{1}(\lvert v\rvert \leq 1)$[2]
Triangular	$(1-\lvert v\rvert)\ \mathbf{1}(\lvert v\rvert \leq 1)$
Epanechnikov	$\frac{3}{4}((1-v^2))\ \mathbf{1}(\lvert v\rvert \leq 1)$

2.2 Statistical Properties

In this section, we derive the bias and variance of the kernel density estimator and show MSE consistency. For this purpose, we introduce the following assumptions about the underlying data and the function $f(x)$:

Assumption 2. *$X_1, ..., X_n$ are i.i.d observations.*

Assumption 3. *$f(x)$ is a three times differentiable PDF.*

Assumption 4. *$f(x)$ is bounded.*

Assumption 5. *x is an interior point in the support of X.*

Using assumptions 1–5, we can derive the bias term for $\hat{f}(x)$:

$$\text{Bias}\left(\hat{f}(x)\right) = \frac{h^2}{2} f^{(2)}(x) \int v^2 k(v)\,dv + O(h^3). \tag{1}$$

The expression indicates the estimator's bias is primarily influenced by $f(x)$'s second derivative, making it biased. Yet, for $h \to 0$, it becomes asymptotically unbiased.
Under the same assumptions we can derive the variance of the estimator by

$$\text{Var}\left(\hat{f}(x)\right) = \frac{1}{nh}\left\{ f(x) \int k^2(v)\,dv + O(h) \right\}. \tag{2}$$

[1] Note that when we derive the bias of the kernel density estimator, we make use of the implication $\int vk(v)\,dv = 0$.

[2] Here, the notation $\mathbf{1}(|v| \leq 1)$ represents the indicator function, taking the value 1 when $|v|$ is less than or equal to 1, and 0 otherwise.

The derivation of the variance shows that it converges to zero if $nh \to \infty$ and $h \to 0$. A detailed derivation of equation (1) and (2), based on the specified assumptions, is provided in the *appendix*. Next, we describe the Mean Squared Error (MSE) asymptotics, a metric for the kernel density estimator's accuracy, representing the squared deviation between the estimated $\hat{f}(x)$ and true $f(x)$ density functions. The MSE at x can be decomposed to:

$$\text{MSE}\left(\hat{f}(x)\right) = \left[\text{Bias}\left(\hat{f}(x)\right)\right]^2 + \text{Var}\left(\hat{f}(x)\right). \tag{3}$$

Plugging in the provided terms for bias and variance, this yields:

$$\text{MSE}\left(\hat{f}(x)\right) = \left[\frac{h^2}{2} f^{(2)}(x) \int v^2 k(v) dv + O(h^3)\right]^2 + \frac{1}{nh}\left\{f(x) \int k^2(v) dv + O(h)\right\}. \tag{4}$$

For the sake of clarity, we will now focus on the term of the squared bias. The squaring of the given bias term results in

$$\left[\frac{h^2}{2} f^{(2)}(x) \int v^2 k(v) dv + O(h^3)\right]^2 = \frac{h^4}{4}\left[f^{(2)}(x) \int v^2 k(v) dv\right]^2 + O(h^6) + O(h^5). \tag{5}$$

As $h \to 0$, higher-order terms such as $O(h^6)$ and $O(h^5)$ diminish more rapidly compared to lower-order terms like $O(h^4)$. Therefore, we can simplify the squared bias as follows, whereas $o(h^4)$ indicates all terms of higher order than h^4:

$$\left[\frac{h^2}{2} f^{(2)}(x) \int v^2 k(v)\, dv + O(h^3)\right]^2 \approx \frac{h^4}{4}\left[f^{(2)}(x) \int v^2 k(v)\, dv\right]^2 + o(h^4). \tag{6}$$

Similarly, for $n \to \infty$ and $h \to 0$, we can identify $(nh)^{-1}$ as the dominant rate for the variance and simplify the expression for the variance:

$$\frac{1}{nh}\left\{f(x) \int k^2(v)\, dv + O(h)\right\} \approx \frac{1}{nh}\left\{f(x) \int k^2(v)\, dv\right\} + o\left(\ (nh)^{-1}\right). \tag{7}$$

Combining the simplified expressions for squared bias (6) and variance (7), we obtain the final form:

$$\begin{aligned}\text{MSE}\left(\hat{f}(x)\right) &= \frac{h^4}{4}\left[f^{(2)}(x) \int v^2 k(v)\, dv\right]^2 + \frac{1}{nh}\left\{f(x) \int k^2(v)\, dv\right\} + o\left(h^4 + (nh)^{-1}\right) \\ &= O\left(h^4 + (nh)^{-1}\right). \end{aligned} \tag{8}$$

Intuitively, the influence of reducing the bandwidth can be described on the one hand as having a finer approximation to $f(x)$ (lower bias), but on the other hand having the risk (for fixed n) of fitting more noise (higher variance). The MSE converges to 0 if $nh \to \infty$ and $h \to 0$, under these conditions $\hat{f}(x)$ is a consistent estimator of $f(x)$, provided the bandwidth does not converge too fast to 0.

2.3 Multivariate kernel density estimator

Our findings can be extended to the multivariate case with some adjustments to our introduced notations. Consider the i.i.d. sample $\{X_i\}_{i=1}^n$ of q-vectors, where $X_i \in \mathbb{R}^q$ (for some $q > 1$) and the probability density function is given by $f(x) = f(x_1, x_2, ..., x_q)$. Let X_{is} be the s-th component of the vector X_i with $s = 1, ..., q$. The kernel density estimator now takes the following form:

$$\hat{f}(x) = \frac{1}{nh_1 \cdots h_q} \sum_{i=1}^{n} \tilde{k}\left(\frac{X_i - x}{h}\right) \tag{9}$$

Where the kernel function $\tilde{k}\left(\frac{X_i - x}{h}\right) = k\left(\frac{X_{i1} - x_1}{h_1}\right) * \cdots * \ k\left(\frac{X_{iq} - x_q}{h_q}\right)$ (referred to as the "product kernel function") and $k(\cdot)$ is the univariate kernel function satisfying our assumption 1. It can be demonstrated[3] that the multivariate kernel density estimator exhibits the following statistical properties:

$$\text{Bias}\left(\hat{f}(x)\right) = \frac{\int v^2 k(v) dv}{2} \sum_{s=1}^{q} h_s^2 f^{(2)}(x) + O\left(\sum_{s=1}^{q} h_s^3\right), \tag{10}$$

$$\text{Var}\left(\hat{f}(x)\right) = \frac{1}{nh_1 \cdots h_q} \left[\left(\int k^2(v) dv\right)^q f(x) + O\left(\sum_{s=1}^{q} h_s^2\right)\right]. \tag{11}$$

Similar to the univariate case, the bias and variance exhibit the same properties. As $max_{1 \le s \le q} h_s \to 0$, the multivariate kernel density estimator, $\hat{f}(x)$, asymptotically approaches an unbiased estimator. Similarly, the multivariate variance converges to 0 when $(nh_1 \cdots h_q) \to \infty$ and $max_{1 \le s \le q} h_s \to 0$. If we again use (3), we can summarize (10) and (11) to

$$\text{MSE}\left(\hat{f}(x)\right) = O\left(\left(\sum_{s=1}^{q} h_s^2\right)^2 + \left(\frac{1}{nh_1 \cdots h_q}\right)\right) \tag{12}$$

As the multivariate MSE shows, if $n \to \infty$, $max_{1 \le s \le q} h_s \to 0$ and $(nh_1 \cdots h_q) \to \infty$ we get $\hat{f}(x) \to_{MSE} f(x)$ and therefore have also a consistent multivariate estimator.

3 Bandwidth Selection

In the forthcoming subsection, our focus will be on determining optimal bandwidths for our kernel density estimator. To achieve this, our aim is to minimize the integrated (mean) squared error (I(M)SE) of $\hat{f}(x)$. It's worth noting that we concentrate on minimizing I(M)SE rather than standard mean squared error (MSE). While MSE quantifies the average squared difference between the estimated density function and the true density function at individual points in the domain, our goal is to assess the integrals of the squared errors across the entire data range, providing a comprehensive measure of the

[3]The proofs for (10) and (11) have been included in the *appendix*.

average quality of the density estimate.

3.1 Rule of Thumb

The bandwidth of the kernel, denoted as h, is a critical parameter in kernel density estimation, significantly influencing the accuracy of the resulting estimate. It serves as a trade-off between under-smoothing and over-smoothing. Specifically, a small h for a given sample size n results in low bias but high variance, potentially introducing spurious data artifacts (under-smoothing). Conversely, a large h leads to high bias but low variance, potentially obscuring underlying data structures (over-smoothing). The challenge lies in finding an optimal bandwidth, h_{opt}, which necessitates optimizing the overall function $\hat{f}$, rather than individual points $\hat{f}(x)$. The IMSE is a common criterion for this purpose:

$$\begin{aligned}\text{IMSE}(\hat{f}) &= \int \text{MSE}(\hat{f}\,(x))dx \\ &= \int E\left(\left(\hat{f}_X(x) - f_X(x)\right)^2\right)dx \\ &= \frac{h^4}{4}\left(\int v^2 k(v)dv\right)^2 \int \left(f_X^{(2)}(x)\right)^2 dx + \frac{1}{nh}\int k^2(v)dv + o\left(h^4 + (nh)^{-1}\right) \quad (13)\end{aligned}$$

where $f^{(2)}$ denotes the second derivative of the true density function f. To determine h_{opt}, we minimize the leading terms of the IMSE with respect to h. Differentiating the IMSE with respect to h and setting the derivative to zero yields the optimal bandwidth condition. The resulting expression for h_{opt} is:

Definition 2. *Optimal bandwidth h_{opt}*

$$\begin{aligned}h_{opt} &= \left(\frac{\int k^2(v)\,dv}{n\left(\int v^2 k(v)\,dv\right)^2 \int \left(f^{(2)}(x)\right)^2\,dx}\right)^{1/5} \\ &= \left(\frac{R(k)}{\sigma_k^4 R(f^{(2)})}\right)^{1/5} n^{-1/5},\end{aligned}$$

where $R(g) = \int g^2(x)\,dx$ and $\sigma_g^2 = \int x^2 g(x)\,dx$.

Silverman (1986) suggests an approach to derive h_{opt} by assuming a parametric family for f. For instance, if f is assumed to be normally distributed with mean μ and variance σ^2, the term $f^{(2)}$ in the expression for h_{opt} can be explicitly determined:

$$\int f^{(2)}(x)^2\,dx = \sigma^{-5}\int \phi^{(2)}(x)^2\,dx = \frac{3}{8}\pi^{-1/2}\sigma^{-5} \approx 0.212\sigma^{-5}. \quad (14)$$

With a Gaussian Kernel $k(v) = \frac{1}{\sqrt{2\pi}}e^{-\frac{v^2}{2}}$, we substitute $k(v)$ and Equation (14) into the

formula for h_{opt}:

$$h_{\text{RT}} = \left(\frac{4}{3}\right)^{\frac{1}{5}} \sigma n^{-\frac{1}{5}} \approx 1.06\sigma n^{-\frac{1}{5}}. \tag{15}$$

Finally, σ can be replaced with the sample standard deviation in practical applications. This method is effective if the underlying distribution closely resembles the assumed normal distribution (Li and Racine 2006).

3.2 Plug-in methods

The key difference between the rule of thumb and plug-in methods lies in how the unknown integral $R(f^{(2)}) = \int \left(f^{(2)}(x)\right)^2$ in h_{opt}, defined in definition 2, is determined. Instead of assuming a specific density for the unknown PDF of f, plug-in methods utilize a second kernel density estimator to more accurately estimate the underlying density $R(f^{(2)})$. Consequently, the estimate $\hat{R}(f^{(2)})$ can be substituted into h_{opt}. This approach is precisely what was developed by Park and Marron (1990). They utilize the proposed estimator of Hall and Marron (1987) who showed that one can estimate $\theta_m = \int \left(f^{(m)}(x)\right)^2 dx$ by slightly changing the form of that unknown integral for the case $m = 2$ to $\hat{R}(f) = R(\hat{f}) - n^{-1}h^{-5}R(k^{(2)})$. When employing the KDE outlined in definition 2, we obtain:

$$\hat{R}(f_\alpha^{(2)}) = (n(n-1))^{-1}\alpha^{-5} \sum_{i \neq j} \sum L^{(4)} \left(\frac{X_i - X_j}{\alpha}\right) = \hat{S}_{ND}(\alpha) \tag{16}$$

Here, α represents a second bandwidth and L another kernel function for the kernel density estimate of $\hat{f}_\alpha$, which is beneficial since we can now separate both analyses from one another. ND in $\hat{S}_{ND}(\alpha)$ represents 'Non-Diagonals', indicating the exclusion of diagonal terms ($i = j$) in the double sum. In order to determine the second bandwidth estimator, Park and Marron (1990) proceeded by optimizing the following:

$$\frac{\partial \,\text{MSE}\left(\hat{S}_{ND}(\alpha)\right)}{\partial\, \alpha} \stackrel{!}{=} 0 \;\Leftrightarrow\; \alpha_1 = C_1(L)C_2(f)n^{-2/13},$$

where $C_1(L) = \left(\frac{18R(L^{(4)})}{\sigma_L^4}\right)^{1/13}$, $C_2(f) = \left(\frac{R(f)}{R^2(f^{(3)})}\right)^{1/13}$. To connect α_1 and h_{opt} note that:

$$h_{opt} = \left(\frac{R(k)}{\sigma_k^4 R(f^{(2)})}\right)^{-1/5} n^{(-1/5)} \;\Leftrightarrow\; n = \frac{R(k)}{\sigma^4 R(f^{(2)})} h_{opt}^{-5}.$$

If we substitute n in α_1 and assume for simplicity $L = k$ we get:

$$\alpha_1(h) = C_1(k)C_2(f)\left(\frac{R(k)}{\sigma^4 R(f^{(2)})}h_{opt}^{-5}\right)^{-2/13} \tag{17}$$

$$= \left(\frac{18R(K^{(4)})\sigma_k^8}{\sigma_k^4 R^2(k)}\right)^{1/13}\left(\frac{R(f)R^2(f^{(2)})}{R^2(f^{(3)})}\right)^{1/13} h_{opt}^{10/13} \tag{18}$$

$$= C_3(k)C_4(f)h_{opt}^{10/13} \tag{19}$$

Given that $\alpha_1(h)$ still relies on the unknown density f, Park and Marron (1990) suggested assuming a density for f at this point, noting that the potential influence of an incorrect f on our target variable h_{opt} diminishes adequately at this stage. Hence, $g_1(x)$ is taken as a constant probability density function[4], and the scaled density is expressed as $g_{\hat{\lambda}}(x) = g_1(x/\hat{\lambda})\frac{1}{\hat{\lambda}}$. Here, we introduce a scaling parameter, $\hat{\lambda}$, to normalize $g_{\hat{\lambda}}(x)$ to have an interquantile range or standard deviation of 1. The estimation of $\hat{\lambda}$ is subsequently performed based on the available data x. Next, substitute f in $C_4(f)$ with $g_{\hat{\lambda}}(x)$ to get $\alpha_1(h) = C_3(k)C_4(g_{\hat{\lambda}})h_{opt}^{10/13}$. Then, plug $\alpha_1(h)$ back into equation (16) to get $\hat{S}_{ND}(\alpha_1(h))$. Ultimately, the optimal bandwidth $\hat{h}_1$ suggested by Park and Marron (1990) is determined as the root (choosing the greatest one if there are multiple) of the following equation:

$$z(\hat{h}_1) = \left(\frac{R(k)}{\sigma_k^4 \hat{S}_{ND}(\alpha_1(\hat{h}_1))}\right)^{1/5} n^{-1/5} - \hat{h}_1 = 0 \tag{20}$$

One could use computational algorithms such as the Newton-Raphson method to find the root of equation (20). In Park and Marron's (1990) Theorem 3.3, the asymptotic performance of $\hat{h}_1$ is characterized as:

$$\hat{h}_1/h_{opt} = 1 + O_p(n^{-4/13})$$

To enhance the precision of bandwidth computation, we will build upon the methodology introduced by Park and Marron (1990) and further illustrate the refined approach suggested by Sheather and Jones (1991a), which enhances Park and Marron's (1990) bandwidth estimator $\hat{h}_1$. The primary objective of their enhanced estimator $\hat{h}_{PI}$ is to address the discontinuity issue that arises when $\hat{S}_{ND}$ (equation (16)) changes signs for h in equation (20), leading to a discontinuity at these points. This discontinuity presents a significant challenge for various root-finding procedures and may result in inaccurate identification of discontinuities as solutions, particularly when the discontinuity jumps over the root. In response Sheather and Jones (1991b) wanted to create a computationally more stable estimator and therefore incorporate diagonal terms in their estimate of

[4]Park and Marron (1990) recommend employing a normal density for g_1. However, they state, as demonstrated in their theorem 3.3, that the choice of density at this juncture is not critical.

$\hat{R}(f^{(2)})$:

$$\hat{S}_D(\alpha) = (n(n-1))^{-1}\alpha^{-5}\sum_{i=1}^{n}\sum_{j=1}^{n} L^{(4)}\left(\frac{X_i - X_j}{\alpha}\right), \tag{21}$$

In this scenario, the subscript D represents 'Diagonals-in', indicating that the estimator is inherently positive by construction. Consequently, it avoids encountering discontinuity issues. Sheather and Jones (1991b) observed that the discrepancy between $\hat{S}_D$ and $\hat{S}_{ND}$ comprises a positive non-stochastic term, contributing to bias. Hence, prioritizing bias minimization with an optimal bandwidth parameter α_2 supersedes achieving a balance in the bias-variance tradeoff for this specific problem[5]. Sheather and Jones (1991b) proposed the following formula for the optimal bandwidth, obtained by differentiating $MSE\left(\hat{S}_D\right)$ with respect to α, with a focus on predominantly reducing bias. This results in $\alpha_2 = \left(\frac{2L^{(4)}(0)}{\sigma_L^2}\right)^{1/7} R^{-1/7}(f^{(3)})n^{-1/7}$. If one would now again combine α_2 with h_{opt} as we did with $\alpha_1(h)$ one gets $\alpha_2(h) = \left(\frac{2L^{(4)}(0)\sigma_k^4}{R(k)}\right)^{1/7}\left(\frac{R(f^{(2)})}{R(f^{(3)})}\right)^{1/7} h^{5/7}$, where Sheather and Jones (1991b) emphasise that the insertion of a scaled model for f, akin to what we implemented for the estimator $\hat{h}_1$ (20), results in insufficient cancellation of the leading bias terms and thus a third stage in the estimation of $R(f^{(2)})$ and $R(f^{(3)})$ using the methods of Park and Marron (1990) or Hall and Marron (1987) is required. Subsequently, we can determine the optimal bandwidth $\hat{h}_{PI}$ using the same root-finding procedures as for example Newton-Raphson to solve again equation (20), but now with $\hat{S}_D(\alpha_2(\hat{h}_{PI}))$ instead of $\hat{S}_{ND}(\alpha_1(\hat{h}_1))$. The authors showed that their estimator is asymptotically superior to $\hat{h}_1$, since

$$\hat{h}_{PI}/h_{opt} = 1 + O_p(n^{-5/14})$$

3.3 Least squares cross-validation

Another method to choose the optimal bandwidth h is least squares cross-validation, which estimates the smoothing parameter h completely automatically on a data-driven basis. The method was developed by Rudemo (1982), Stone (1984) and Bowman (1984). In contrast to the plug-in method, this method minimizes the (estimated) integrated squared error,

$$\text{ISE} = \int \left[\hat{f}(x) - f(x)\right]^2 dx = \int \hat{f}(x)^2\,dx - 2\int \hat{f}(x)f(x)\,dx + \int f(x)^2\,dx,$$

[5]Please see Sheather and Jones (1991b) for a more comprehensive explanation.

where the optimal choice of the parameter h depends on the particular sample[6]. The first integral can be estimated by plugging in our estimator for $\hat{f}(x)$ and rearranging:

$$\int \hat{f}(x)^2 \, dx = \frac{1}{n^2 h^2} \sum_{i=1}^{n} \sum_{j=1}^{n} \int k\left(\frac{X_i - x}{h}\right) k\left(\frac{X_j - x}{h}\right) \, dx.$$

The second integral contains the unknown term $f(x)$, but $\int \hat{f}(x) f(x) \, dx$ can be rewritten as $\mathrm{E}_X\left(\hat{f}(x)\right)$ and estimated by its sample average: $\frac{1}{n}\sum_{i=1}^{n} \hat{f}_{-i}(X_i)$, where

$$\hat{f}_{-i}(X_i) = \frac{1}{(n-1)h} \sum_{j=1, j\neq i}^{n} k\left(\frac{X_i - X_j}{h}\right)$$

represents the usual kernel density estimator without including observation i to estimate $f(X_i)$ in an unbiased fashion[7] Our minimization problem is independent of the last ISE integral since it does not depend on h. Using the introduced estimators we obtain our loss function

$$T = \frac{1}{n^2 h^2} \sum_{i=1}^{n} \sum_{j=1}^{n} \int k\left(\frac{X_i - x}{h}\right) k\left(\frac{X_j - x}{h}\right) \, dx - \frac{2}{n(n-1)h} \sum_{i=1}^{n} \sum_{j=1, j\neq i}^{n} k\left(\frac{X_i - X_j}{h}\right),$$

where the optimal h is numerically chosen to minimize T. Define the minimizer of T as h_{cv} and the minimizer of the ISE as $\tilde{h}_{opt}$. Härdle, Hall, and Marron (1988) demonstrate that

$$\frac{h_{cv}}{\tilde{h}_{opt}} = 1 + O_p\left(n^{-1/10}\right),$$

which converges (in probability) to 1 if $n \to \infty$, but at a comparable slow rate. The literature on the least squares cross validation attempts to increase the convergence rate but involves a further set of restrictive assumptions.

4 Confidence Intervals

In this chapter, we discuss how confidence intervals can be calculated for KDE. The first problem is to estimate the distribution of $\hat{f}(x)$ appropriately in a finite sample context. In order to estimate the quantiles of the confidence intervals, it is used that $\hat{f}(x)$ is asymptotically normally distributed (as $nh \to \infty$, $h \to 0$ and $n \to \infty$)[8]. The second difficulty lies in the fact that the kernel density estimator is biased as long as we have $h \nrightarrow 0$, which is the classic case when you want to determine an I(M)SE-optimal bandwidth.

[6]This means that an h is chosen which selects $\hat{f}(x)$ as close as possible to $f(x)$ for a specific sample instead of the mean over all samples we could have obtained according to the IMSE.

[7]The leave-one-out estimation is necessary because the expectation operator assumes that X and the X_j's are independent of eachother.

[8]For the derivation, reference is made to Li and Racine (2006, pp. 28-30).

So far, two different approaches can be distinguished, which are described in detail below. Beforehand a new assumption is required for the construction of the asymptotic confidence intervals:

Assumption 6. *The kernel function $k(v)$ has support* $[0, 1]$.

We assume that assumptions 1 to 6 hold and newly introduced kernel functions additionally fulfill assumptions 1 and 6. The presentation of the methods is analogous to Calonico, Cattaneo, and Farrell (2018) with the difference that the smoothness and the degree of the kernel function are fixed (see assumptions 1).

4.1 Bias correction

The bias correction method adjusts the point estimate $\hat{f}(x)$ by subtracting an estimate of the leading bias term. We define the estimate of the leading bias term based on equation 1 as

$$\hat{\mathrm{B}}(x) = \frac{h^2}{2} \hat{f}^{(2)}(x) \int v^2 k(v)\, dv.$$

The second derivative of $\hat{f}(x)$ can be expressed as

$$\hat{f}^{(2)}(x) = \frac{1}{nb^3} \sum_{i=1}^{n} L^{(2)} \left(\frac{X_i - x}{b} \right),$$

where $L^{(2)}$ denotes the second derivative of the new introduced kernel function L along with the bandwidth $b > 0$. If we define $\rho = \frac{h}{b}$, the bias corrected estimate can now be described by

$$\hat{f}(x) - \hat{\mathrm{B}}(x) = \frac{1}{nh} \sum_{i=1}^{n} \left(K(u) - \rho^3 L^{(2)}(\rho u) \int v^2 k(v)\, dv \right).$$

4.1.1 Traditional bias correction by Hall (1992)

To derive a confidence interval around $f(x)$ for the traditional bias correction method (BC), the following studentized T-statistic is proposed in the literature:

$$T_{bc}(x) = \frac{\sqrt{nh} \left(\hat{f}(x) - \hat{\mathrm{B}}(x) - f(x) \right)}{\hat{\sigma}_{bc}},$$

where $\hat{\sigma}_{bc}^2 = (nh)\widehat{\mathrm{Var}} \left(\hat{f}(x) \right)$. The variance of $\hat{f}(x)$ is calculated by replacing the expected values in the variance formula with sample averages. This procedure is repeated in the following chapters. Using normal approximations for the T-statistics, we can determine

our bias-corrected confidence interval for each x by

$$I_{bc}(x) = \left[\hat{f}(x) - \hat{\mathrm{B}}(x) - z_{1-\frac{\alpha}{2}}\frac{\hat{\sigma}_{bc}}{\sqrt{nh}}, \hat{f}(x) - \hat{\mathrm{B}}(x) + z_{1-\frac{\alpha}{2}}\frac{\hat{\sigma}_{bc}}{\sqrt{nh}}\right].$$

In the next step, we will illustrate the asymptotic properties of the coverage probability of $I_{bc}(x)$ and specify conditions that are necessary to obtain an asymptotically valid confidence interval. Define for this purpose the scaled bias corrected estimator by $\eta_{bc} = \sqrt{nh}\left(E\left(\hat{f}(x) - \hat{\mathrm{B}}(x)\right) - f(x)\right)$. If $nh/\log(nh) \to \infty, \eta_{bc} \to 0$ and $p \to 0$, we have

$$\begin{aligned} P\left(f(x) \in I_{bc}(x)\right) = 1 - \alpha + \left\{\frac{1}{nh}q_1(K) + \eta_{bc}^2 q_2(K) + \frac{\eta_{bc}}{\sqrt{nh}}q_3(K)\right\}\frac{\phi(z_{\frac{\alpha}{2}})}{f(x)}\left(1 + o(1)\right) \\ +\rho^3(\Omega_1 + \rho^2\Omega_2)\phi(z_{\frac{\alpha}{2}})z_{\frac{\alpha}{2}}\left(1 + o(1)\right), \end{aligned} \tag{22}$$

where $\phi(z)$ is the Standard Normal distribution at quantile z. Ω_1 the covariance of $\hat{\mathrm{B}}(x)$ and $\hat{f}(x)$, Ω_2 the variance of $\hat{\mathrm{B}}(x)$, and $q_1(k)$, $q_2(k)$ and $q_3(k)$ are decribed in detail in the appendix.

4.1.2 Robust bias correction by Calonico, Cattaneo, and Farrell (2018)

Calonico, Cattaneo, and Farrell (2018) propose a robust bias corrected method (RBC) to estimate confidence intervals. Here, the authors suggest the modified studentized statistic:

$$T_{rbc}(x) = \frac{\sqrt{nh}\left(\hat{f}(x) - \hat{\mathrm{B}}(x) - f(x)\right)}{\hat{\sigma}_{rbc}},$$

where $\hat{\sigma}_{rbc}^2 = (nh)\widehat{\mathrm{Var}}\left(\hat{f}(x) - \hat{\mathrm{B}}(x)\right)$. In contrast to the classical studentized statistic, the authors consider not only the variability of $\hat{f}(x)$ but also that of $\hat{\mathrm{B}}(x)$ to closely imitate the finite-sample behaviour of bias correction. To capture the behaviour of the bias correction in finite samples more precisely, the authors allow ρ to converge to an arbitrary (non-negative) finite limit, so that the bias correction is first-order important.[9] Based on normal approximations for the T-statistics, we can derive our robust bias corrected confidence interval for each x by

$$I_{rbc}(x) = \left[\hat{f}(x) - \hat{\mathrm{B}}(x) - z_{1-\frac{\alpha}{2}}\frac{\hat{\sigma}_{rbc}}{\sqrt{nh}}, \hat{f}(x) - \hat{\mathrm{B}}(x) + z_{1-\frac{\alpha}{2}}\frac{\hat{\sigma}_{rbc}}{\sqrt{nh}}\right].$$

[9] The condition represents a significantly weaker condition compared to the traditional approach where we only get valid confidence intervals if $\rho \to 0$.

If $nh/\log(nh) \to \infty, \eta_{bc} \to 0$ and $p \to \overline{p} < \infty$, the asymptotic coverage probability is

$$P\left(f(x) \in I_{rbc}(x)\right) = 1 - \alpha + \left\{\frac{1}{nh} q_1(K) + \eta_{bc}^2 q_2(K) + \frac{\eta_{bc}}{\sqrt{nh}} q_3(K)\right\} \frac{\phi(z_{\frac{\alpha}{2}})}{f(x)} \left(1 + o(1)\right) \tag{23}$$

4.2 Undersmoothing

Another approach, well supported in the literature, is Undersmoothing (US). This method involves selecting a bandwidth smaller than that which minimizes the IMSE by choosing a faster convergence rate. For example, in the case of the Silverman estimator, a faster convergence rate can be achieved by tuning the estimator's rate depending on a hyperparameter: $h_{\text{silverman}}(\lambda) = 1.06\hat{\sigma} n^{-\frac{1}{5\lambda}}$ for $\lambda \in (0, 1)$. Lowering λ will make the bias converge to zero more rapidly. The test statistic without bias becomes:

$$T_{us}(x) = \frac{\sqrt{nh}\left(\hat{f}(x) - f(x)\right)}{\hat{\sigma}_{us}},$$

with $\hat{\sigma}_{us}^2 = \widehat{\text{Var}}\left(\hat{f}(x)\right)$. Again, the normal approximation of the statistic $T_{us}(x)$ provides the upper and lower limits of the confidence interval:

$$I_{us}(x) = \left[\hat{f}(x) - z_{1-\frac{\alpha}{2}} \frac{\hat{\sigma}_{us}}{\sqrt{nh}}, \hat{f}(x) + z_{1-\frac{\alpha}{2}} \frac{\hat{\sigma}_{us}}{\sqrt{nh}}\right].$$

Scaling the bias of the density estimator $\eta_{us} = \sqrt{nh}\left(E\left(\hat{f}(x)\right) - f(x)\right)$ and letting $\frac{nh}{\log(nh)} \to \infty$ and $\eta_{us} \to 0$, leads us to the asymptotic coverage probability of $I_{us}(x)$:

$$P\left(f(x) \in I_{us}(x)\right) = 1 - \alpha + \left\{\frac{1}{nh} q_1(K) + \eta_{us}^2 q_2(K) + \frac{\eta_{us}}{\sqrt{nh}} q_3(K)\right\} \frac{\phi(z_{\frac{\alpha}{2}})}{f(x)} \left(1 + o(1)\right). \tag{24}$$

5 Monte Carlo Simulations

The aim of the following Monte Carlo Simulation is to contrast the described methodologies for constructing confidence intervals and estimating bandwidth within a finite sample context. The first step is to analyze in more detail how the choice of methods influences the coverage probability of the confidence intervals and from which sample size one can speak of valid confidence intervals (chapter 5.1). In the next step, we examine the extent to which the interval length of the confidence intervals differs (chapter 5.2). Finally, we investigate if the results for the confidence interval method robust bias correction changes when we additionally undersmooth our estimate for the leading bias term $\hat{\text{B}}(x)$ chapter 5.3). To compare the different methods, 5000 realizations of the following four distributions are simulated:

1. Model (Gaussian Density): $x \sim \mathcal{N}(0,1)$
2. Model (Skewed Unimodal Density): $x \sim \frac{1}{5}\mathcal{N}(0,1) + \frac{1}{5}\mathcal{N}\left(\frac{1}{2}, \left(\frac{2}{3}\right)^2\right) + \frac{3}{5}\mathcal{N}\left(\frac{13}{12}, \left(\frac{5}{9}\right)^2\right)$
3. Model (Bimodal Density): $x \sim \frac{1}{2}\mathcal{N}\left(-1, \left(\frac{2}{3}\right)^2\right) + \frac{1}{2}\mathcal{N}\left(1, \left(\frac{2}{3}\right)^2\right)$
4. Model (Asymmetric Bimodal Density): $x \sim \frac{3}{4}\mathcal{N}(0,1) + \frac{1}{4}\mathcal{N}\left(\left(\frac{3}{2}\right), \left(\frac{1}{3}\right)^2\right)$.

You can find visualizations of the densities discussed above in the appendix (figure 2). The density is evaluated at $x = \{-2, -1, 0, 1, 2\}$ for sample sizes from 25 to 500 in increments of 25. The bandwidth h is calculated in three different ways:

1. IMSE optimal rule of thumb estimator (Silverman 1986): $\hat{h}_{RT}$
2. IMSE optimal plug-in estimator (Sheather and Jones 1991b): $\hat{h}_{PI}$
3. ISE optimal cross-validation estimator (Rudemo 1982): $\hat{h}_{CV}$

As recommended by Calonico, Cattaneo, and Farrell (2018), the bandwidth for calculating the leading bias term of the bias-corrected confidence interval methods is computed initially by setting $\hat{b} = \hat{h}$ in chapter 5.1 and 5.2. In addition, we choose $\alpha = 5\%$ and fix all kernels to the Epanechnikov kernel.

5.1 Analysis of coverage probability

Upon reviewing the coverage probability, we consistently observe that higher coverage is achieved when the observed density function has a higher probability. This trend holds true regardless of the distribution being examined. As demonstrated by equations (22), (23), and (24) in Chapter 4, the probability of the true density lying within our confidence interval depends on the true density function $f(x)$ in the denominator. Consequently, a higher density associated with this point results in an increased coverage probability. One might think of this as indicating a faster approach towards infinity at these points, given the greater number of observations clustered around them. This phenomenon is illustrated, for instance, in figure 3, where we quickly achieve the theoretically correct coverage probability of 95% for points x at -1, 0, and 1 (with $50 \lesssim n \lesssim 100$), while it progresses more slowly for x at -2 and 2 (with $n \approx 500$).

In our examination of BC for estimating confidence intervals, we observe that valid confidence intervals are not achieved for $n \leq 500$ across all examined distributional models (see for example figure 11 or figure 15). This observation aligns with theoretical expectations, as in this subsection, we set $\hat{h} = \hat{b}$, causing $\hat{\rho} = \frac{\hat{h}}{\hat{b}}$ not to converge to 0. Hence, for finite sample sizes, the likelihood of our confidence interval containing $f(x)$ diminishes, as implied by equation (22). This reduction in probability stems from the additional noise introduced, represented by the term $\rho^3(\Omega_1 + \rho^2\Omega_2)\phi(z_{\frac{\alpha}{2}})z_{\frac{\alpha}{2}}(1 + o(1))$, which, in our scenario, is negative.

In the context of employing RBC to compute confidence intervals, we observed that we achieve valid confidence intervals for all models at the latest point for $n \approx 500$, except for model 2 with $x = -2$ (for model 2 coverage probabilities see figure 7). This observation aligns with theoretical expectations since we now incorporate the variability of our estimator in computing its standard errors. Consequently, we achieve, at least asymptotically, valid confidence intervals for any ρ, as demonstrated in equation (23) (even in the case where $\hat{b} = \hat{h}$). In our study, we already attain valid confidence intervals for $n \leq 500$. We suppose that the reason model 2 with $x = -2$ exhibits invalid confidence intervals is due to its very low probability mass at this specific point, resulting in a scarcity of observation points at $x = -2$ and thus slower convergence rates for valid confidence intervals. This situation is in line with the general observations we described earlier. Additionally, we noted a marginal impact of the chosen bandwidth on the determination of valid confidence intervals. It was evident that the rule of thumb method suggested by Silverman (1986) appeared to exhibit a slight advantage for evaluation points associated with a low probability mass. This phenomenon is best illustrated in Figure 7, where the Silverman bandwidth (depicted by the yellow line) slightly outperforms other bandwidth-selecting methods, particularly for points such as $x = \{-2, -1\}$, which, within model 2, possess the lowest probabilities. We believe that the rule of thumb method tends to outperform slightly due to its streamlined computational process, resulting in reduced estimation errors, particularly in situations with limited data points where the probability distribution is small.

In our investigation of US, we focus on a modified rule of thumb for bandwidth selection, as previously discussed in chapter 4.2, where we review λ values ranging from $\lambda = \{0.6, 0.7, 0.8, 0.9, 1\}$. Decreasing λ corresponds to reducing $\hat{h}_{\text{silverman}}(\lambda)$. We observed that valid confidence intervals are achieved for all distributions, reaching this validity at the latest point for $n \approx 500$, except for model 2 at $x = -2$ (for coverage probabilities of model 2, refer to figure 8). The speed at which the confidence interval approaches the desired theoretical threshold of 95% depends heavily on the tuning parameter λ. A λ closer to 1, akin to employing the standard rule of thumb bandwidth estimator, consistently yields the best results. However, this advantage diminishes as n approaches 500. This phenomenon can be explained by equation (24). Asymptotically we have that, $nh \to 0$, but in our finite sample analysis, we encounter a challenge: with a fixed, finite n, the term $\left\{ \frac{1}{nh} q_1(K) + \eta_{us}^2 q_2(K) + \frac{\eta_{us}}{\sqrt{nh}} q_3(K) \right\}$ disrupts the probability of including the true $f(x)$ within our confidence interval as h decreases, i.e., as λ decreases. While attempting to reduce bias by opting for a smaller h, a narrower bandwidth introduces more uncertainty regarding $\hat{f}(x)$. The confidence interval for model 2 at $x = -2$ is not valid for the same reasons encountered for RBC.

The analysis of the Best Model Comparison suggests that there is negligible disparity between RBC and US (with $\lambda = 1$) in terms of their rate of convergence to the 95% line. However, both methods surpass bias correction in all observed scenarios (refer to

figures 6 or 14). When we examine robust bias correction, it becomes apparent that any advantage gained from bias calculation is effectively nullified by the uncertainty associated with such calculations, particularly in contexts with small sample sizes. This observation clarifies why robust bias correction closely resembles undersmoothing with $\lambda = 1$ in terms of coverage probability.

5.2 Analysis of interval length

In the next analysis, the interval length is analyzed in more detail. The first thing to notice here is that the interval length depends strongly on the evaluation point. Evaluation points with a lower true probability $f(x)$ show lower confidence interval lengths for all sample sizes and distributions considered. The effect is also independent of the bandwidth estimators. The observation can be explained theoretically by the fact that the probability $f(x)$ is in the nominator of the (asymptotic) variance of $f(x)$ (see equation 26).

The confidence interval methods are also affected by the choice of bandwidth estimators. For BC and RBC, it can be observed that using $\hat{h}_{RT}$ leads to a smaller confidence interval length for all sample sizes, whereas there is no difference when using $\hat{h}_{PI}$ or $\hat{h}_{CV}$. One possible explanation for the smaller interval length due to $\hat{h}_{RT}$ is that $\hat{h}_{RT}$ estimates comparatively few parameters and can therefore have a lower variance.

Concerning US, the parameter λ influences the interval length. A smaller λ shows a larger confidence interval length for all sample sizes and distributions. The observation corresponds to the theoretical prediction of the Bias-variance trade-off, as a reduction in the bias of $\hat{f}(x)$ through a lower λ should be accompanied by an increase in the variance $\hat{f}(x)$.

By comparing the confidence interval methods, a larger interval length is shown for RBC. The comparison is based on the use of Silverman's rule of thumb estimator $\hat{h}_{RT}$, as the estimator shows the smallest interval length for all methods, and fixing $\lambda = 1$ for US. The result that RBC leads to a larger interval length can theoretically be explained by the fact that $\widehat{\text{Var}}\left(\hat{f}(x) - \hat{\text{B}}(x)\right)$ can be decomposed into $\widehat{\text{Var}}\left(\hat{f}(x)\right) + \widehat{\text{Var}}\left(\hat{B}(x)\right) - 2 * \widehat{\text{Cov}}\left(\hat{f}(x), \hat{B}(x)\right)$, whereby the additional variance of the bias term can dominate. The intuitive explanation here is that the calculation of an additional estimate such as the leading bias term can lead to more uncertainty. Another observation is that the differences in interval length between the confidence interval methods become smaller when an evaluation point with a lower density is considered. Similar to the beginning of the subsection, this can be explained by the fact that the variance of $\hat{f}(x)$ at evaluation points having a low probability is lower. Therefore the differences in the interval lengths are also smaller.

5.3 Relaxing b=h

In the next simulation, we will investigate to what extent the results for RBC change when we undersmooth our estimate for the leading bias term $\hat{\mathrm{B}}(x)$. To reduce the dimension of the analysis, Silverman's rule of thumb bandwidth estimator is used, which has shown the best results in the previous analyses. The bandwidth b is estimated by $\hat{b}_{\mathrm{RT}}(\eta) = 1.06\hat{\sigma}n^{-\frac{1}{5\eta}}$ for $\eta \in \{0.6, 0.7, 0.8, 0.9, 1\}$ such that $\hat{b}_{\mathrm{RT}}(\eta) \leq \hat{h}_{\mathrm{RT}}$.

In the context of employing robust bias correction and undersmoothing our estimate for bias correction, we observe that we achieve valid confidence intervals for all η and models at the latest point for $n = 500$, except for model 2 with $x = -2$ (for model 2 coverage probabilities see figure 9). Furthermore, we observe a slight influence of η on the determination of accurate confidence intervals. It became apparent that lowering η seemed to offer a minor benefit for assessing points linked with a low probability mass. For example, if we take a closer look $x = 2$ or $x = -2$ for Model 1, one can see that the coverage probability of about 0.93 is achieved for $\eta = 0.6$ and $n \approx 200$, whereas this is the case for $\eta = 1$ and $n \approx 500$. Consistent with theoretical expectations, a reduction in the parameter η leads to a less biased estimation of the second derivative, thus promoting a more accurate estimation of the coverage probability.

Analogous to the effect of the US parameter λ (from chapter 5.2), a lower η shows a larger confidence interval length for all sample sizes. This effect holds for all four data-generating processes. This result can be explained by the Bias-variance trade-off, as we observe a higher variance of $\hat{\mathrm{B}}(x)$ if we reduce our bias of $\hat{\mathrm{B}}(x)$ by choosing a lower η. In summary, a higher coverage probability cannot be achieved if the interval length is not increased at the same time [10]. Interestingly, the same trade-off cannot be found when we undersmooth our main bandwidth h, where an increase in λ leads to a longer interval but does not improve the coverage rate.

6 Real data application

6.1 Data Collection

This chapter builds upon the theoretical foundations of kernel density estimation through an application to real data. To this end, we have utilized data from the Consumer Expenditure Public Use Microdata Survey (CE Pumd) (U.S. Bureau of Labor Statistics 2023). Focusing on the univariate case, our analysis centers on the variable "Total amount of family income after estimated taxes in the last 12 months (Imputed or collected data)." The CE Pumd's tri-monthly rotating panel design facilitates the imputation of first-quarter 2023 income for households recorded in the last months of 2022. By restricting our dataset to these most recent observations, we have compiled a dataset consisting of

[10] Therefore, we have opted to continue selecting $\hat{b} = \hat{h}$ as the best choice for robust bias correction in our later real data application.

$n = 4795$ records of household incomes after tax.

6.2 Application of MC-findings

Building on our findings from chapter 5.1 and 5.2, we utilize the models identified as best suited in terms of coverage probability and interval length. Firstly, US with $\lambda = 1$ is used together with the Silverman rule of thumb estimator $\hat{h}_{RT}$ to estimate $f(x)$. In addition, RBC with $\eta = 1$ is applied using the same bandwidth $\hat{h}_{RT}$. Consistently, we employ the Epanechnikov kernel function and set the significance level at $\alpha = 0.05$. Both methods will be evaluating 200 equidistant evaluation points linearly spaced between the minimum and maximum observed family income values.

6.3 Analysis

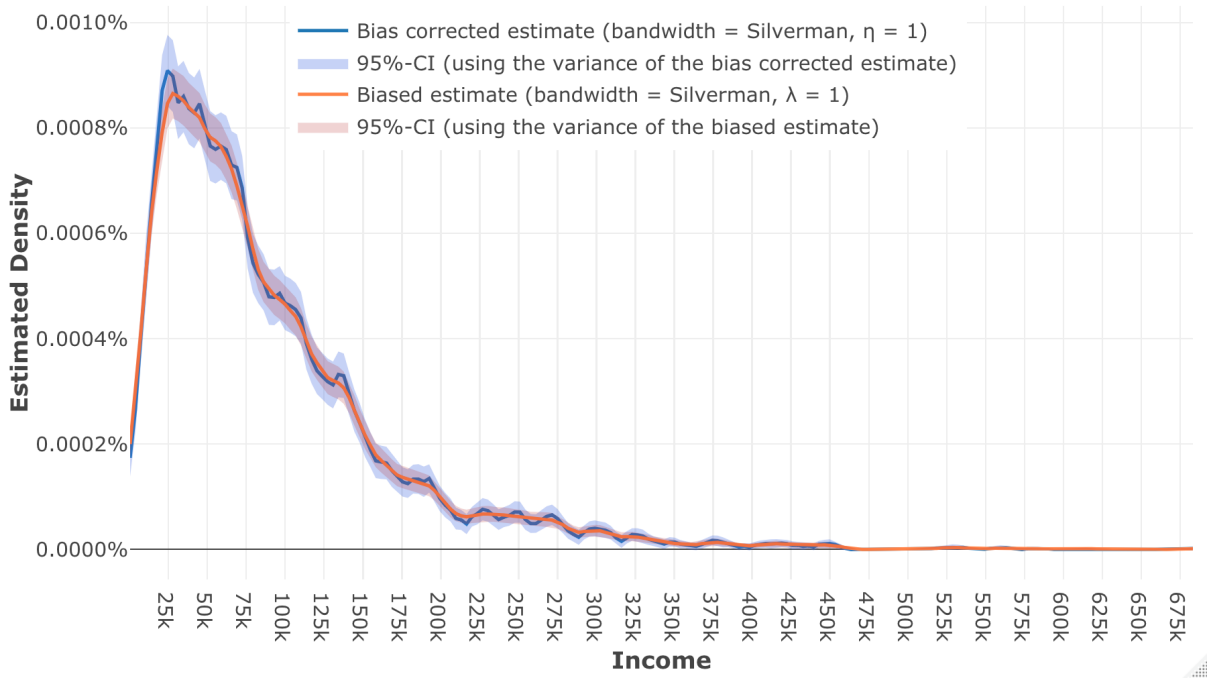

Figure 1: KDE estimation of family income after taxes ($n = 4795$)

The plotted graphs reveal a peak in the distribution of family income after taxes between \$30,000 and \$40,000, indicating a higher concentration at the lower end of the income spectrum. While this general trend is consistent across both models, the bias-corrected estimates exhibit smoother curves and more conservative confidence intervals. This visual examination aligns with the findings from our Monte-Carlo Simulation outlined in chapter 5.2. First, we observe a smoother curve for our RBC method, since we subtract the bias at each $\hat{f}(x)$. Second, the confidence interval lengths fluctuate with evaluation points, showing shorter intervals at lower $f(x)$ values[11]. Last, the implementation of RBC results

[11]Since we do not know $f(x)$, it is implicitly assumed here that $f(x) \approx \hat{f}(x)$.

in wider intervals compared to US with $\lambda = 1$, attributable to the additional estimation of the bias term.

7 Conclusion

Based on the analyzed techniques for constructing confidence intervals, computing kernel density bandwidths, and simulating four distributions, our paper delineates four main observations:

1. Silverman's rule of thumb estimator $\hat{h}_{RT}$ may be a favourable choice for bandwidth selection since the estimator demonstrates faster convergence rates for coverage probability and smaller confidence intervals compared to the alternative bandwidth estimation methods $\hat{h}_{PI}$ and $\hat{h}_{CV}$ (independent of the confidence interval method).

2. RBC and US provide valid confidence intervals for all models (and bandwidth estimators) at the latest point for $n \approx 500$, except for model 2 with $x = -2$. However, this is not the case for BC, where we consistently underestimate the 95% coverage rate.

3. Concerning US, lowering λ yields lower convergence rates of coverage probability and larger confidence interval lengths. Thus, the suggested optimal choice would be $\lambda = 1$.

4. The optimal choice among the confidence interval construction techniques emerges as US with $\lambda = 1$. This preference is substantiated by the observation that, while upholding a comparable coverage probability (as RBC), US with $\lambda = 1$ results in shorter interval lengths compared to employing RBC (using $\hat{h}_{RT}$).

These findings collectively indicate that both the foundational principle of the undersmoothing method, which advocates for the selection of a smaller bandwidth to mitigate bias and the explicit consideration of bias calculation, do not confer advantages over employing a simple test statistic utilizing the optimal bandwidth $\hat{h}_{RT}$ and disregarding bias.

The following limitations merit consideration for future investigations. Only distributions without boundary points were examined (see assumption 5). Additionally, only the Epanechnikov kernel was utilized, chosen based on our (compact kernel) assumption 6. Furthermore, an IMSE- or ISE-optimal bandwidth estimator was employed rather than identifying a pointwise best MSE-optimal bandwidth. The last point should also emphasize that our recommendation using undersmoothing with $\lambda = 1$ might not be optimal when one considers a pointwise MSE-optimal bandwidth estimator.

References

Bowman, A. (1984). "An alternative method of cross-validation for the smoothing of density estimates". In: *Biometrika* 71, pp. 353–360.

Calonico, Sebastian, Matias Cattaneo, and Max Farrell (2018). "On the Effect of Bias Estimation on Coverage Accuracy in Nonparametric Inference". In: *Journal of the American Statistical Association* 113.522, pp. 767–779.

Hall, P. (1992). "Effect of Bias Estimation on Coverage Accuracy of Bootstrap Confidence Intervals for a Probability Density". In: *The Annals of Statistics* 20.2, pp. 675–694.

Hall, P. and J.S. Marron (1987). "Estimation of Integrated Squared Density Derivatives". In: *Statistics & Probability Letters* 6, pp. 109–115.

Härdle, W., P. Hall, and J.S. Marron (1988). "How Far are Automatically Chosen Regression Smoothing Parameters from their Optimum?" In: *Journal of The American Statistical Association* 83, pp. 86–101.

Li, Q. and J. S. Racine (2006). *Nonparametric Econometrics: Theory and Practice*. Princeton and Oxford: Princeton University Press.

Park, U. and J.S. Marron (1990). "Comparison of Data-Driven Bandwidth Selectors". In: *Journal of the American Statistical Association* 85, pp. 66–72.

Rudemo, M. (1982). "Empirical choice of histograms and kernel density estimators". In: *Scandinavian Journal of Statistics* 9, pp. 65–78.

Sheather, S.J. and M.C. Jones (1991a). "A Reliable Data-Based Bandwidth Selection Method for Kernel Density Estimaton". In: *Journal of the Royal Statistical Society* 53. 1991B, pp. 683–690.

— (1991b). "Using non-stochastic terms to advantage in kernel-based estimation of integrated squared density derivatives". In: *Statistics & Probability Letters* 11. 1991A, pp. 511–514.

Silverman, B.W. (1986). *Density Estimation for Statistics and Data Analysis*. New York: Chapman and Hall.

Stone, C. J. (1984). "An asymptotically optimal window selection rule for kernel density estimates". In: *Annals of Statistics* 12, pp. 1285–1297.

U.S. Bureau of Labor Statistics (2023). *Consumer Expenditure Public Use Microdata*. Accessed on: 2024-01-05. URL: `https://www.bls.gov/cex/pumd_data.htm`.

A Appendix

A.1 Theory related appendix

Proof 1. Bias derivation (equation (1)).

$$
\begin{aligned}
\text{Bias}\left(\hat{f}(x)\right) &= E\left[\frac{1}{nh}\sum_{i=1}^{n} k\left(\frac{X_i - x}{h}\right)\right] - f(x) \\
&= h^{-1}E\left[k\left(\frac{X_1 - x}{h}\right)\right] - f(x) \\
&\text{(by identical distribution of assumption 2)} \\
&= h^{-1}\int f(x_1)k\left(\frac{x_1 - x}{h}\right)\,dx_1 - f(x) \\
&= h^{-1}\int f(x + hv)k(v)h\,dv - f(x) \\
&\text{(define } v = (x_1 - x)/h\text{)} \\
&= \int\left[f(x) + f^{(1)}(x)hv + \frac{1}{2}f^{(2)}(x)h^2v^2 + \frac{1}{6}f^{(3)}(\tilde{x})h^3v^3\right]k(v)\,dv - f(x) \\
&\text{(by Taylor approximation at } x\text{, assumption 3 and where } \tilde{x} \in [x, x + hv]\text{)} \\
&= f(x) + 0 + \frac{h^2}{2}f^{(2)}(x)\int v^2k(v)\,dv + \frac{h^3}{6}f^{(3)}(\tilde{x})\int v^3k(v)\,dv - f(x) \\
&\text{(by assumption 1)} \\
&= \frac{h^2}{2}f^{(2)}(x)\int v^2k(v)\,dv + O(h^3) \qquad (25)
\end{aligned}
$$

□

Proof 2. Variance derivation (equation (2)).

$$
\begin{aligned}
\operatorname{Var}\left(\hat{f}(x)\right) &= \operatorname{Var}\left[\frac{1}{nh}\sum_{i=1}^{n} k\left(\frac{X_i - x}{h}\right)\right] \\
&= \frac{1}{n^2h^2}\left\{\sum_{i=1}^{n} \operatorname{Var}\left[k\left(\frac{X_i - x}{h}\right)\right] + 0\right\} \\
&\text{(by independence of assumption 2)} \\
&= \frac{1}{nh^2}\operatorname{Var}\left[k\left(\frac{X_1 - x}{h}\right)\right] \\
&\text{(by indentical distribution of assumption 2)} \\
&= \frac{1}{nh^2}\left\{\mathrm{E}\left[k^2\left(\frac{X_1 - x}{h}\right)\right] - \mathrm{E}\left[k\left(\frac{X_1 - x}{h}\right)\right]^2\right\} \\
&= \frac{1}{nh^2}\left\{\int f(x_1)k^2\left(\frac{x_1 - x}{h}\right)\,dx_1 - \left[\int f(x_1)k\left(\frac{x_1 - x}{h}\right)\,dx_1\right]^2\right\} \\
&= \frac{1}{nh^2}\left\{h\int f(x + hv)k^2(v)\,dv - \left[h\int f(x + hv)k(v)\,dv\right]^2\right\} \\
&\text{(using } v = (x_1 - x)/h\text{)} \\
&= \frac{1}{nh^2}\left\{\int f(x_1)k^2\left(\frac{x_1 - x}{h}\right)\,dx_1 - \left[\int f(x_1)k\left(\frac{x_1 - x}{h}\right)\,dx_1\right]^2\right\} \\
&= \frac{1}{nh^2}\left\{h\int \left[f(x) + f^{(1)}(\hat{x})hv\right]k^2(v)\,dv - O(h^2)\right\} \\
&\text{(by Taylor approximation at } x\text{, assumption 3 and where } \hat{x} \in [x, x + hv]\text{)} \\
&= \frac{1}{nh}\left\{f(x)\int k^2(v)\,dv + O\left(h\int |v|k^2(v)\,dv\right) - O(h)\right\} \\
&= \frac{1}{nh}\left\{f(x)\int k^2(v)\,dv + O(h)\right\} \qquad (26)
\end{aligned}
$$

□

Proof 3. Multivariate bias (equation (10)).

$$\begin{aligned}
\text{Bias}\left(\hat{f}(x)\right) &= E\left(\hat{f}(x)\right) - f(x) \\
&= E\left[\frac{1}{nh_1\cdots h_q} k\left(\frac{X_{i1}-x_1}{h_1}\right) \text{x}\cdots\text{x}\; k\left(\frac{X_{iq}-x_q}{h_q}\right)\right] - f(x) \\
&= \int \frac{1}{nh_1\cdots h_q} k\left(\frac{x_{i1}-x_1}{h_1}\right) \text{x}\cdots\text{x}\; k\left(\frac{x_{iq}-x_q}{h_q}\right) f(x_i)\; dx_i - f(x) \\
&= \int f\left(x+hv\right) k(v)\; dv - f(x) \\
&\text{(Since we have that} f\left(x+hv\right) = \\
&f(x) + \sum_{s=1}^{q} f^{(s)}(x) h_s v_s + \frac{1}{2}\sum_{s=1}^{q}\sum_{t=1}^{q} f^{(st)}(x) h_s h_t v_s v_t + O\left(\sum_{s=1}^{q} h_s^3\right) \text{ we get)} \\
&= \int \left[f(x) + \sum_{s=1}^{q} f^{(s)}(x) h_s v_s + \frac{1}{2}\sum_{s=1}^{q}\sum_{t=1}^{q} f^{(st)}(x) h_s h_t v_s v_t\right] k(v) dv \\
&- f(x) + O\left(\sum_{s=1}^{q} h_s^3\right) \\
&= \frac{\int v^2 k(v) dv}{2} \sum_{s=1}^{q} h_s^2 f^{(ss)}(x) + O\left(\sum_{s=1}^{q} h_s^3\right)
\end{aligned}$$

□

Proof 4. Multivariate variance (equation (11)).

$$\begin{aligned}
\text{Var}\left(\hat{f}(x)\right) &= \text{Var}\left[\frac{1}{nh_1\cdots h_q}\sum_{k=1}^{n} k\left(\frac{X_{i1}-x_1}{h_1}\right) \text{x}\cdots\text{x}\; k\left(\frac{X_{iq}-x_q}{h_q}\right)\right] \\
&= \frac{1}{nh_1^2\cdots h_q^2}\text{Var}\left[\sum_{k=1}^{n} k\left(\frac{X_{i1}-x_1}{h_1}\right) \text{x}\cdots\text{x}\; k\left(\frac{X_{iq}-x_q}{h_q}\right)\right] \\
&= \frac{1}{nh_1^2\cdots h_q^2} E\left[k\left(\frac{X_{i1}-x_1}{h_1}\right) \text{x}\cdots\text{x}\; k\left(\frac{X_{iq}-x_q}{h_q}\right)\right]^2 \\
&- \frac{1}{nh_1^2\cdots h_q^2}\left(E\left[k\left(\frac{X_{i1}-x_1}{h_1}\right) \text{x}\cdots\text{x}\; k\left(\frac{X_{iq}-x_q}{h_q}\right)\right]\right)^2 \\
&= \frac{1}{nh_1^2\cdots h_q^2}\int \left[k\left(\frac{x_{i1}-x_1}{h_1}\right) \text{x}\cdots\text{x}\; k\left(\frac{x_{iq}-x_q}{h_q}\right)\right]^2 f(x_i) dx_i + o\left(\frac{1}{nh_1\cdots h_q}\right) \\
&= \frac{1}{nh_1\cdots h_q}\left[\left(\int k^2(v) dv\right)^q f(x) + O\left(\sum_{s=1}^{q} h_s^2\right)\right]
\end{aligned}$$

□

Definition 3. *Description of the terms in equations (22), (23), and (24).*

$$
\begin{aligned}
q_1(k) &= \left(\int k(v)^2 dv\right)^{-2} \left(\int k(v)^4 dv\right) \left(z_{\frac{\alpha}{2}}^3 - 3z_{\frac{\alpha}{2}}\right)/6 - \left(\int k(v)^2 dv\right)^{-3} \left(\int k(v)^3 dv\right)^2 \\
&\quad \times \left[2z^3/3 + \left(z_{\frac{\alpha}{2}}^5 - 10z_{\frac{\alpha}{2}}^3 + 15z_{\frac{\alpha}{2}}\right)/9\right] \\
q_2(k) &= -\left(\int k(v)^2 dv\right)^{-1} z_{\frac{\alpha}{2}} \\
q_3(k) &= \left(\int k(v)^2 dv\right)^{-2} \left(\int k(v)^3 dv\right) \left(2z_{\frac{\alpha}{2}}^3/3\right) \\
\Omega_1 &= -2\frac{1/2\left(\int v^2 k(v) dv\right)}{1/hE\left[\left(k(X_{h,i}) - E\left(k(X_{h,i})\right)^2\right]} \\
&\quad \times \left\{\int f(x - vh)k(v)L^2(v\rho)dv - b\int f(x - vh)k(v)dv \int f(x - vb)L^2(v)dv\right\} \\
\Omega_2 &= 1/2\left(\int v^2 k(v) dv\right)^2 \left(\int k(v)^2 dv\right)^{-2} \left(\int L(v)^2 dv\right)
\end{aligned}
$$

A.2 Monte Carlo Simulation Results

Here, we present the outcomes of our Monte Carlo Simulation:

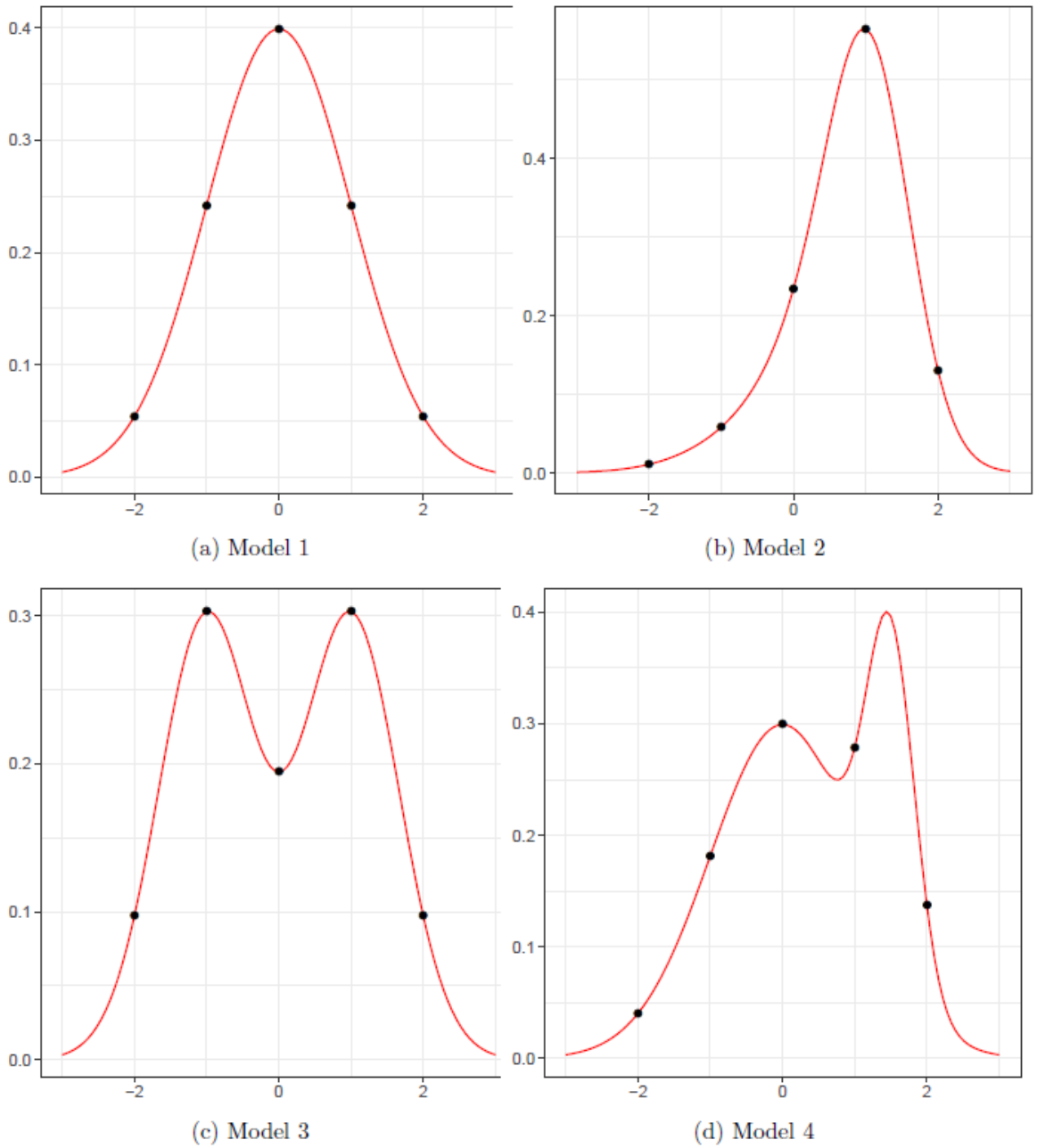

Figure 2: Data generating density models with evaluating points at -2, -1, 0, 1, and 2.

A.2.1 Figures related to Coverage Probability

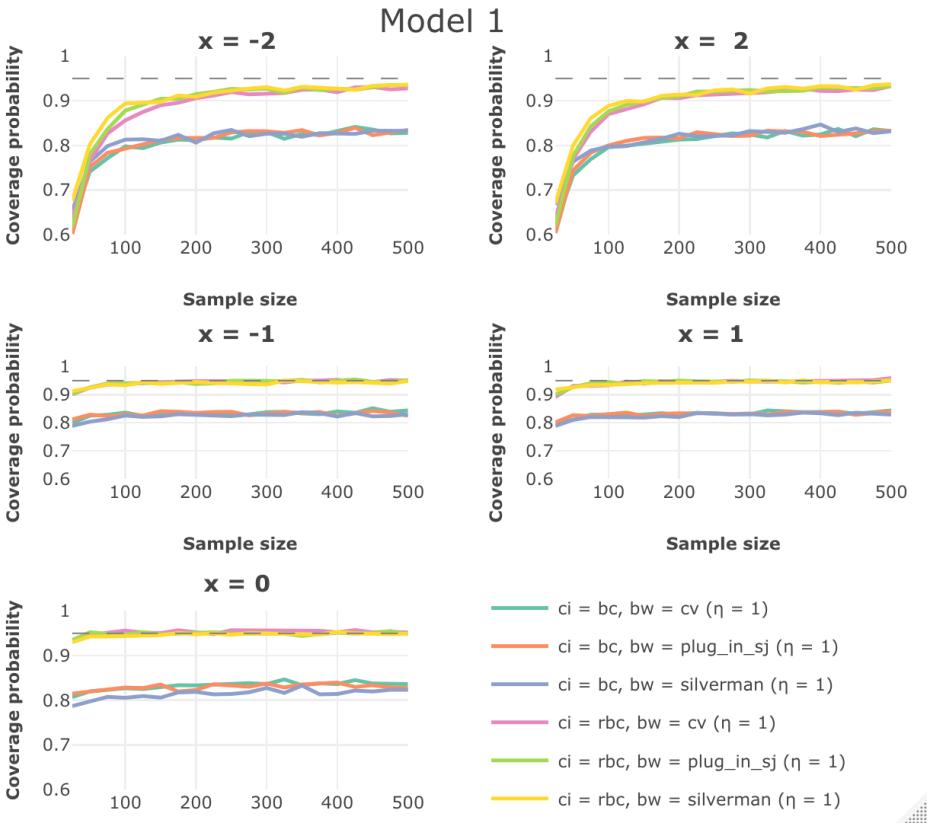

Figure 3: Coverage probability for RBC and BC (Model 1).

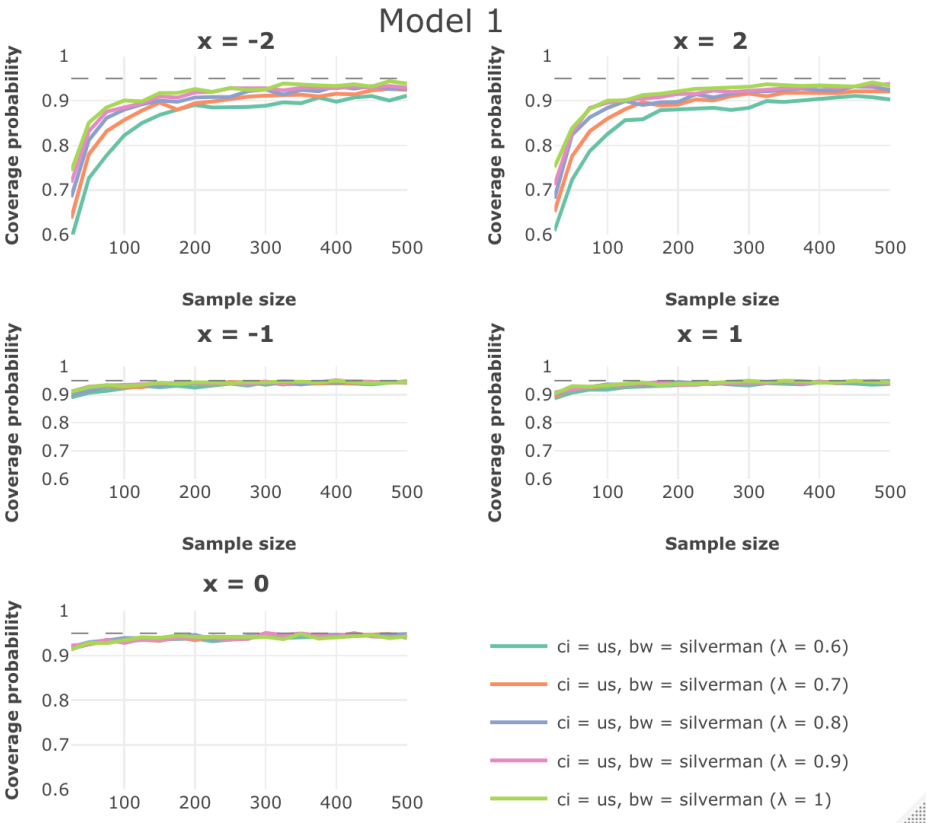

Figure 4: Coverage probability for US using $\hat{h}_{\mathrm{RT}}(\lambda)$ (Model 1).

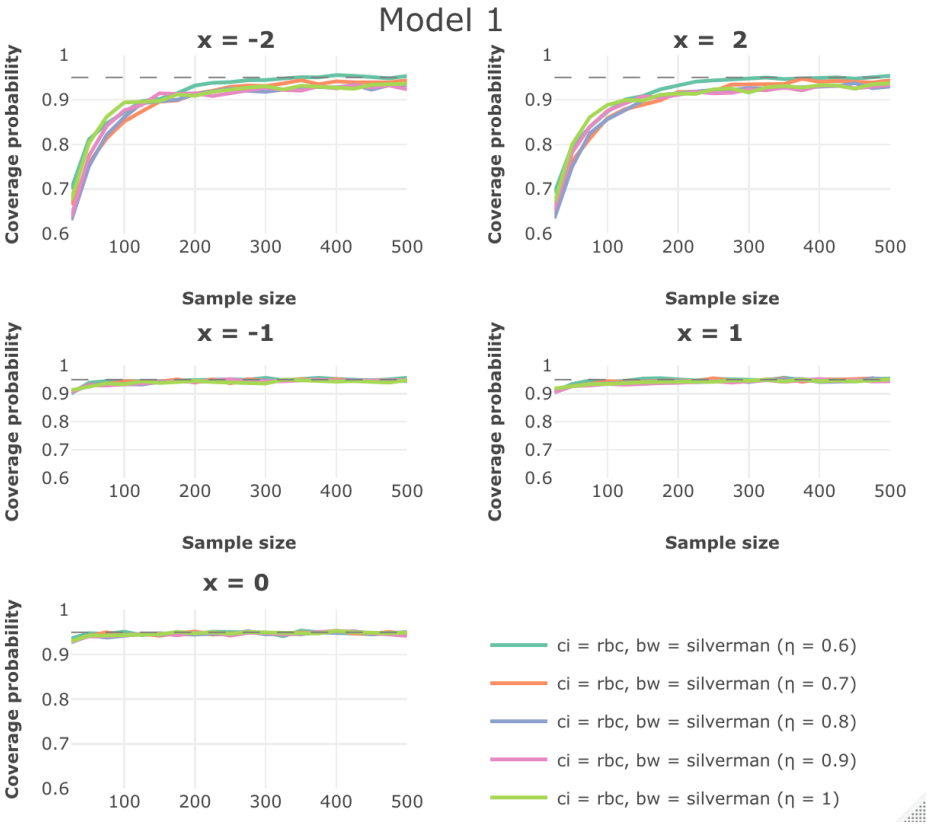

Figure 5: Coverage probability for RBC using $\hat{b}_{\mathrm{RT}}(\eta)$ (Model 1).

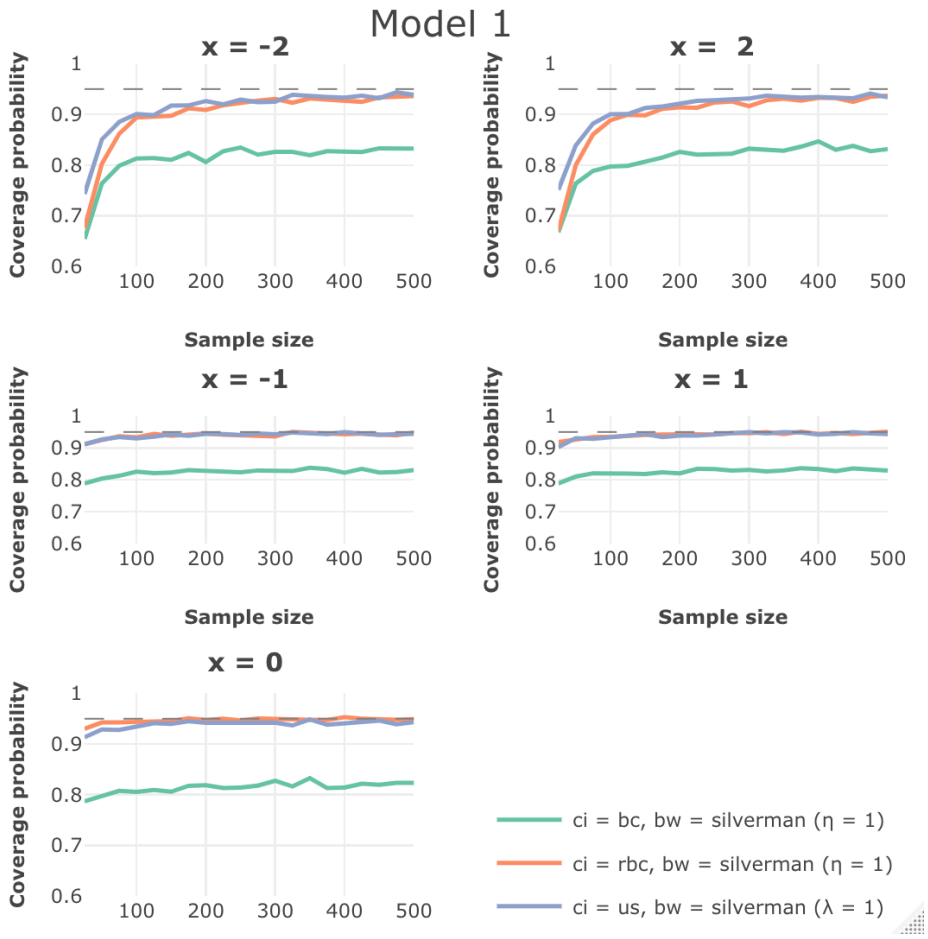

Figure 6: Coverage probability for best combinations (Model 1).

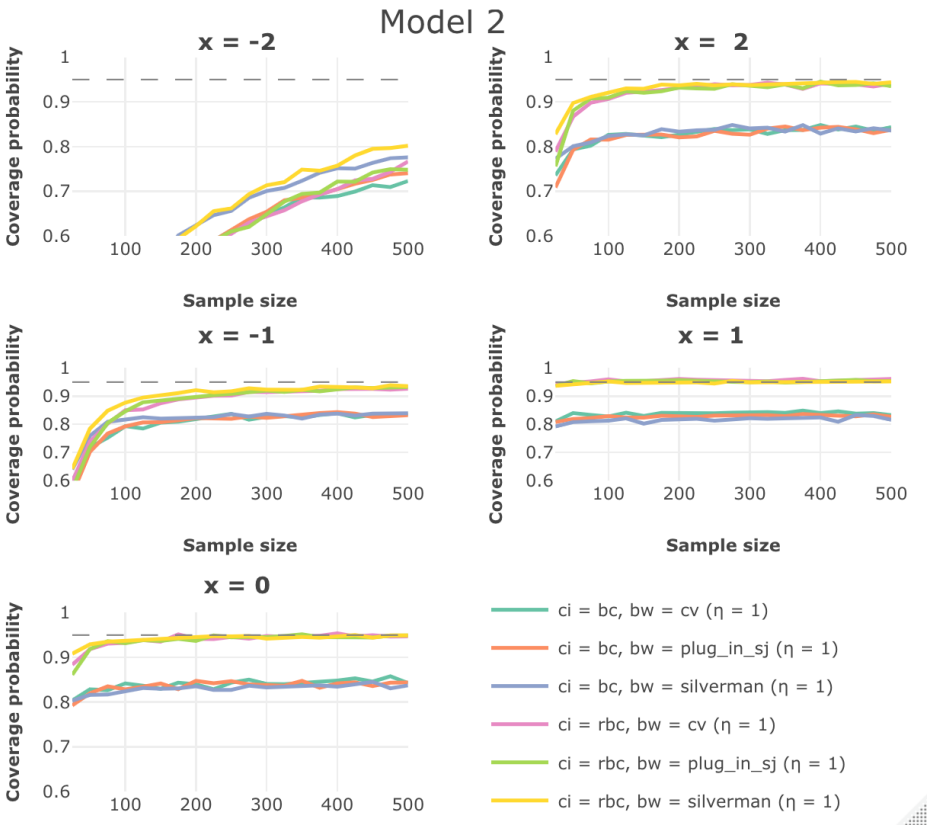

Figure 7: Coverage probability for RBC and BC (Model 2).

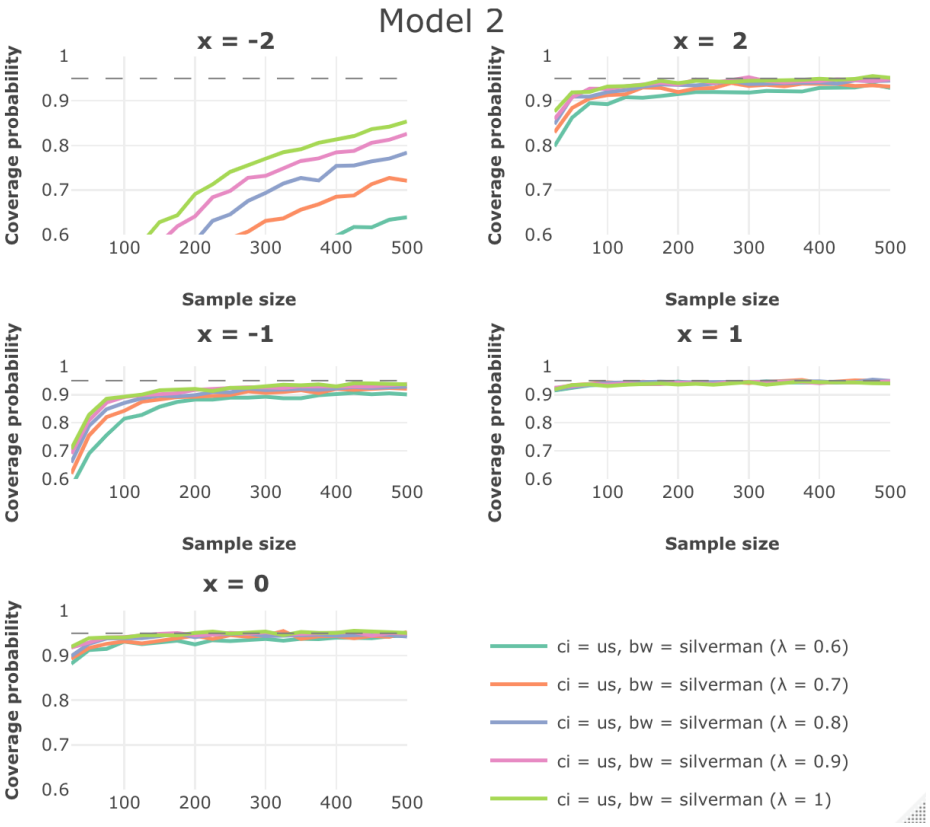

Figure 8: Coverage probability for US using $\hat{h}_{\mathrm{RT}}(\lambda)$ (Model 2).

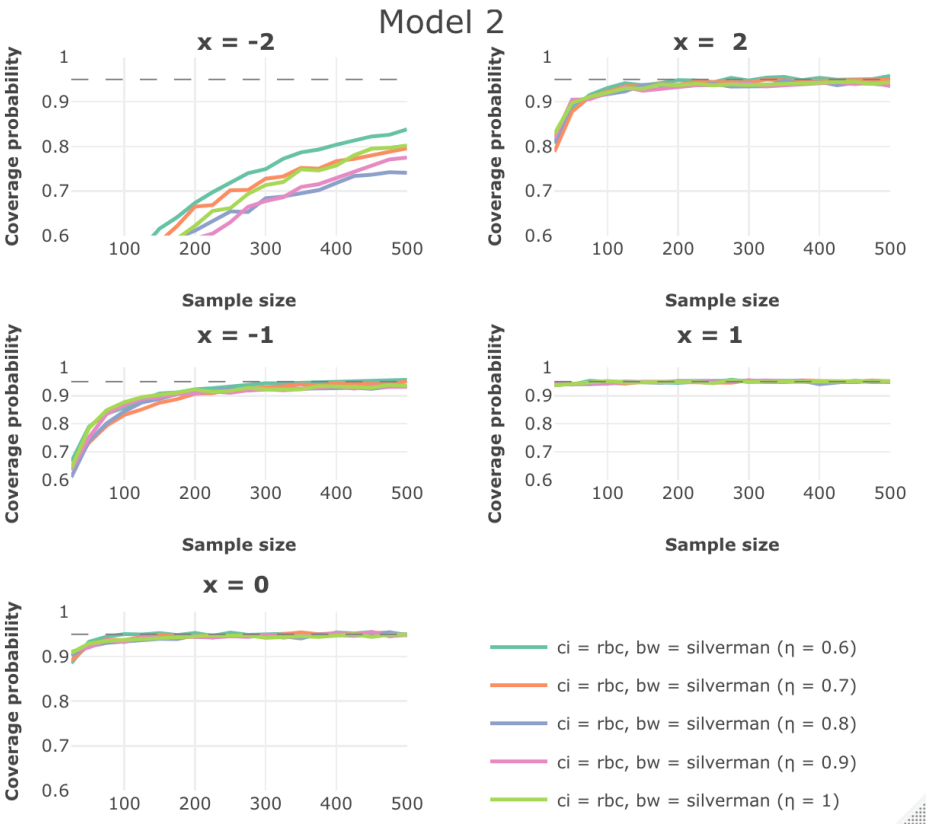

Figure 9: Coverage probability for RBC using $\hat{b}_{\mathrm{RT}}(\eta)$ (Model 2).

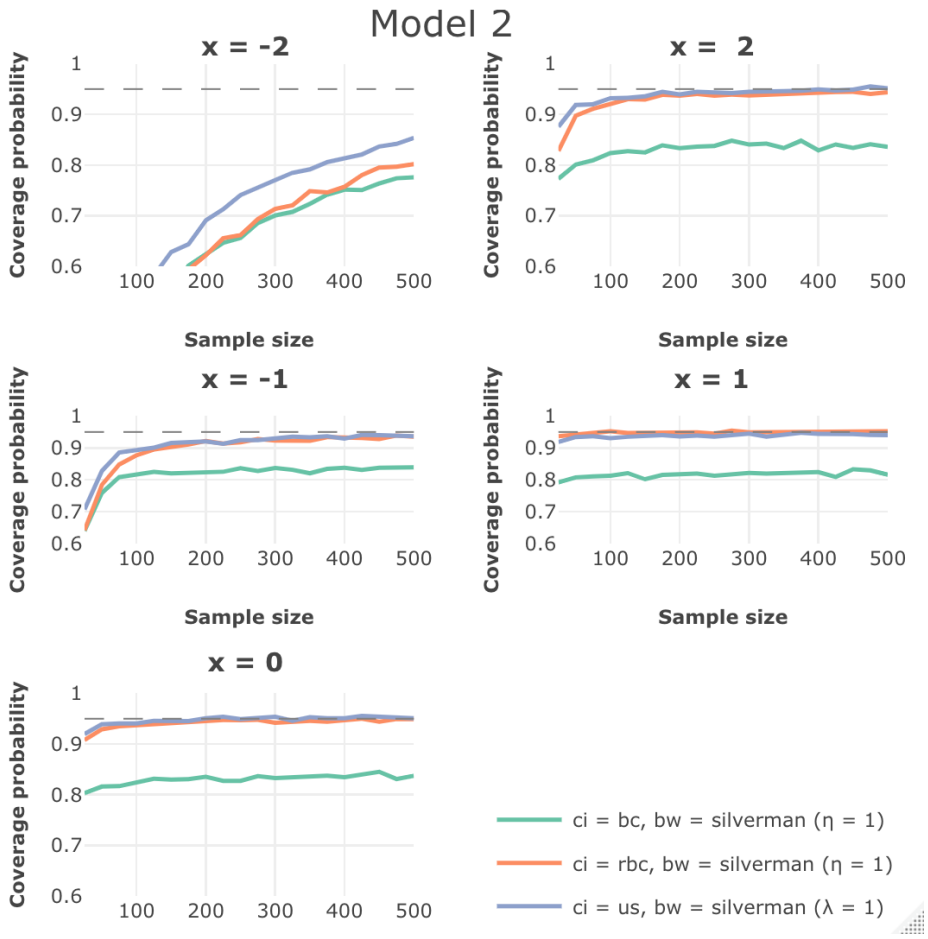

Figure 10: Coverage probability for best combinations (Model 2).

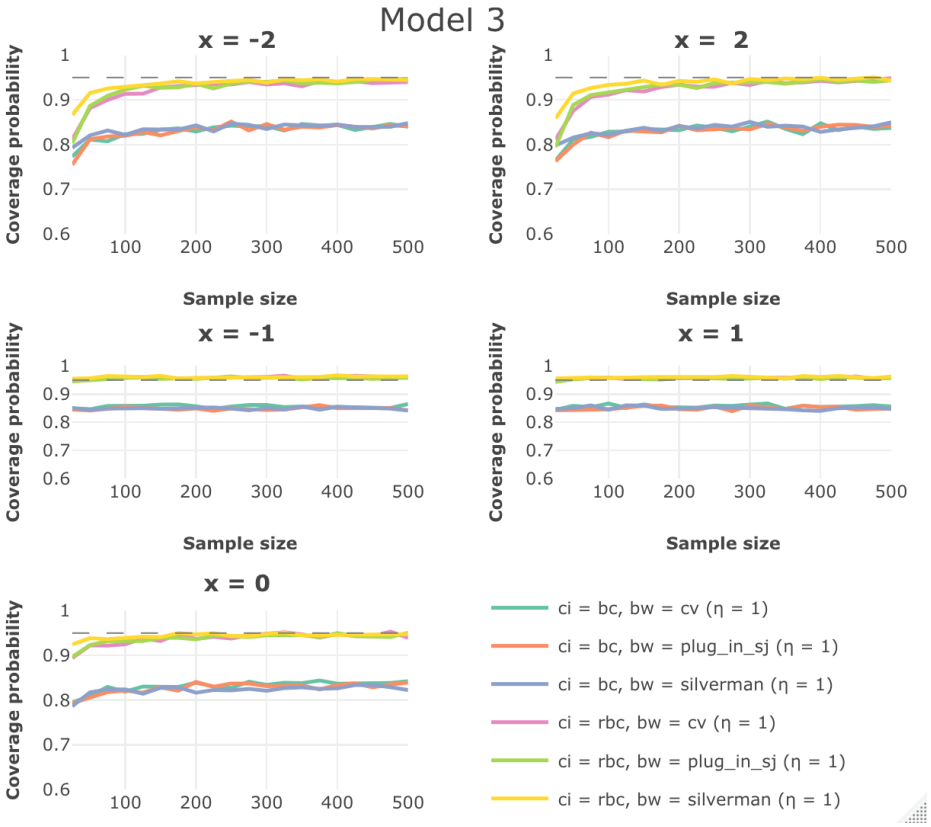

Figure 11: Coverage probability for RBC and BC (Model 3).

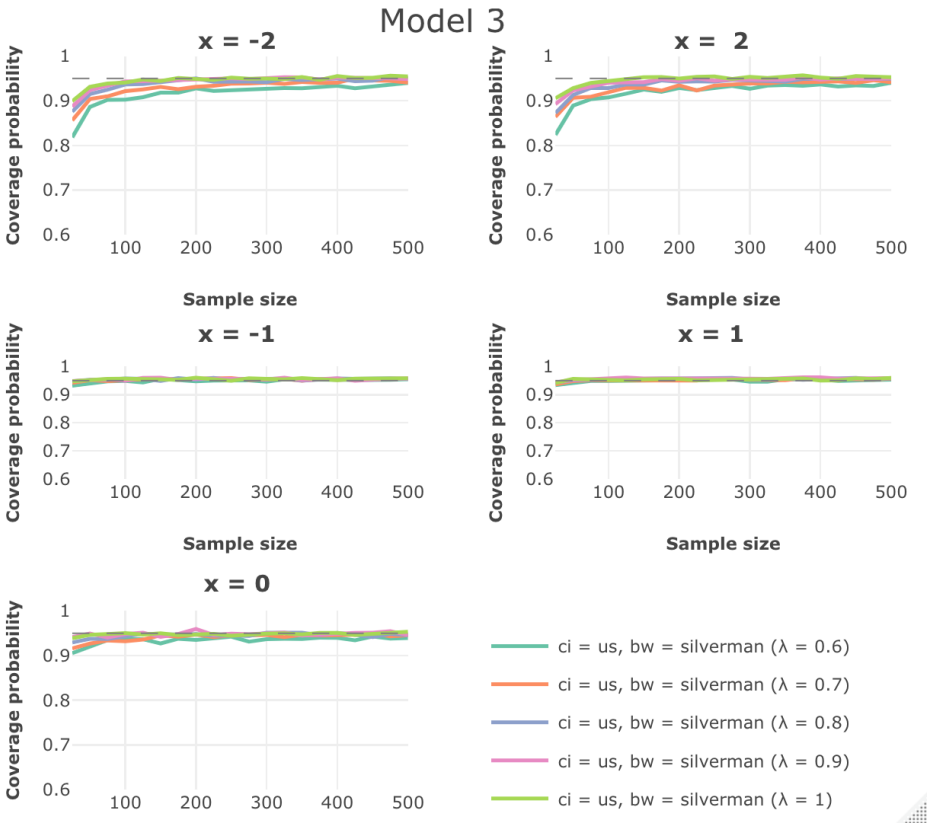

Figure 12: Coverage probability for US using $\hat{h}_{\mathrm{RT}}(\lambda)$ (Model 3).

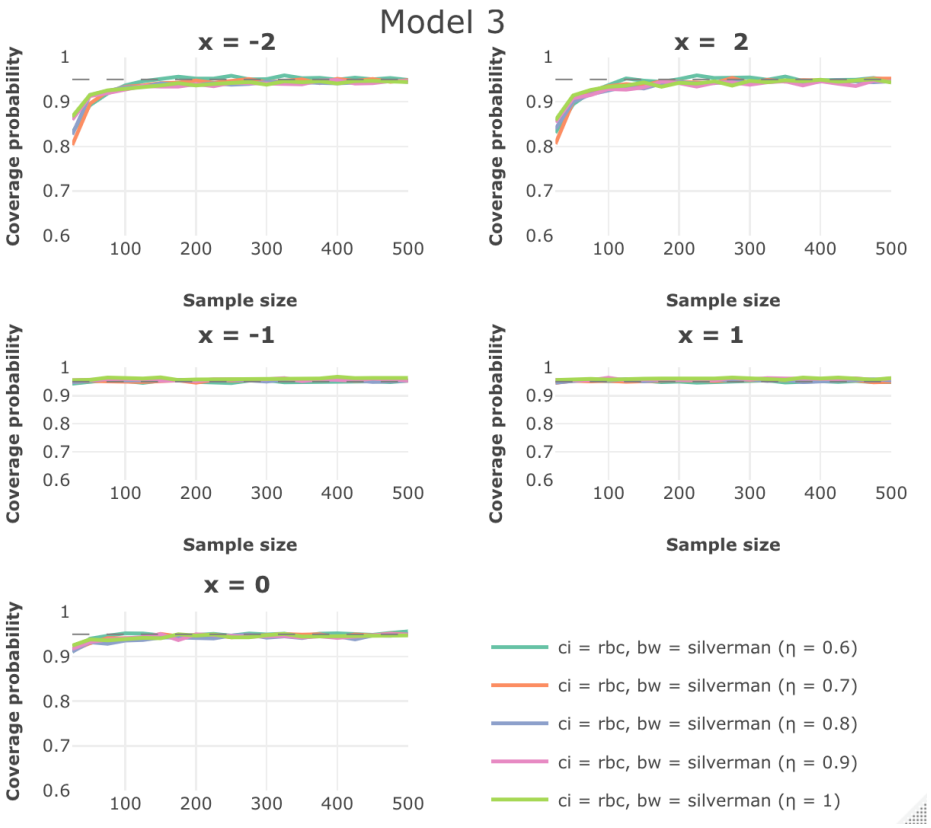

Figure 13: Coverage probability for RBC using $\hat{b}_{\mathrm{RT}}(\eta)$ (Model 3).

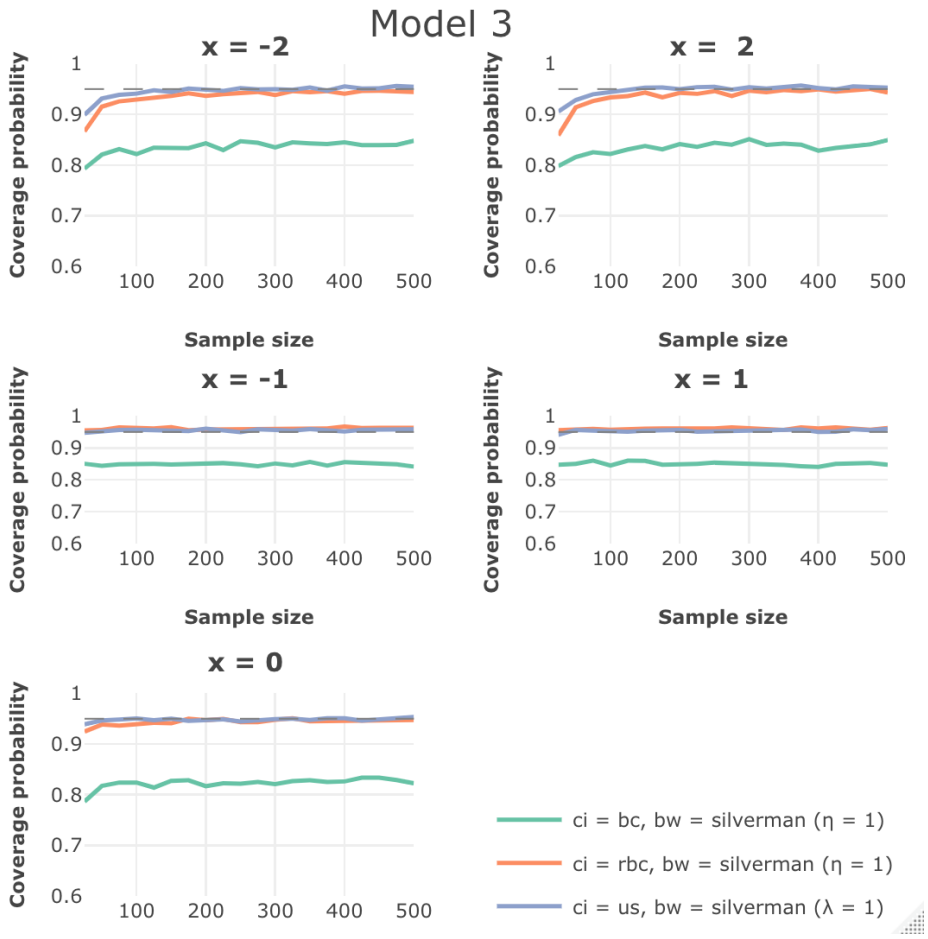

Figure 14: Coverage probability for best combinations (Model 3).

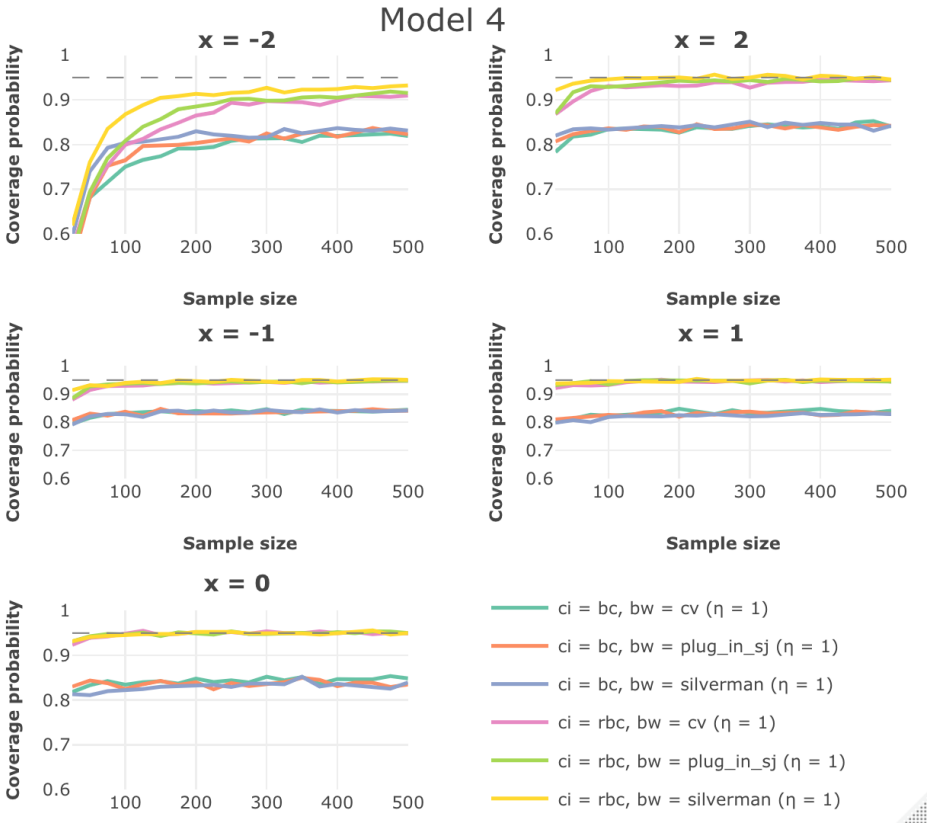

Figure 15: Coverage probability for RBC and BC (Model 4).

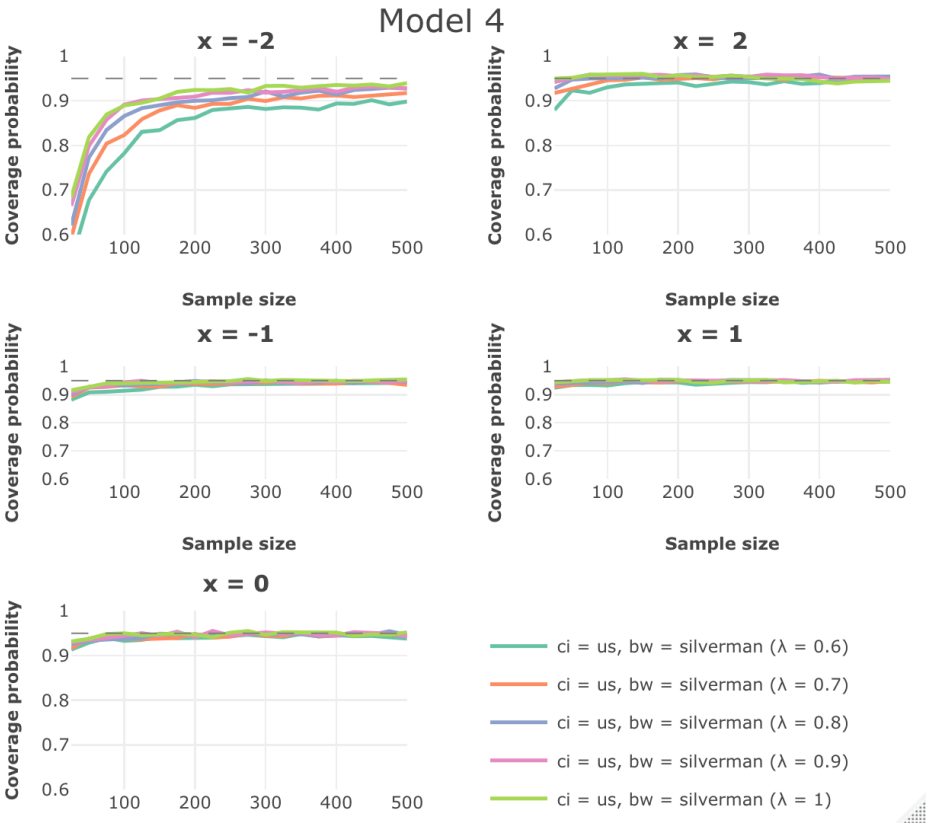

Figure 16: Coverage probability for US using $\hat{h}_{\mathrm{RT}}(\lambda)$ (Model 4).

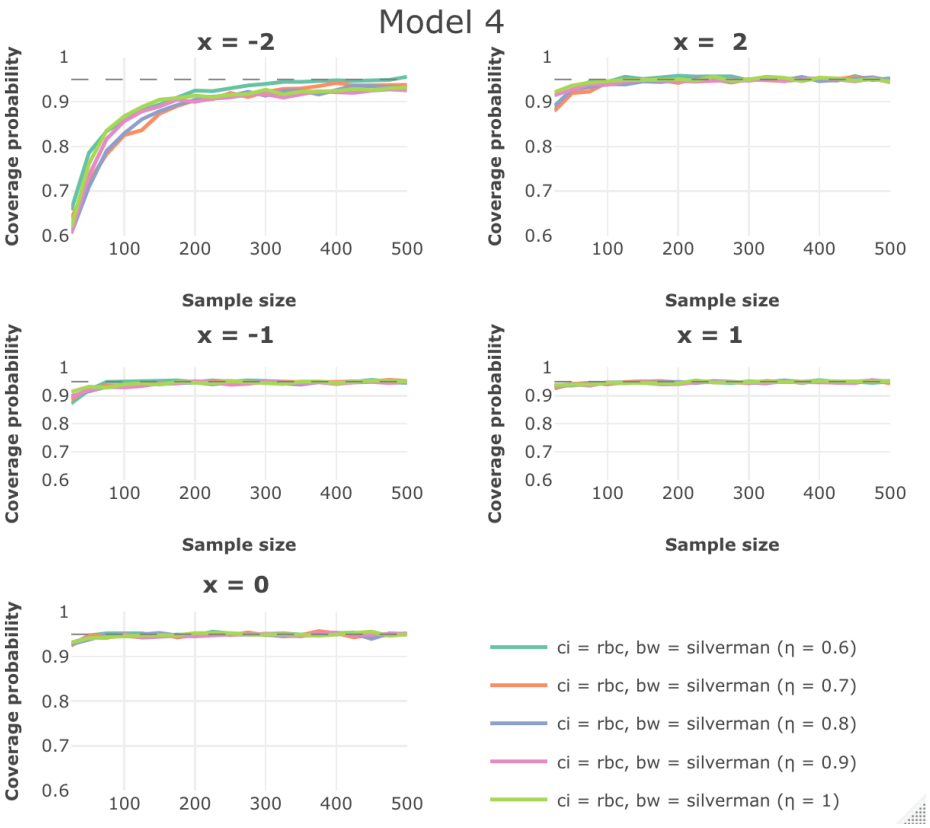

Figure 17: Coverage probability for RBC using $\hat{b}_{\mathrm{RT}}(\eta)$ (Model 4).

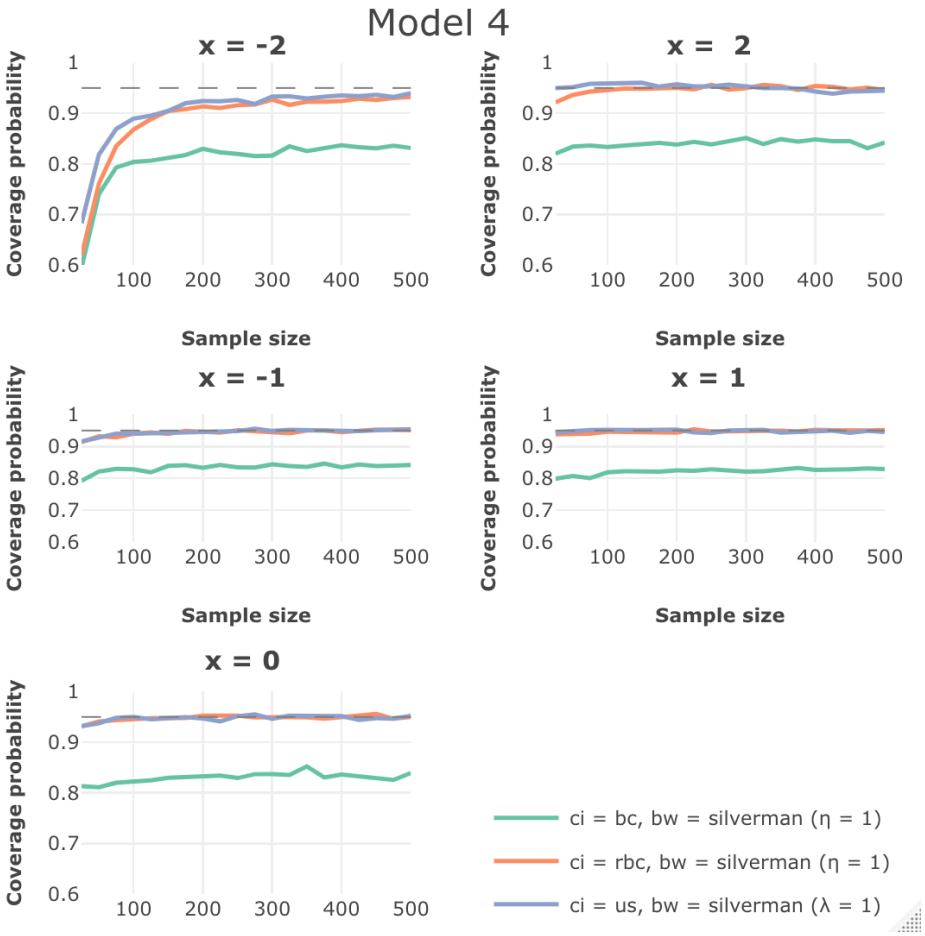

Figure 18: Coverage probability for best combinations (Model 4).

A.2.2 Figures related to Interval Lengths

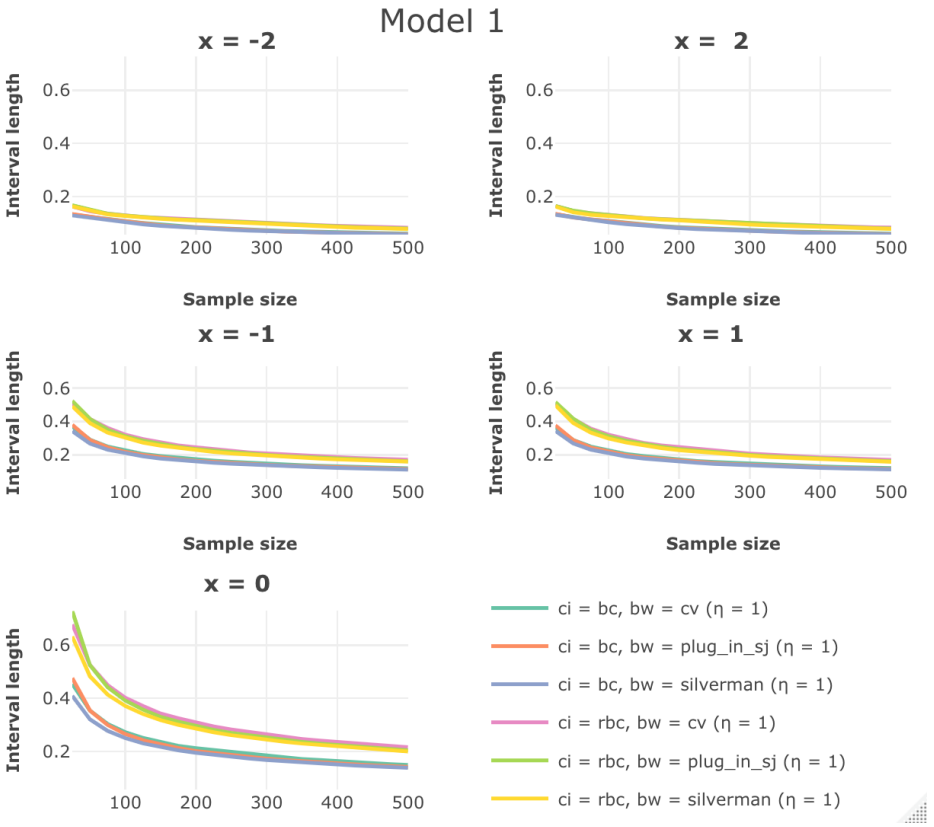

Figure 19: Interval length for RBC and BC (Model 1).

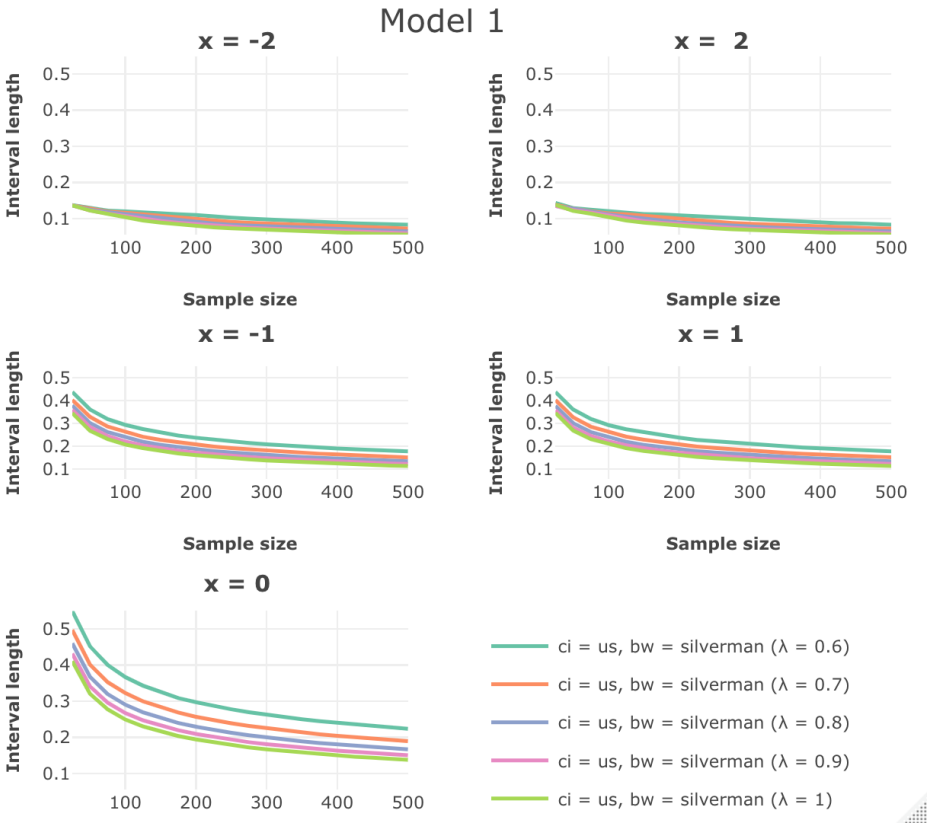

Figure 20: Interval length for US using $\hat{h}_{\mathrm{RT}}(\lambda)$ (Model 1).

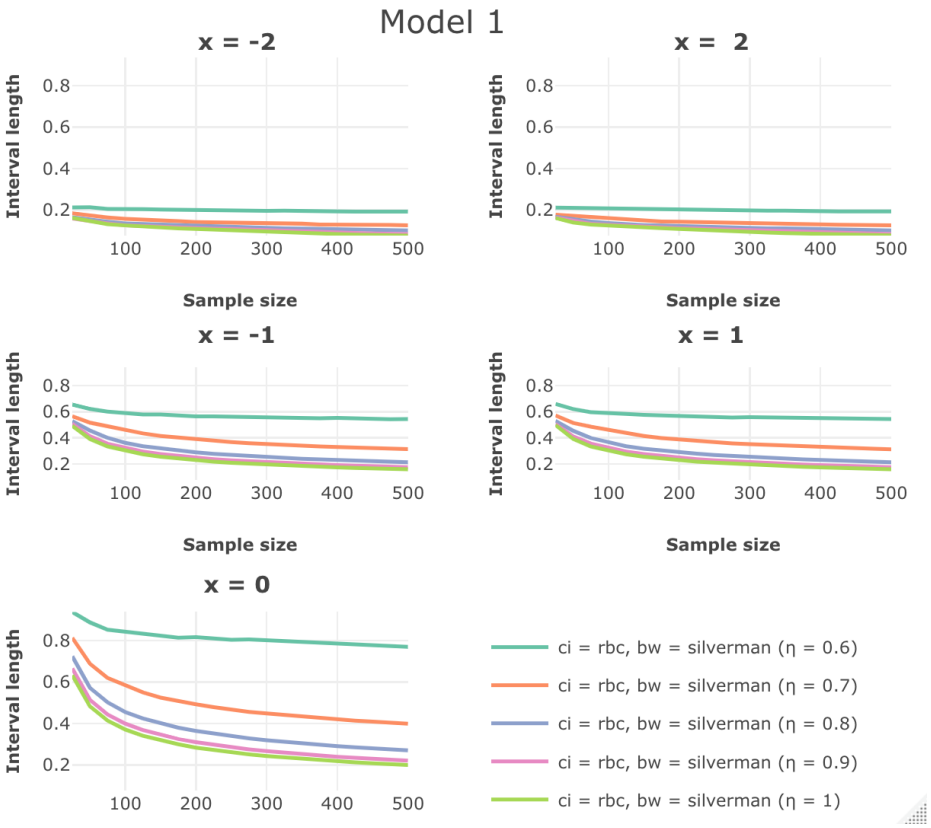

Figure 21: Interval length for RBC using $\hat{b}_{\mathrm{RT}}(\eta)$ (Model 1).

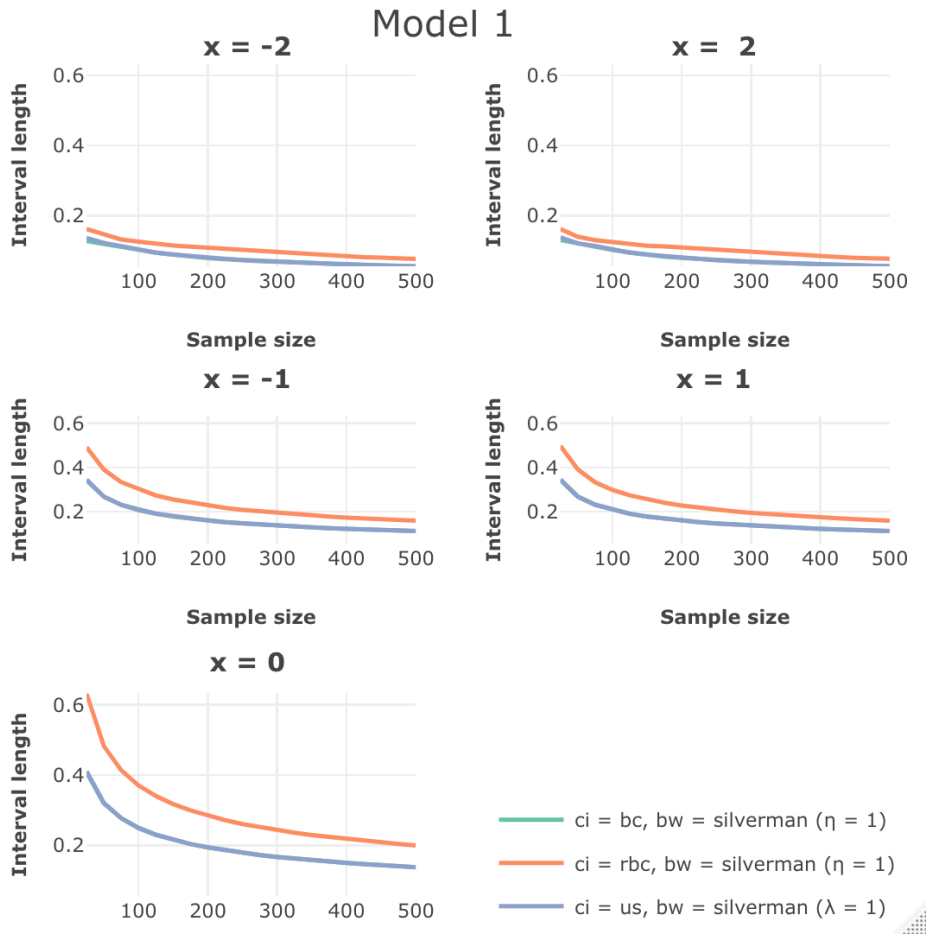

Figure 22: Interval length for best combinations (Model 1).

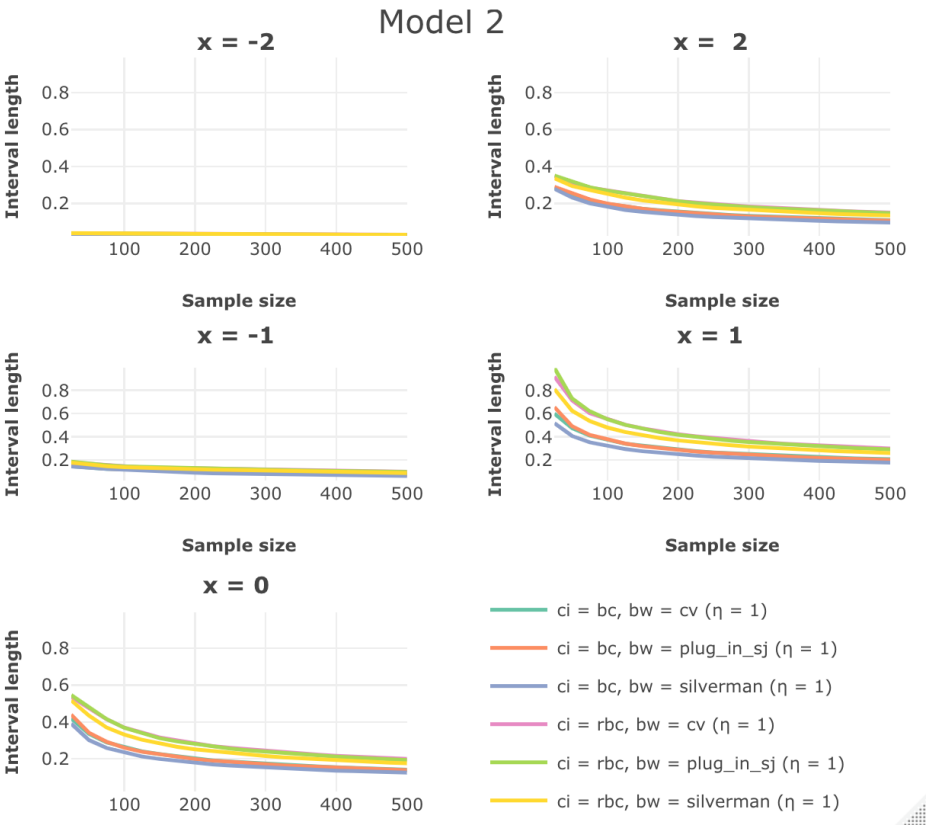

Figure 23: Interval length for RBC and BC (Model 2).

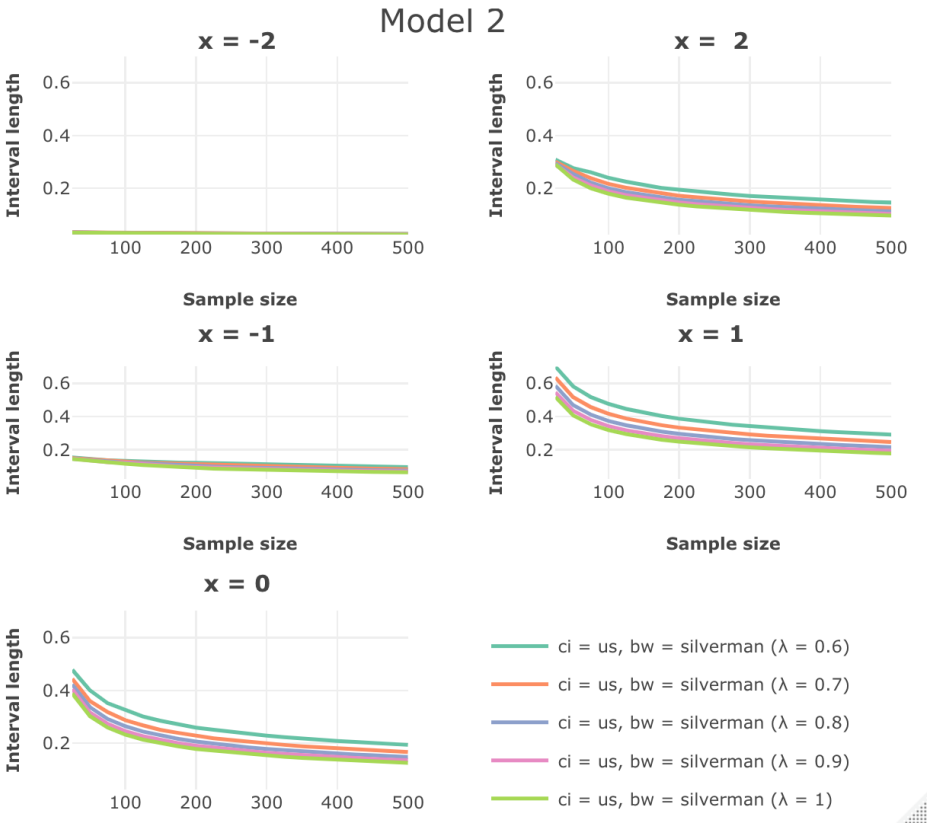

Figure 24: Interval length for US using $\hat{h}_{\mathrm{RT}}(\lambda)$ (Model 2).

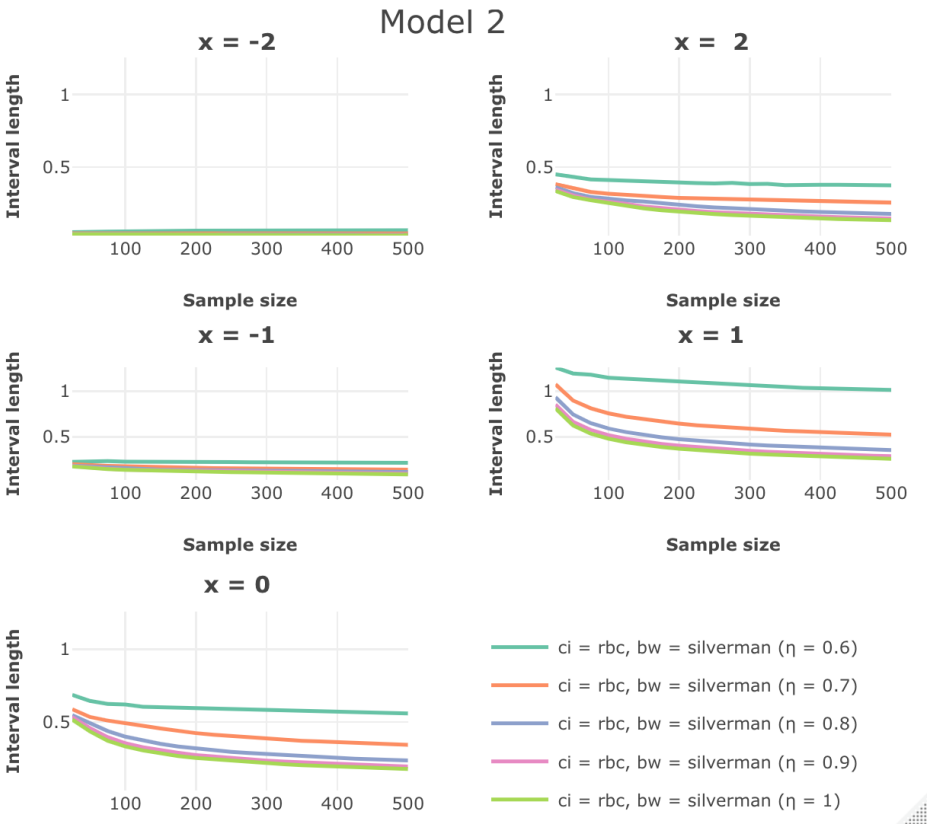

Figure 25: Interval length for RBC using $\hat{b}_{\mathrm{RT}}(\eta)$ (Model 2).

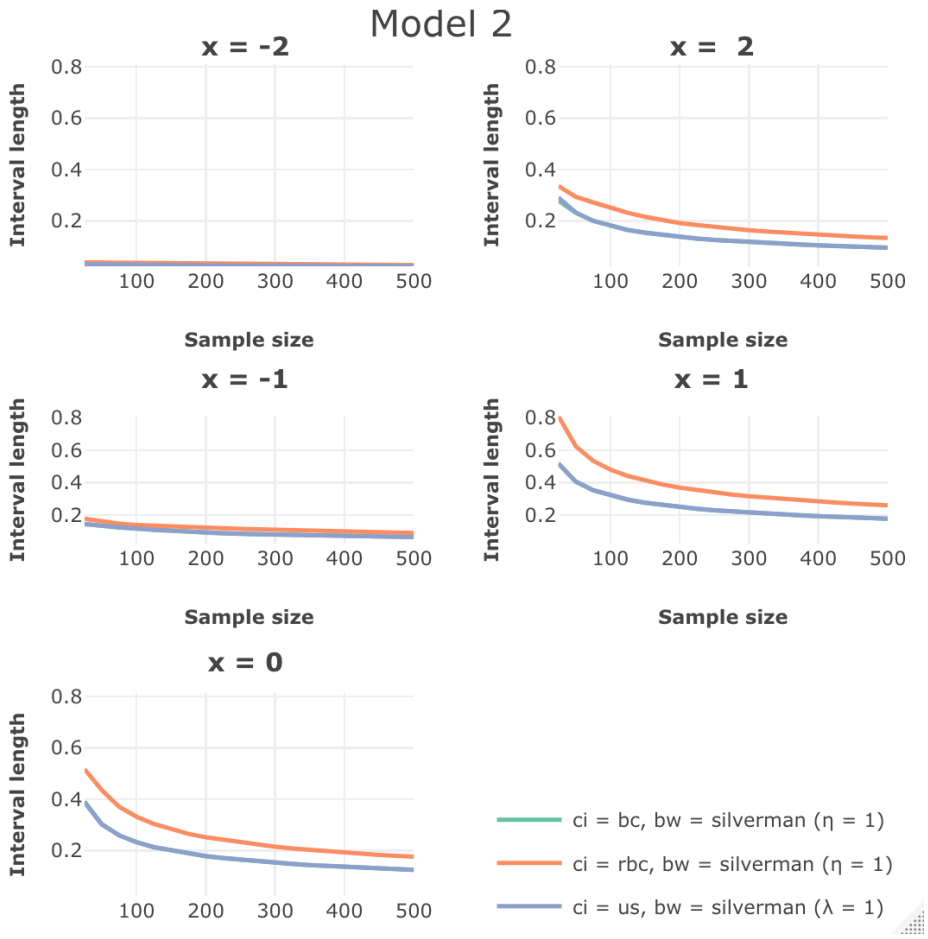

Figure 26: Interval length for best combinations (Model 2).

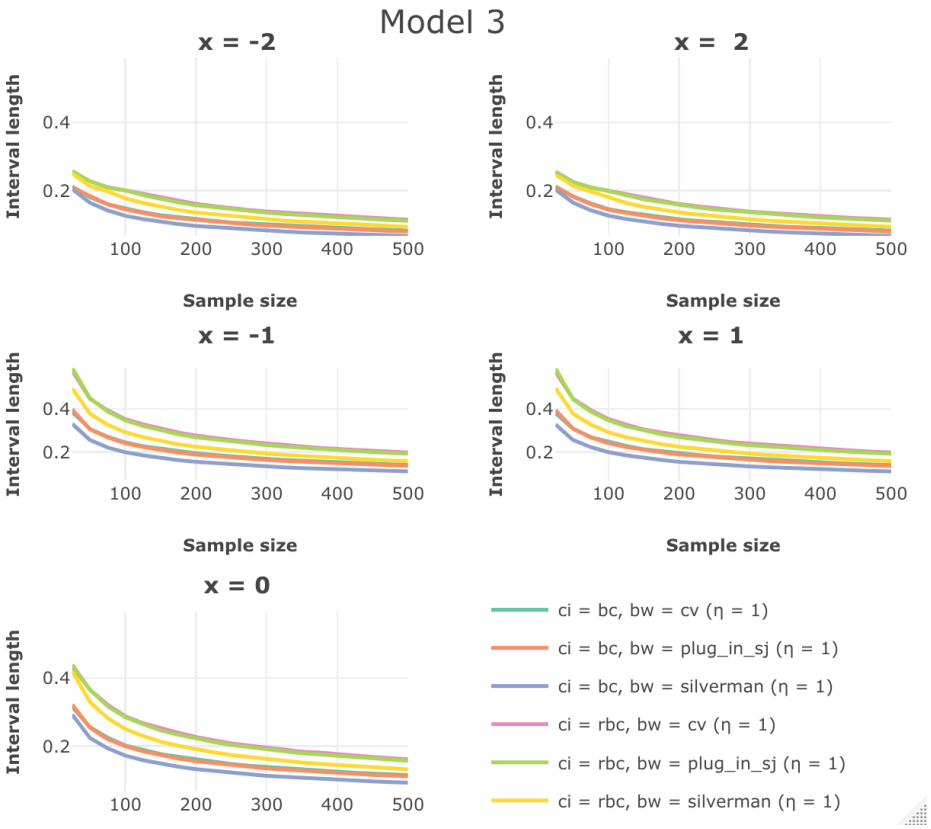

Figure 27: Interval length for RBC and BC (Model 3).

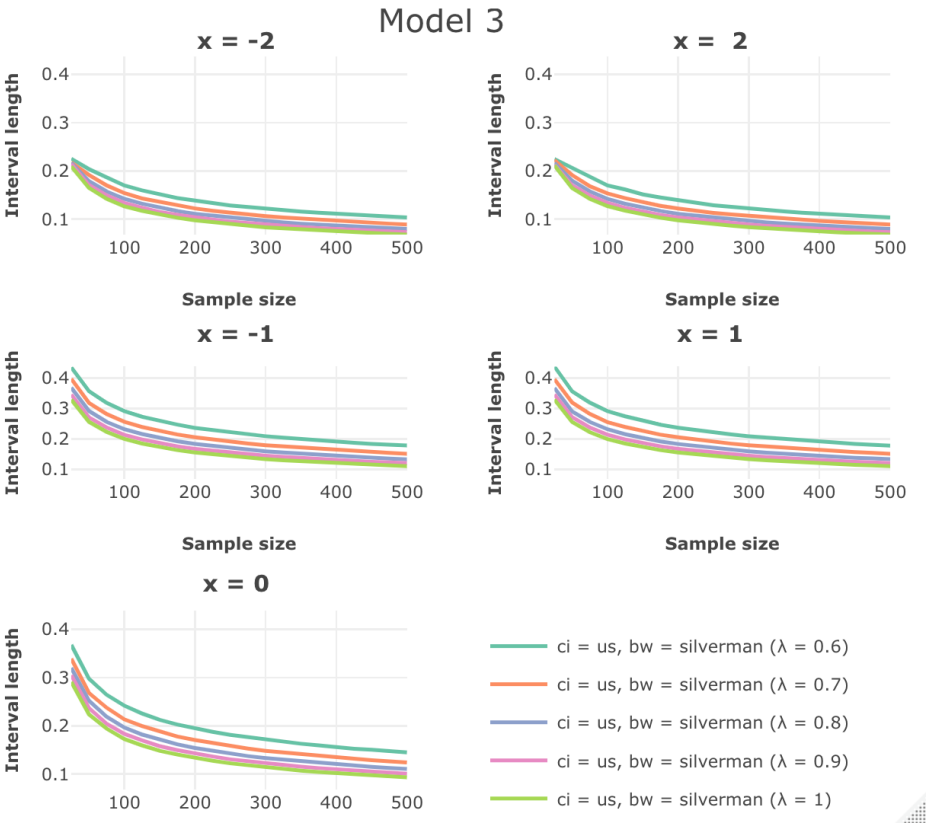

Figure 28: Interval length for US using $\hat{h}_{\mathrm{RT}}(\lambda)$ (Model 3).

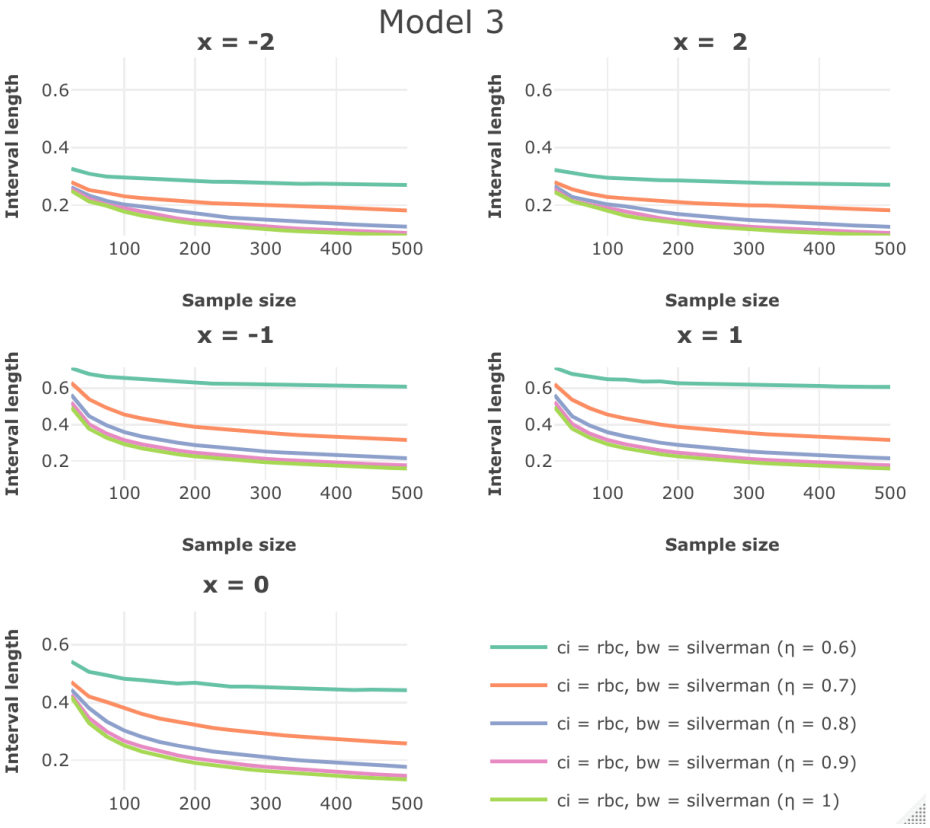

Figure 29: Interval length for RBC using $\hat{b}_{\mathrm{RT}}(\eta)$ (Model 3).

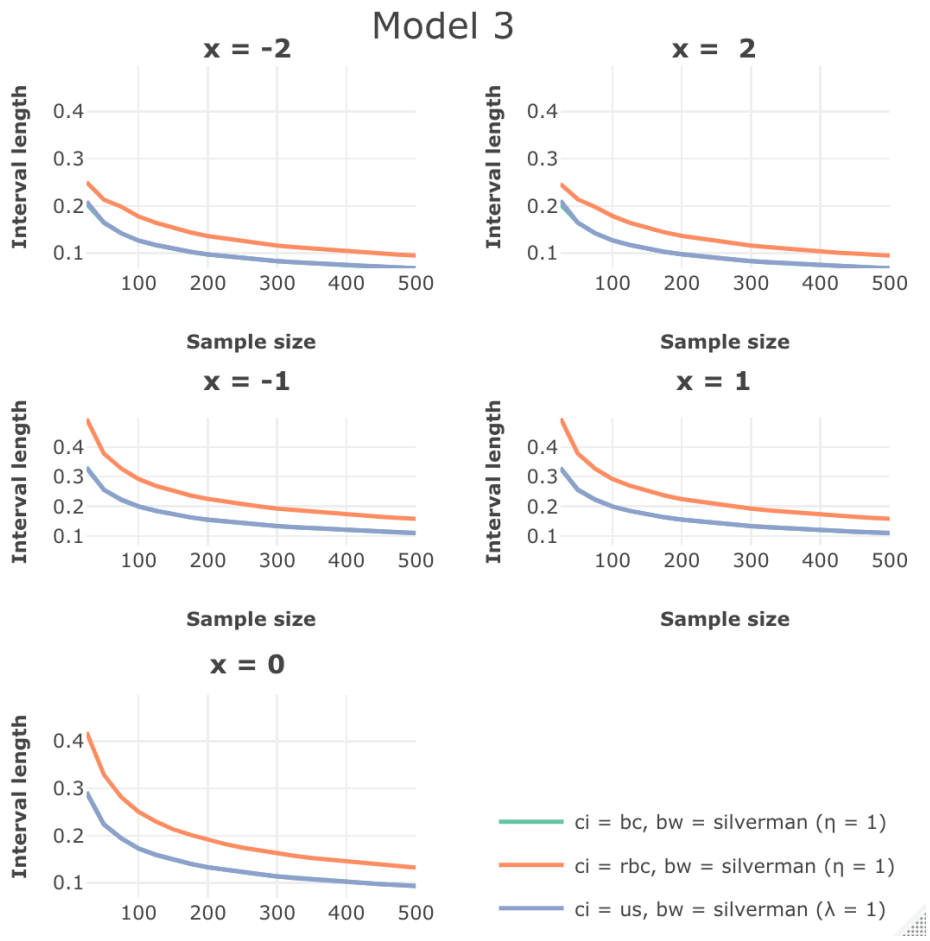

Figure 30: Interval length for best combinations (Model 3).

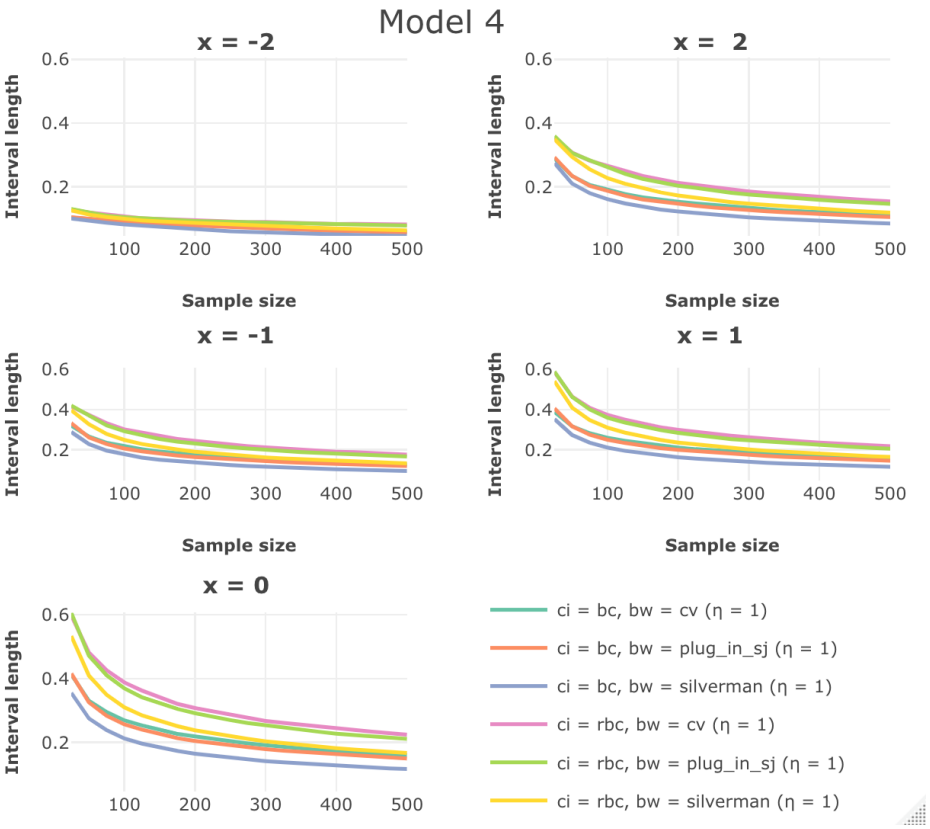

Figure 31: Interval length for RBC and BC (Model 4).

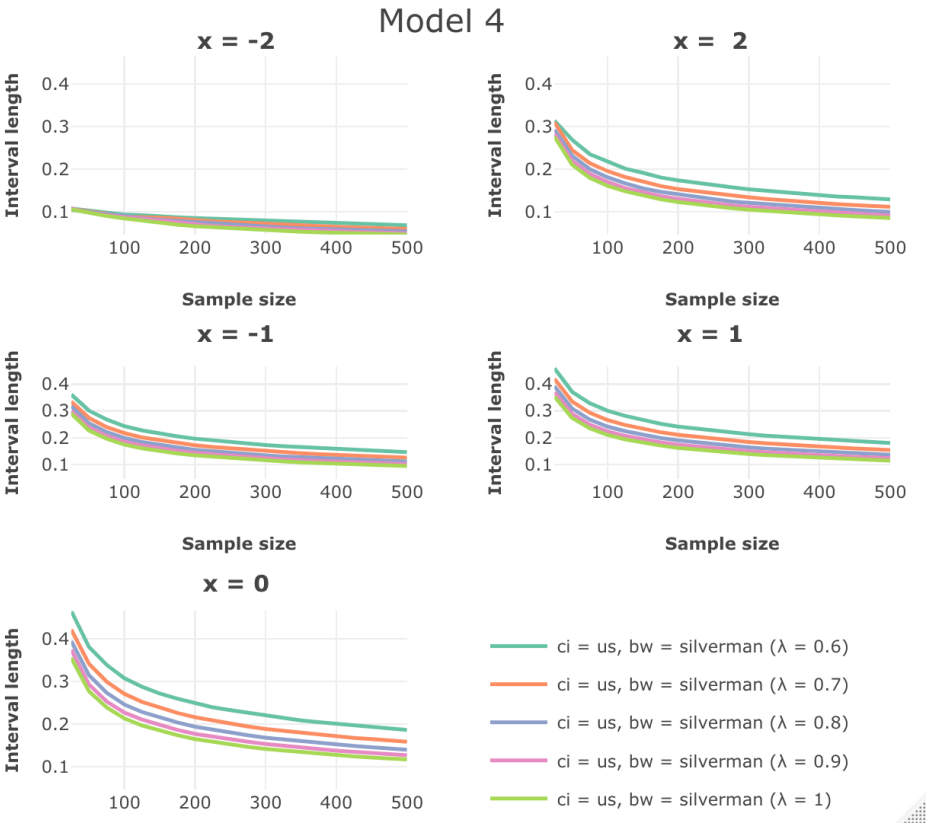

Figure 32: Interval length for US using $\hat{h}_{\mathrm{RT}}(\lambda)$ (Model 4).

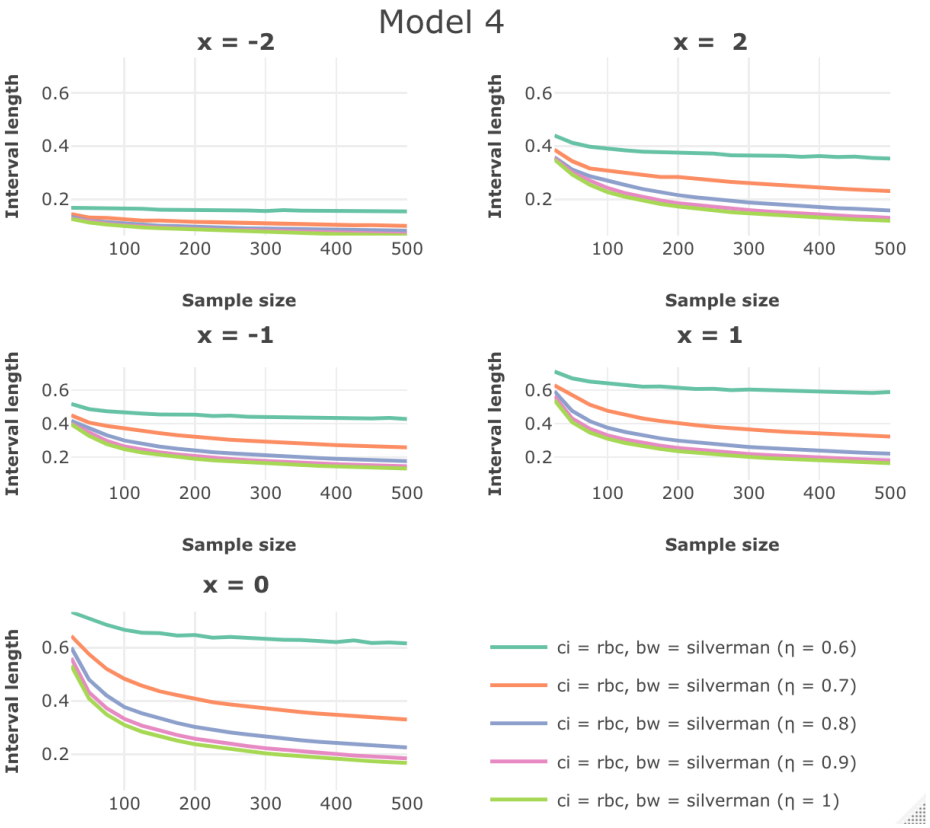

Figure 33: Interval length for RBC using $\hat{b}_{\mathrm{RT}}(\eta)$ (Model 4).

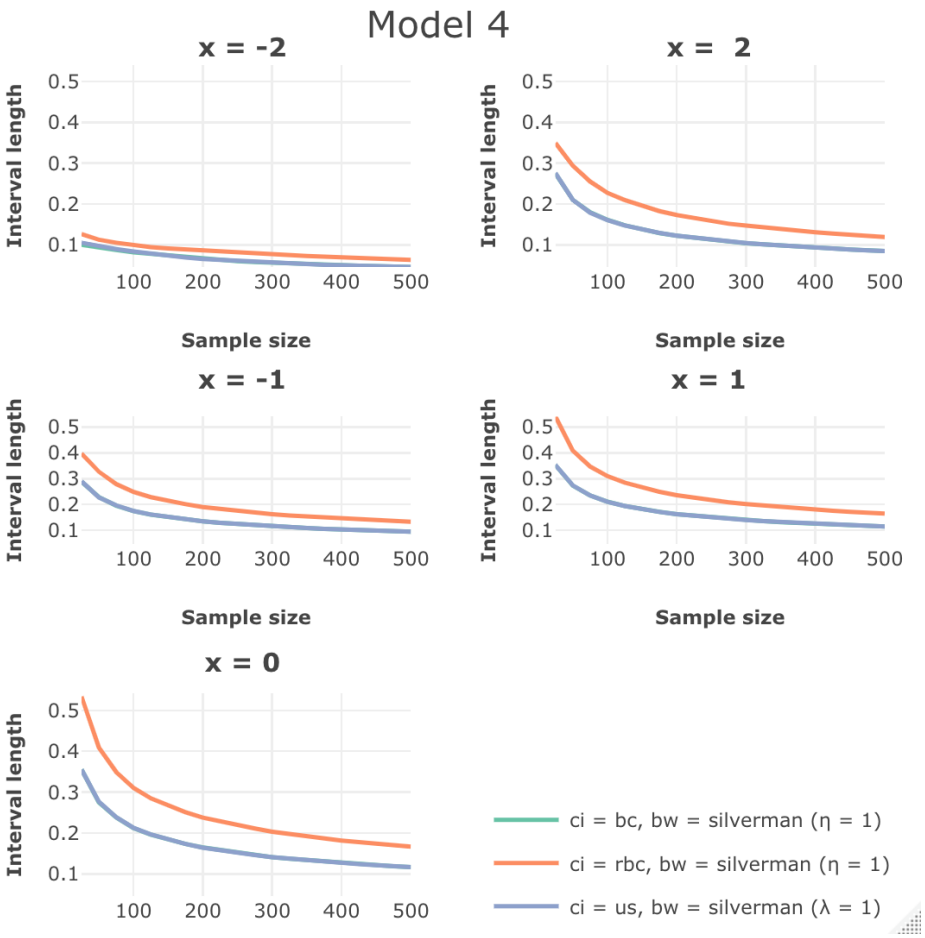

Figure 34: Interval length for best combinations (Model 4).