UTILIZING OBSERVATIONS AND RANKS IN KERNEL ESTIMATION

By

NICHOLAS ARMSTRONG KAUKIS

Bachelors of Arts in Statistics
Oklahoma State University
Stillwater, Oklahoma
2009

Master of Science in Statistics
Oklahoma State University
Stillwater, Oklahoma
2014

Submitted to the Faculty of the
Graduate College of the
Oklahoma State University
in partial fulfillment of
the requirements for
the Degree of
DOCTOR OF PHILOSOPHY
December, 2016

UTILIZING OBSERVATIONS AND RANKS

IN KERNEL ESTIMATION

Dissertation Approved:

Dr. Ibrahim Ahmad

Dissertation Adviser

Dr. Melinda McCann

Committee Chair

Dr. Carla Goad

Dr. Ye Liang

Dr. Jay Schweig

Outside Committee Member

Name: NICHOLAS KAUKIS

Date of Degree: DECEMBER, 2016

Title of Study: UTILIZING OBSERVATIONS AND RANKS IN KERNEL ESTIMATION

Major Field: STATISTICS

Abstract: For univariate i.i.d. samples, analyses are usually performed on observations themselves, or on the ranks of the observations when they are ordered. This is true in both the parametric and non-parametric setting. However, observations and ranks can be utilized simultaneously. This work concentrates on developing methods that combine observations and ranks in kernel-based estimation methods. The standard kernel density estimator, transformation kernel density estimator, and the Nadaraya-Watson kernel regression estimator are modified using ranks. Excluding the modified transformation kernel density case, the asymptotic properties are investigated for the estimators. For all estimators, various bandwidth selection methods are explored. Simulation is used to compare the modified estimators to the standard estimators. It is demonstrated that when samples are taken from skewed distributions, the modified kernel density estimator outperforms the standard estimator. Additionally, the modified Nadaraya-Watson outperforms the standard estimator in a variety of cases. The modified transformation density estimator did not outperform the unmodified estimator.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER I

LITERATURE REVIEW

1.1 Kernel Density Estimation

Consider a random sample $X_1, X_2, \ldots, X_n$. It is typically assumed that the random sample is taken such that each X_i comes from the same population and the value of X_i is characterized by the probability density function $f(x)$. In general, $f(x)$ may be considered unknown or may be assumed to be of one several different families of density functions. If $f(x)$ is unknown, an estimate of it may be desired. The simplest way to estimate $f(x)$ is to use a histogram that is scaled to 'density' on the vertical axis. Typically, A_{h1}, ..., A_{hk} bins of equal length h which may be referred as the *bandwidth* or *binwidth* are created so that the bins cover the entire range of the random sample. Then the number of values falling into a given bin are counted which is then converted to a 'density' given by

$$\frac{\sum_{i=1}^{n} I(x_i \in A_{hj})}{nh}.$$

These densities are plotted on the vertical axis across the range of a bin so that the area underneath the bin heights is equal to one.

An example histogram for data using varying bandwidths is given in Figure 1.1. It can be seen that the bandwidth will affect how much detail can be seen in the histogram. If h is too large, there is not as much detail about the relative shape of the estimated density formed by the data, whereas if h is too small, too much variability in the data can be seen and too many peaks appear in the histogram. Work by Rosenblatt et al. (1956) and Parzen

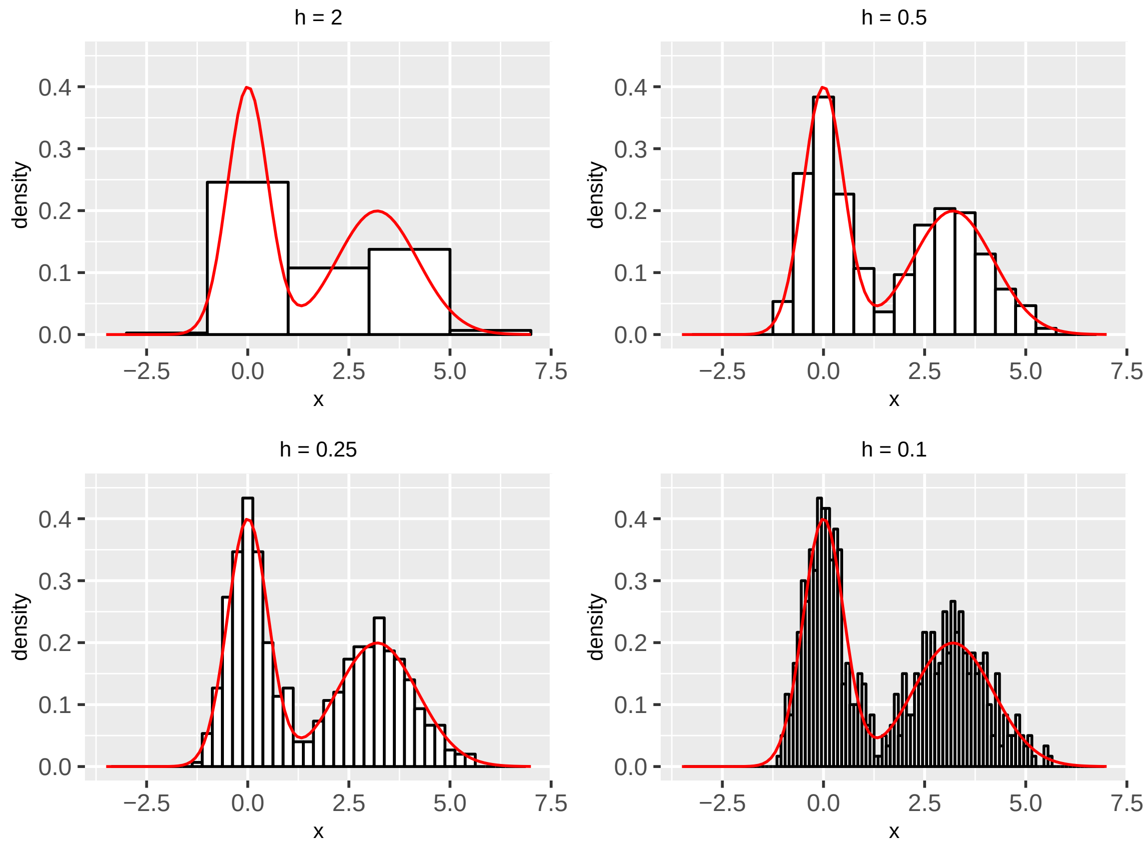


Figure 1.1: Depiction of a 4 histograms with 4 different bandwidths. Data are 600 random values from $0.5 \cdot \text{Norm}(0, 0.5^2) + 0.5 \cdot \text{Norm}(3.2, 1^2)$ mixture distribution.

(1962) generalized this idea of density estimation into the following form

$$\hat{f}_h(x) = \frac{1}{nh} \sum_{i=1}^{n} K\left(\frac{x - X_i}{h}\right), \tag{I.1}$$

where K is the "kernel function" and h is the "bandwidth". The central idea is to create a "bin" around every possible value of x and count the number of values in that bin weighted by K. K is typically chosen such that it is the density function of some known distribution. Often, the following assumptions are made for K:

1. $K(u) \geq 0, -\infty < u < \infty$

2. $\int K(u)\, du = 1$*

3. $\int uK(u)\,du = 0$*

4. $\sigma_K^2 := \int u^2 K(u)\,du < \infty$*.

Note that for all integrals in this paper, they are assumed to be from $-\infty$ to ∞ unless stated otherwise. While assumption 1 is utilized to ensure non-negativity of the estimate, there has been work (Bartlett 1963; Jones and Signorini 1997) showing that improvements based on an error criterion, discussed later in this paper, can be made by utilizing kernels that violate this non-negativity assumption.

As in the histogram case, one of the most important issues is the choice of the bandwidth h. This importance can be seen in Figure 1.1. As before, if h is too large, the finer details of the density estimate are lost, whereas if h is too small, too much fine structure is represented. These situations are referred to as oversmoothing and undersmoothing, respectively. There are a variety of methods for automatic selection of h which are based on the statistical properties of the kernel density estimator. More comprehensive reviews of the results of kernel density estimation are available by Wand and Jones (1994), and Silverman (1986).

1.1.1 Kernel Density Estimation: Statistical Properties

The kernel density estimate can be evaluated using the standard criteria of bias, variance, and mean squared error (MSE). Through a Taylor series approximation it can be seen that the expectation of $\hat{f}_h(x)$ is

$$\begin{aligned} Bias(\hat{f}_h(x)) &= \int K\left(\frac{x-y}{h}\right) f(y)\,dy - f(x) \\ &= \int K(u) f(x-uh)\,du - f(x) \\ &= \int K(u)\left(f(x) - uhf'(x) + \frac{u^2h^2}{2} f''(x)\right) du + o(h^2) - f(x) \\ &= \frac{\sigma_K^2 h^2}{2} f''(x) + o(h^2). \end{aligned} \tag{I.2}$$

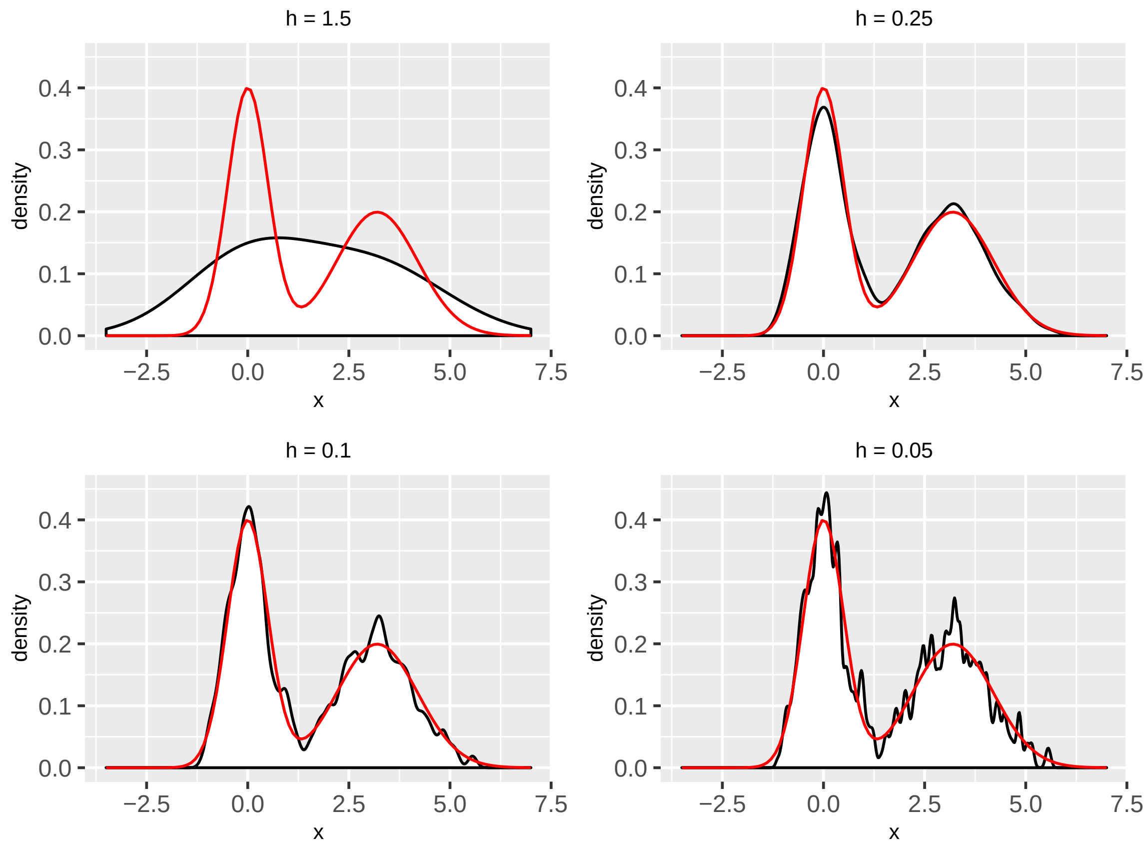


Figure 1.2: Depiction of a kernel density estimation of the same data as in Figure 1.1 with four different bandwidths.

In a similar manner, the variance of $\hat{f}_h(x)$ can be derived:

$$\begin{aligned} Var(\hat{f}_h(x)) &= \frac{1}{nh}\int K(u)^2 f(x+uh)\,dy - \frac{1}{n}E(\hat{f}_h(x))^2 \\ &= \frac{1}{nh}\int K(u)^2\,(f(x)+o(1))\,dy - \frac{1}{n}\left(\frac{\sigma_K^2 h^2}{2}f''(x)+o(h^2)\right) \\ &= \frac{R(K)}{nh}f(x)+o\left((nh)^{-1}\right). \end{aligned} \tag{I.3}$$

Note that here, the notation $R(K) = \int K(x)^2\,dx$ is used.

Taking the results of (II.20) and (I.3), the MSE is given by

$$MSE(\hat{f}_h(x)) = \frac{R(K)}{nh}f(x) + \frac{\sigma_K^4 h^4}{4}f''(x)^2 + o\left((nh)^{-1}+h^4\right). \tag{I.4}$$

Note that this gives a point-wise measurement of error. In density estimation, the interest would be on a measurement of error over all possible values of x. This is the motivation behind what is referred to as the mean integrated squared error (MISE).

$$MISE(\hat{f}_h(x)) := \int E[(\hat{f}_h(x) - f(x))^2]\, dx \tag{I.5}$$

Using the above definition. The MISE of $\hat{f}_h(x)$ is derived by integrating (I.4).

$$MISE(\hat{f}_h(x)) := \frac{R(K)}{nh} + \frac{\sigma_K^4 h^4}{4} R(f'') + o\left((nh)^{-1} + h^4\right)$$

For application, however, the Asymptotic MISE (AMISE) is used for analyses and is the basis for some methods of bandwidth selection. It is given by

$$AMISE(\hat{f}_h(x)) := \frac{R(K)}{nh} + \frac{\sigma_K^4 h^4}{4} R(f''). \tag{I.6}$$

1.1.2 **Kernel Density Estimation: Bandwidth Selection**

There are several automatic selection methods available for kernel density estimation (KDE). Three main types will be discussed: rule-of-thumb plug-in, cross-validation, and direct plug-in. The rule of thumb plug-in method is the simplest method, it is attributed to Silverman (1986). It is based on the idea of minimizing the AMISE given by I.6 with respect to h. This leads to the following solution for h:

$$h_{AMISE} = \left(\frac{R(K)}{\sigma_K^2 R(f'') n} \right)^{1/5}. \tag{I.7}$$

The major issue with this optimal h is that it requires the use of f'' which necessitates direct knowledge of the density function being estimated. This is the main issue that selection rules must overcome.

The Silverman Rule-of-Thumb estimator assumes that the density being estimated is normally distributed with some mean μ and variance σ^2. This leads to the following optimal h:

$$h = \left(\frac{8\sqrt{\pi}R(K)}{3\sigma_K^2 n} \right)^{1/5} \sigma.$$

Note that this formula requires knowledge of the true value of σ which may not be available. Consequently a reasonable estimate of σ is usually utilized such as the sample standard deviation s or $\hat{\sigma}_{IQR}$, an estimate based on scaling the interquartile range of the sample. Whichever estimate of σ is used, designate it $\hat{\sigma}$. Thus the Silverman Rule-of-Thumb bandwidth selector is

$$h_{ROT} = \left(\frac{8\sqrt{\pi}R(K)}{3\sigma_K^2 n} \right)^{1/5} \hat{\sigma}. \tag{I.8}$$

Another bandwidth selector is the least squared cross-validation (LSCV) selector which is due to the work of Rudemo (1982) and Bowman (1984). The basis for an LSCV bandwidth is to find a way to estimate the parts of the expanded MISE (I.5) that depend on h, and choose h such that the estimated portion of the MISE is minimized. This gives the LSCV bandwidth h_{LSCV} of

$$h_{LSCV} = \arg\min_h \left(\int \hat{f}_h(x)^2 \, dx - \frac{2}{n} \sum_{i=1}^{n} \hat{f}_{-i,h}(X_i) \right), \tag{I.9}$$

where $\hat{f}_{-i,h}(x)$ is the density estimate created by omitting the i^{th} value from the random sample, the leave-one-out cross-validation kernel density estimator of $f(x)$.

Combining this idea of cross-validation with plug-in estimation results in the biased cross-validation (BCV) bandwidth selection of h from Scott and Terrell (1987). The BCV selection method involves optimizing the AMISE (I.6) via an estimator of $R(f'')$ denoted by $\widehat{R(f'')}$. This is given by

$$\widehat{R(f'')} = \frac{1}{n^2} \sum\sum_{i \neq j} K'' * K'' \left(\frac{X_i - X_j}{h} \right).$$

Note that $f * g$ is the convolution of two functions f and g. This estimator $\widehat{R(f'')}$ is used to estimate the AMISE which gives the form for the BCV bandwidth h_{BCV}

$$h_{BCV} = \arg\min_h \left(\frac{R(K)}{nh} + \frac{\sigma_K^4 h^4}{4} \widehat{R(f'')} \right) \tag{I.10}$$

Besides these bandwidth selectors, another type of data-driven bandwidth is that of Park and Marron (1990) with modest refinements by Sheather and Jones (1991). The approach is similar in some respects to both the biased cross-validation (I.10) and the rule-of-thumb idea (I.8). The main hurdle in trying to estimate the asymptotically optimal bandwidth h_{AMISE} is in estimating the $R(f'')$ term in (I.7). The plug-in bandwidth h_{SJ} is given as outlined by Sheather and Jones (1991) by the solution to

$$\left(\frac{R(K)}{\sigma_K^4 \hat{S}(\hat{\alpha}(h))} \right)^{1/5} - h = 0 \tag{I.11}$$

where

$$\hat{S}(\alpha) = \frac{1}{n(n-1)\alpha^5} \sum_{i=1}^{n} \sum_{j=1}^{n} \phi^{iv} \left(\frac{X_i - X_j}{\alpha} \right)$$

and

$$\hat{\alpha}(h) = 1.357 \left(\frac{\hat{S}(a)}{\hat{T}(b)} \right).$$

$\hat{T}$ is defined similarly to $\hat{S}$,

$$\hat{T}(b) = -\frac{1}{n(n-1)b^{-7}} \sum_{i=1}^{n} \sum_{j=1}^{n} \phi^{vi} \left(\frac{X_i - X_j}{b} \right).$$

Using $\hat{\lambda}$ to denote the sample inter-quartile range,

$$a = 0.920\hat{\lambda}n^{-1/7} \quad ; \quad b = 0.912\hat{\lambda}n^{-1/9}.$$

Thus, obtaining h_{SJ} is a root finding procedure where a numerical approach must be utilized.

A review of the performance of these bandwidth selectors and others is given by Jones, Marron, and Sheather (1996). The findings were that h_{ROT} tends to oversmooth, h_{LSCV} is too variable and tends to undersmooth, h_{BCV} tends to oversmooth, and h_{SJ} typically performs best. These findings can change depending on the complexity of the underlying density, but similar results were found by Cao, Cuevas, and Manteiga (1994) and Park, Turlach, et al. (1992). Confirmation of these conclusions are given Loader (1999) where it was indicated that h_{SJ} is less variable but it still tends to be too large whenever h_{AMISE}, the optimal bandwidth, is quite small.

A more recent, and possibly the most comprehensive review, of bandwidth selection methods is given by Heidenreich, Schindler, and Sperlich (2013). Their examinations indicated that small modifications of the previously mentioned methods seem best. For example, a modification called one-sided cross-validation, initially proposed for regression problems by Hart and Yi (1998), stabilized the LSCV bandwidth technique and did not increase the bias much. This is taken a step further with Savchuk, Hart, and Sheather (2010), where a estimator is proposed based off of ”indirect cross-validation.” Heidenreich, Schindler, and Sperlich (2013) indicate that the indirect cross-validation method and a refinement of the Sheather-Jones plug-in bandwidth typically perform best, though when sample size increases, cross-validation techniques become computationally expensive.

1.2 **Kernel Regression**

Another application of the usage of kernels is through that of kernel regression. Let $(X_i, Y_i), i = 1, \ldots, n$ be an IID random sample from a bivariate distribution where Y_i is described by the following model:

$$Y_i = g(X_i) + \varepsilon_i.$$

The function $g(x) = \mathrm{E}[Y|X = x]$ is a smooth function, not necessarily linear, and ε_i is a random variable centered at zero with finite variance σ_ε^2. The objective is typically to estimate $g(x)$. One approach to estimating this function using kernels is from the work of Nadaraya (1964) and Watson (1964). The estimator proposed is

$$\hat{g}_{NW}(x) = \frac{\sum_{i=1}^n K\left(\frac{x-X_i}{h}\right) Y_i}{\sum_{i=1}^n K\left(\frac{x-X_i}{h}\right)}. \tag{I.12}$$

The intuition behind this estimator lies in trying to estimate the conditional mean,

$$\mathrm{E}[Y|X = x] = \frac{\int y f(x,y) dy}{f(x)}.$$

The idea is to estimate the displayed densities. The estimation of $f(x)$, the density of X, takes the form of the previously discussed KDE methods. For the joint density $f(x,y)$, an approach is used where $\hat{f}(x,y) = \frac{1}{nh^2} \sum_{i=1}^n K\left(\frac{x-X_i}{h}\right) K\left(\frac{y-Y_i}{h}\right)$. Integrating over $\hat{f}(x,y)$ leaves us with the Nadaraya-Wayson estimator. Defining

$$W_{hi}(x) = \frac{K\left(\frac{x-X_i}{h}\right)}{\sum_{i=1}^n K\left(\frac{x-X_i}{h}\right)},$$

(I.12) can be rewritten as

$$\hat{g}_{NW}(x) = \sum_{i=1}^{n} W_{hi}(x) Y_i, \tag{I.13}$$

so that any estimate of $g(x)$ is a weighted average of the Y_i values.

1.2.1 Nadaraya-Watson Estimator: Statistical Properties

A comprehensive review of kernel regression estimators, of which the Nadaraya-Watson estimator is a special case, can be found in Wand and Jones (1994) or more extensively in Li and Racine (2007) and Hardle et al. (2012). A brief summary of those works will be given in this document. The main results will be displayed without proof, as they require more extensive and tedious workings. The main points of discussion for the Nadaraya-Watson estimator will be the bias, variance, and then bandwidth selection.

First, the bias is given as follows.

$$Bias(\hat{g}_{NW}(x)) = \left(\frac{\sigma_K^2 g''(x)}{2} + \frac{g'(x) f'(x)}{f(x)} \right) h^2 + o(h^2) + O\left(\frac{1}{nh} \right) \tag{I.14}$$

The form of this bias can be seen as structurally similar to that of the bias of $\hat{f}_h(x)$ in (II.20), but complexity is added by the inclusion of the g function. As long as it is assumed that $h \to 0$ and $nh \to \infty$ as $n \to \infty$, it can be seen that the bias goes to zero.

A similar parallel can be seen in the variance of $\hat{g}_{NW}(x)$,

$$Var(\hat{g}_{NW}(x)) = \frac{1}{nh} \cdot \frac{\sigma_\varepsilon^2}{f(x)} R(K) + o\left(\frac{1}{nh} \right). \tag{I.15}$$

Details for the bias and variance results were first characterized by Collomb (1976).

1.2.2 Nadaraya-Watson Estimator: Bandwidth Selection

Just like with the density estimation case, an important attribute of kernel-based regression is the choice of the bandwidth. The choice of bandwidth in the regression case is mainly based in the estimation and subsequent optimization of the average squared error (ASE), which is defined in the following:

$$ASE(\hat{g}_{NW}(x)) = \frac{1}{n}\sum_{i=1}^{n}\left(g(X_i) - \hat{g}_{NW}(X_i)\right)^2.$$

An issue with the ASE is that it requires knowledge of the true value of $g(X_i)$. The function g is not known, and we only know Y_i, which is $g(X_i)$ convoluted with the error term. An obvious solution is to estimate $g(X_i)$ with $Y_i = g(X_i) + \varepsilon_i$. This gives the residual sum of squares (RSS),

$$RSS(\hat{g}_{NW}(x)) = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{g}_{NW}(X_i))^2. \tag{I.16}$$

Now, the desired objective is the minimization of this value by the choice of h. Unfortunately, $RSS \to 0$ as $h \to 0$ since this will guarantee that $\hat{g}_{NW}(X_i) = Y_i$. To counteract this, a penalization function is applied to (I.16). Another approach is to use cross-validation techniques similar to those utilized in the density estimation bandwidth selection.

The cross-validation bandwidth selection, initially explored by Clark (1977) for the regression estimation case, is based on using a leave-one-out estimator. Let $\hat{g}_{-i,h}(x)$ define the leave-one-out Nadaraya-Watson estimator where the i^{th} value is removed from the sample. Then the cross-validation bandwidth selector is chosen as the value of h such that

$$CV(h) = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{g}_{-i,h}(x))^2, \tag{I.17}$$

is minimized.

An alternative approach is to penalize the RSS in (I.16) by multiplying each term in the sum by a function $\Xi(u)$ such that the first order Taylor expansion of Ξ is $\Xi(u) = 1 + 2u + O(u^2)$ whenever $u \to 0$. The value substituted for u is $\frac{1}{n}W_{hi}(x)$. Denote a penalized sum of squares by G, then the objective is to minimize

$$G(h) = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{g}(X_i))^2 \Xi\left(\frac{1}{n}W_{hi}(X_i)\right)$$

with respect to h. Several approaches to the choice of Ξ have been proposed. A very direct approach is to use

$$\Xi_S(u) = 1 + 2u,$$

which is due to Shibata (1981). Work from Craven and Wahba (1978) gives the generalized cross-validation penalization function of

$$\Xi_{GCV} = \frac{1}{(1-u)^2}.$$

Li and Racine (2007) in their section on bandwidth selection recommend the use of a modified penalization based on the Akaike information criterion (AIC) due to Akaike (1974). The modified AIC is given by Hurvich, Simonoff, and Tsai (1998) and is denoted by AIC_C,

$$AIC_C(h) = \log(\hat{\sigma}^2) + \frac{1 + \frac{\sum_{i=1}^{n} W_{hi}(X_i)}{n}}{1 - \frac{\sum_{i=1}^{n} W_{hi}(X_i) + 2}{n}}, \tag{I.18}$$

where $\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{g}(x))^2$.

There are many different approaches to bandwidth selection and a comprehensive review of them by Köhler, Schindler, and Sperlich (2014) provides many more examples and details. The most widely used methods are briefly summarized above.

CHAPTER II

RANK WEIGHTED KERNEL DENSITY ESTIMATION

The objective now is to take the density and regression estimated techniques discussed in the previous chapter and improve them by including additional information that is always available when a random sample is taken, the rankings. Typically, in parametric estimation, the values of a sample are all that are used; when dealing with classical nonparametric techniques, non-parametric techniques such as the Kruskal-Wallis Test (Kruskal and Wallis 1952) or the Mann-Whitney-U Test (Mann and Whitney 1947), only the sample ranks are used. In hopes of trying to gain improved estimation abilities, here a technique for combining the information of the ranks and sample values will be presented. First, the density estimation case will be discussed where the bias, variance, MSE, and MISE will be derived. This will lead to discussion of bandwidth selection for the rank weighted kernel density estimator discussed below. The next chapter will be similar but focuses on introducing ranks into the estimator for the kernel regression case.

Assume an i.i.d. random sample $X_1, X_2, \ldots, X_n$ is taken from a ”well” behaving distribution with CDF $F(x)$ and density $f(x)$. By well behaved distribution, it is meant that the variance is finite and the density is continuous. Additionally, the derivatives of the density up to and including the second derivative are assumed to be continuous. The estimator being proposed in this chapter uses the order statistics $X_{(1)}, X_{(2)}, \ldots, X_{(n)}$, where the density of the i^{th} order statistic is denoted by $f_{(i)}(x)$ which is

$$f_{(i)}(x) = \frac{n!}{(i-1)!\,(n-i)!}\left[F(x)\right]^{i-1}\left[1-F(x)\right]^{n-k} f(x)$$

The proposed rank weighted kernel density estimator of $f(x)$ is:

$$\hat{f}_h(x) = \frac{1}{nh}\sum_{i=1}^{n}\frac{1}{\sqrt{i}}K\left(\frac{x-X_{(i)}}{\sqrt{i}h}\right)$$

Note that standard assumptions on the kernel K are being made, namely $\int K(z)\,dz = 1$, $\int zK(z)\,dz = 0$, and $\int z^2K(z)\,dz = \sigma_K^2 < \infty$. An obvious concern is to make sure that the estimator integrates to one. Here, that is relatively simple to show.

$$\begin{aligned}\int \hat{f}_h(x)\,dx &= \frac{1}{nh}\sum_{i=1}^{n}\frac{1}{\sqrt{i}}\int K\left(\frac{x-X_{(i)}}{\sqrt{i}h}\right)dx \\ &= \frac{1}{nh}\sum_{i=1}^{n}\int\frac{1}{\sqrt{i}}K(y)\sqrt{i}h\,dy \\ &= \frac{1}{n}\sum_{i=1}^{n}\int K(y)\,dy \\ &= 1\end{aligned}$$

Next some motivation for this estimator will be presented. Let $f(x)$ represent the density function of a right skewed distribution such as the exponential distribution or what is referred to as the skewed normal distribution. Plots of these density functions are given in Figure 2.1.

When the density is high, it is expected that several values in a random sample will be in that region, and when the density is low, far less values are expected in the region. If there are a lot of values within a small range in a random sample, a smaller bandwidth would be desirable. As data gets more sparse, a larger bandwidth is more desirable. A small bandwidth displays the finer structure of a distribution, such as sharp points or steeper slopes like in the skewed densities in Figure 2.1. A larger bandwidth is preferred in low density regions, where a low number of observations are expected, because a large bandwidth flattens out the kernel and prevents spikes from forming where there would not be spikes in the true density.

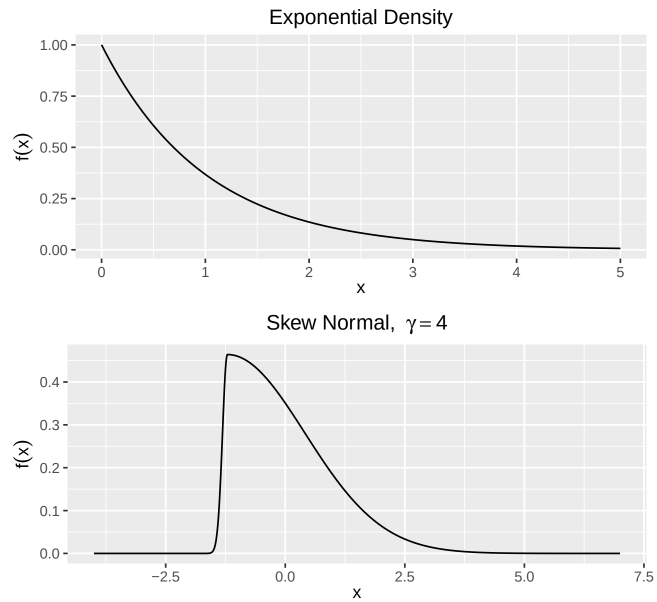


Figure 2.1: Plots of the exponential distribution and skew normal distribution.

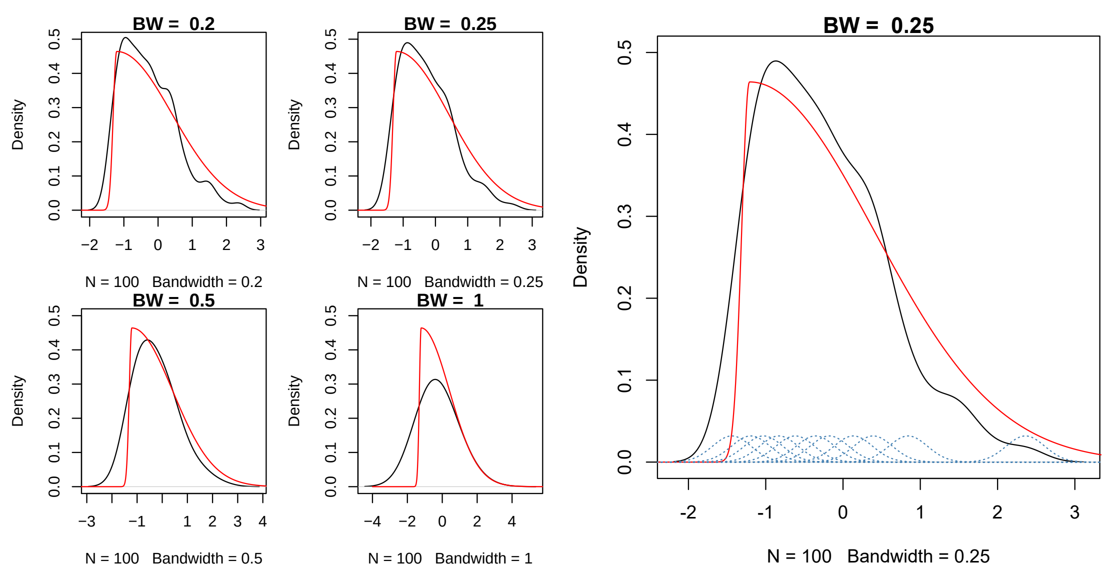


Figure 2.2: Density estimates for a set of data simulated from the skew normal distribution under four different bandwidths. True distribution is in red, and standard density estimate is in black. The picture on the right shows $\frac{1}{nh} \cdot K\left(\frac{(x-X_i)}{h}\right)$ for 11 observations ($i = 1, 10, 20, \cdots, 100$ in the ordered sample) in a sample of 100 used to create the density estimate. This depicts the weight given to sample values when trying to estimate the density at x. These are given in blue dotted lines.

Figure 2.2 tries to illustrate this trade off in fine structure versus bumpiness of density estimates for the standard density estimator. When the bandwidths are smaller like in the top two density estimates on the left image, the estimate may be fairly accurate around the peak but as x increases, what should be a smooth tail can be somewhat bumpy in the density estimate. In the density estimates with larger bandwidths on the bottom of the left image in Figure 2.2,the tail is now smooth but at lower values of x, the density estimate does not capture the form of the target distribution.

Because there are changes in the density value, there are changes in how the random sample is clustered. This means it is desirable to have a bandwidth that adapts to the degree of clustering in the data. With right skewed distributions, this simply means there should be a larger bandwidth as the estimation point x increases. Using the rank weights emulates this idea. If one were to be estimating a distribution with a heavy right tail and an extreme value is obtained, it would not be desirable to use the same bandwidth as where the mode of the distribution would be since that would induce a spike in the density estimate. Essentially, issues may arise when trying to estimate data from distributions with thicker tails.

For a right skewed distribution, the expected structure for a sample would be a clustering of observations with lower values and sparse observations at higher values. This makes it seem appropriate to start with a smaller bandwidth and then move to a larger bandwidth as there is progression along the x-axis, which is indicated by increasing ranks where the bandwidth given to a particular rank i is $\sqrt{i}h$. In this form, the square root of the rank is a scaling for h that changes, and $\sqrt{i}h$ is the actual bandwidth used for any given value of i in the ordered sample. Look to Figure 2.3 for a demonstration of how the increasing bandwidth works for the skew normal. The image on the right shows how the kernel function for a given observation flattens as the rank increases. On the left side of Figure 2.3, it can seen that when the bandwidth is too small in the top two density estimates, there is rather extreme bumpiness at lower values of x but smoothing in the tail. In the bottom two density estimates, there is a smooth estimate in both the high density region and right hand tail of the target distribution in red.

Looking at Figure 2.2, one can get better estimates of the tail at the sacrifice of the

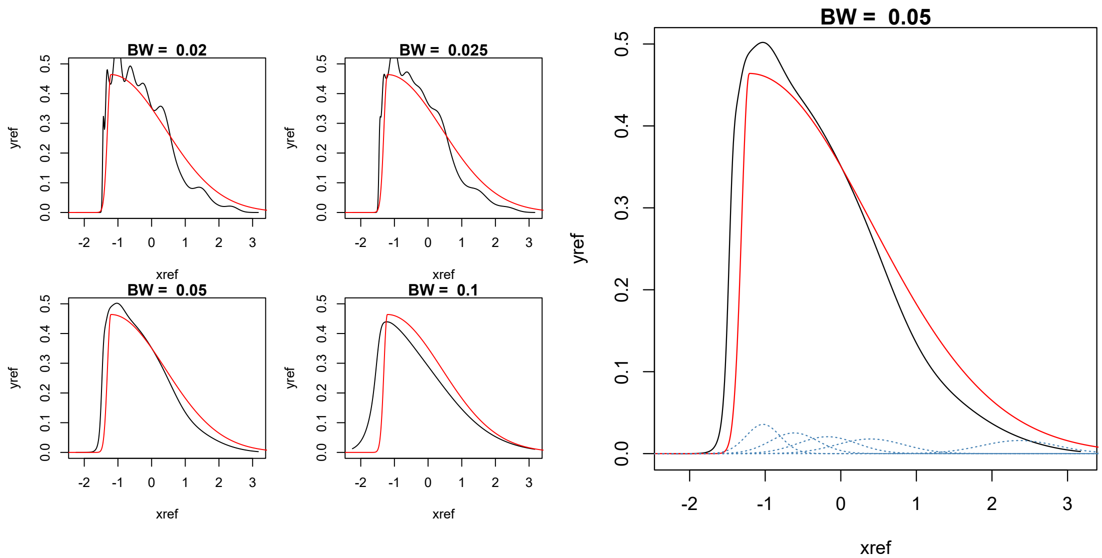


Figure 2.3: Similar to Figure 2.2 but for the rank weighted density estimator which is the black line. Again, the image on the right shows how the kernel function weights observations in blue dotted lines.

estimate near the mode. With Figure 2.3, smooth estimation of both the mode and tails can be achieved.

In the next sections, the asymptotic statistical properties of this estimator will be discussed. First a derivation of the form of the bias will be given, and then the variance will be derived. This will provide the foundation for deriving the asymptotic mean squared error (MSE) and subsequently the asymptotic mean integrated squared error (AMISE).

2.1 Statistical Properties

Here, the asymptotic form of the bias for the rank weighted kernel density estimator will be derived. First, note that unless otherwise stated, integrals are from $-\infty$ to ∞. The derivation is quite similar to that of the standard density estimation case. Additional complications arise due to the inclusion of the ranks which means that the order statistic distributions must be considered. To get the bias, first the following expectation will be derived.

The assumptions of the kernel function K are the same as before:

1. $K(u) \geq 0, -\infty < u < \infty$

2. $\int K(u)\,du = 1$

3. $\int uK(u)\,du = 0$

4. $\sigma_K^2 := \int u^2 K(u)\,du < \infty$.

Using these assumptions and some mild assumptions on $f(x)$, the following theorem is derived.

Theorem 2.1.1**.** *Let assumptions 1 - 4 on K be true. Let $f(x)$ be a probability density function with finite mean and variance. Additionally, assume that $f(x)$, and its first two derivatives are continuous. Then*

$$Bias(\hat{f}_h(x)) = \frac{\sigma_K^2 h^2}{2}\left[f^{''}(x) + (n-1)\left(f^{''}(x)F(x) + 3f(x)f^{'}(x)\right)\right] + o(nh^2).$$

Additionally, assume that $h \to 0$ such that $nh \to \infty$, but $nh^2 \to 0$. Then $Bias(\hat{f}_h(x)) \to 0$.

Proof.

$$\begin{aligned}
\frac{1}{\sqrt{i}h}\mathrm{E}\left[K\left(\frac{x - X_{(i)}}{\sqrt{i}h}\right)\right] &= \frac{1}{\sqrt{i}h}\int K\left(\frac{x-y}{\sqrt{i}h}\right) f_{(i)}(y)\,dy \\
&= \int K(z)\, f_{(i)}(x - z\sqrt{i}h)\,dz \\
&= \int K(z)\left\{f_{(i)}(x) - z\sqrt{i}h\frac{d}{dx}f_{(i)}(x) + \frac{z^2 ih^2}{2}\frac{d^2}{dx^2}f_{(i)}(x) + o(ih^2)\right\}dz \\
&= f_{(i)}(x) + \sigma_K^2\frac{ih^2}{2}\frac{d^2}{dx^2}f_{(i)}(x) + o(ih^2)
\end{aligned}$$

Then the expected value of $\hat{f}_h(x)$ is

$$\begin{aligned}
\mathrm{E}\left[\hat{f}_h(x)\right] &= \frac{1}{nh}\sum_{i=1}^{n}\frac{1}{\sqrt{i}}\mathrm{E}\left[K\left(\frac{x - X_{(i)}}{\sqrt{i}h}\right)\right] \\
&= \frac{1}{n}\sum_{i=1}^{n}\left\{f_{(i)}(x) + \frac{\sigma_K^2 ih^2}{2}\frac{d^2}{dx^2}f_{(i)}(x) + o(ih^2)\right\}
\end{aligned}$$

$$
\begin{aligned}
&= f(x) + \frac{\sigma_K^2 h^2}{2n} \frac{d^2}{dx^2} \sum_{i=1}^{n} i f_{(i)}(x) + o(nh^2) \\
&= f(x) + \frac{\sigma_K^2 h^2}{2n} \frac{d^2}{dx^2} \sum_{i=1}^{n} i \frac{n!}{(i-1)!\,(n-i)!} [F(x)]^{i-1} [1-F(x)]^{n-i} f(x) + o(nh^2) \\
&= f(x) + \frac{\sigma_K^2 h^2}{2n} \frac{d^2}{dx^2} \frac{f(x)}{F(x)} \sum_{i=0}^{n} i^2 \frac{n!}{i!\,(n-i)!} [F(x)]^{i} [1-F(x)]^{n-i} + o(nh^2) \\
&= f(x) + \frac{\sigma_K^2 h^2}{2n} \frac{d^2}{dx^2} \frac{f(x)}{F(x)} \sum_{i=0}^{n} i^2 \binom{n}{i} [F(x)]^{i} [1-F(x)]^{n-i} + o(nh^2)
\end{aligned}
$$

Note that the sum is now in the form of the second moment of a binomial random variable of size n, and probability $F(x)$.

$$
\begin{aligned}
&= f(x) + \frac{\sigma_K^2 h^2}{2n} \frac{d^2}{dx^2} \frac{f(x)}{F(x)} \left[nF(x)\,(1-F(x)) + n^2 F(x)^2\right] + o(nh^2) \\
&= f(x) + \frac{\sigma_K^2 h^2}{2n} \frac{d^2}{dx^2} \frac{f(x)}{F(x)} nF(x)\,[1-F(x)+nF(x)] + o(nh^2) \\
&= f(x) + \frac{\sigma_K^2 h^2}{2} \frac{d^2}{dx^2} \frac{f(x)}{F(x)} F(x)\,[1-F(x)+nF(x)] + o(nh^2) \\
&= f(x) + \frac{\sigma_K^2 h^2}{2} \frac{d^2}{dx^2} f(x)\,[1+(n-1)F(x)] + o(nh^2) \\
&= f(x) + \frac{\sigma_K^2 h^2}{2} \left[f''(x) + (n-1)\left(f''(x)F(x) + 3f(x)f'(x)\right)\right] + o(nh^2). \qquad \text{(II.1)}
\end{aligned}
$$

Thus, the bias of the estimator is

$$
\begin{aligned}
Bias(\hat{f}_h(x)) &= \frac{\sigma_K^2 h^2}{2} \left[f''(x) + (n-1)\left(f''(x)F(x) + 3f(x)f'(x)\right)\right] + o(nh^2) \qquad \text{(II.2)} \\
&= O(nh^2) \\
&\to 0.
\end{aligned}
$$

□

Next, the variance of the rank weighted kernel density estimator will be derived. There are many steps to this derivation. They will be broken down into several pieces and the final form will be presented at the end of this section. The main issue in this process is that by invoking the ranks and subsequent order statistics, the independence of the values is lost which complicates the form of the variance since it can no longer be assumed that the covariances are zero. This is in reference to the basic formula for the variance of the sum of random variables:

$$Var(\sum_{i=1}^{n} X_i) = \sum_{i=1}^{n} Var(X_i) + 2\sum\sum_{i<j} Cov(X_i, X_j).$$

If the independence assumption were available, the covariance term on the right would be 0.

Theorem 2.1.2**.** *Under the same assumptions of theorem* 2.1.1, the variance of $\hat{f}_h(x)$ is

$$Var(\hat{f}_h(x)) = \frac{R(K)}{n^2 h}\frac{f(x)}{F(x)}\,\mathrm{E}\left[\sqrt{B_{n,F(x)}}\right] + o(\sqrt{n}h)$$

which converges to 0.

Proof. To begin the variance calculation process, the format of the variance is expanded.

$$\begin{aligned}
Var(\hat{f}_h(x)) &= \frac{1}{n^2h^2} Var(\sum_{i=1}^{n}\frac{1}{\sqrt{i}}K\left(\frac{x-X_{(i)}}{\sqrt{i}h}\right)) \\
&= \sum_{i=1}^{n}\frac{1}{n^2h^2 i} Var(K\left(\frac{x-X_{(i)}}{\sqrt{i}h}\right)) + \frac{2}{n^2h^2\sqrt{ij}} Cov(K\left(\frac{x-X_{(i)}}{\sqrt{i}h}\right), K\left(\frac{x-X_{(j)}}{\sqrt{j}h}\right)) \\
&= \sum_{i=1}^{n}\frac{1}{n^2h^2 i}\left\{\mathrm{E}\left[\left(K\left(\frac{x-X_{(i)}}{\sqrt{i}h}\right)\right)^2\right] - \mathrm{E}\left[K\left(\frac{x-X_{(i)}}{\sqrt{i}h}\right)\right]^2\right\} \\
&\quad + \sum\sum_{i<j}\frac{2}{n^2h^2\sqrt{ij}}\left\{\mathrm{E}\left[K\left(\frac{x-X_{(i)}}{\sqrt{i}h}\right)K\left(\frac{x-X_{(j)}}{\sqrt{j}h}\right)\right] - \right. \\
&\qquad\qquad \left. \mathrm{E}\left[K\left(\frac{x-X_{(i)}}{\sqrt{i}h}\right)\right]\mathrm{E}\left[K\left(\frac{x-X_{(j)}}{\sqrt{j}h}\right)\right]\right\} \\
&= \sum_{i=1}^{n}\frac{1}{n^2h^2 i}\mathrm{E}\left[\left(K\left(\frac{x-X_{(i)}}{\sqrt{i}h}\right)\right)^2\right] + \sum\sum_{i<j}\frac{2}{n^2h^2\sqrt{ij}}\mathrm{E}\left[K\left(\frac{x-X_{(i)}}{\sqrt{i}h}\right)K\left(\frac{x-X_{(j)}}{\sqrt{j}h}\right)\right] \\
&\quad - \sum_{i=1}^{n}\sum_{j=1}^{n}\frac{1}{n^2\sqrt{ij}}\mathrm{E}\left[K\left(\frac{x-X_{(i)}}{\sqrt{i}h}\right)\right]\mathrm{E}\left[K\left(\frac{x-X_{(j)}}{\sqrt{j}h}\right)\right] \qquad \text{(II.3)}
\end{aligned}$$

Each sum of (II.3) will be handled separately and in the order given. In the following derivations, the standard notation,

$$R(f) = \int f(x)^2 dx$$

will be used.

The first sum can be simplified using the following.

$$\begin{aligned}
\frac{1}{h^2 i}\mathrm{E}\left[\left(K\left(\frac{x-X_{(i)}}{\sqrt{i}h}\right)\right)^2\right] &= \frac{1}{h^2 i}\int\left(K\left(\frac{x-y}{\sqrt{i}h}\right)\right)^2 f_{(i)}(y)\,dy \\
&= \frac{1}{h\sqrt{i}}\int (K(z))^2 f_{(i)}(x - z\sqrt{i}h)\,dz \\
&= \frac{1}{h\sqrt{i}}\int (K(z))^2 \left(f_{(i)}(x) + o(1)\right),dz \\
&= \frac{1}{h\sqrt{i}} R(K) f_{(i)}(x) + o\left(\frac{1}{h\sqrt{i}}\right) \qquad \text{(II.4)}
\end{aligned}$$

The notation $B_{n,p}$ will be used to denote a random variable that follows a Binomial(n,p) distribution. To finish work on the first sum, the following can be used.

$$\begin{aligned}
\sum_{i=1}^{n}\frac{1}{\sqrt{i}} f_{(i)}(x) &= \sum_{i=1}^{n}\frac{1}{\sqrt{i}}\frac{n!}{(i-1)!\,(n-i)!}\left[F(x)\right]^{i-1}\left[1-F(x)\right]^{n-i} f(x) \\
&= \frac{f(x)}{F(x)}\sum_{i=0}^{n}\sqrt{i}\frac{n!}{(i)!\,(n-i)!}\left[F(x)\right]^{i}\left[1-F(x)\right]^{n-i} \\
&= \frac{f(x)}{F(x)}\mathrm{E}\left[\sqrt{B_{n,F(x)}}\right] \qquad \text{(II.5)}
\end{aligned}$$

As for the little-o term in (II.4), note that

$$\sum_{i=1}^{n}\frac{1}{\sqrt{i}} = o(n),$$

though sharper bounds may be possible for the little-o error. It may be possible to use

$o(\sqrt{n})$. This gives the following for the first sum in (II.3).

$$\frac{1}{n^2h^2}\sum_{i=1}^{n}\frac{1}{i}\operatorname{E}\left[\left(K\left(\frac{x-X_{(i)}}{\sqrt{i}h}\right)\right)^2\right] = \frac{R(K)}{n^2h}\frac{f(x)}{F(x)}\operatorname{E}\left[\sqrt{B_{n,F(x)}}\right]+o\left(\frac{1}{nh}\right) \tag{II.6}$$

$$\leq \frac{R(K)}{n^2h}\frac{f(x)}{F(x)}\sqrt{\operatorname{E}\left[B_{n,F(x)}\right]}+o\left(\frac{1}{nh}\right)$$

$$= \frac{R(K)}{n^{\frac{3}{2}}h}\frac{f(x)}{\sqrt{F(x)}}+o\left(\frac{1}{nh}\right) \tag{II.7}$$

Where the inequality leading to (II.7) follows from application of Jensen's inequality.

Next, work on the second sum of (II.3) will be presented. This will require the use of the joint distribution of two order statistics. The joint density of two order statistics $X_{(i)}$ and $X_{(j)}$, $i<j$, is

$$f_{(i,j)}(s,t) = \frac{n!}{(i-1)!\,(j-i-1)!\,(n-j)!}[F(s)]^{i-1}[F(t)-F(s)]^{j-i-1}[1-F(t)]^{n-j}f(s)f(t)$$

where $s<t$ and 0 otherwise.

The next expectation is given by the following.

$$\begin{aligned}
&\frac{1}{h^2\sqrt{ij}}\operatorname{E}\left[K\left(\frac{x-X_{(i)}}{\sqrt{i}h}\right)K\left(\frac{x-X_{(j)}}{\sqrt{j}h}\right)\right]\\
&=\frac{1}{h^2\sqrt{ij}}\int\int_{-\infty}^{t}K\left(\frac{x-s}{h\sqrt{i}}\right)K\left(\frac{x-t}{h\sqrt{j}}\right)f_{(i,j)}(s,t)\,ds\,dt\\
&=\int\int_{-\infty}^{v}K(u)K(v)f_{(i,j)}(x-uh\sqrt{i},x-vh\sqrt{j})\,du\,dv\\
&=\int\int_{-\infty}^{v}K(u)K(v)\left[f_{(i,j)}(x,x)-uh\sqrt{i}\frac{\partial}{\partial s}f_{(i,j)}(s,t)\Big|_{s=t=x}-vh\sqrt{j}\frac{\partial}{\partial t}f_{(i,j)}(s,t)\Big|_{s=t=x}\right.\\
&\qquad\left.+o\left(h(\sqrt{i}+\sqrt{j})\right)\right]du\,dv
\end{aligned} \tag{II.8}$$

Next, let $W(u)$ be the CDF of the distribution with density $K()$ and let U and $V\sim W(u)$

$$\begin{aligned}
\int\int_{-\infty}^{v}uK(u)K(v)\,du\,db &= \int uK(u)\int_{u}^{\infty}K(v)\,dv\,du\\
&=\int uK(u)(1-W(u))du
\end{aligned}$$

$$
\begin{aligned}
&= -\mathrm{E}\left[UW(U)\right] \\
&=: -\gamma
\end{aligned}
\tag{II.9}
$$

This is a probability weighted moment (PWM), see Greenwood et al. (1979), which has general form $\mathrm{E}\left[X^i F(x)^j (1-F(x))^k\right]$ for a distribution with CDF $F(x)$. Here the notation γ will be used to denote the PWM with $i=1, j=1, k=0$ for the distribution with density given by the kernel K.

Next, in a similar manner as above,

$$
\begin{aligned}
\int\int_{-\infty}^{v} vK(u)K(v)\,dv\,du &= \int vK(v)W(v)du \\
&= \mathrm{E}\left[VW(V)\right] \\
&= \gamma
\end{aligned}
\tag{II.10}
$$

One last item to note from (II.8) is that

$$
\begin{aligned}
\int\int_{-\infty}^{v} K(u)K(v)\,du\,dv &= \int W(v)K(v)dv \\
&= E[W(V)] \\
&= \frac{1}{2}
\end{aligned}
\tag{II.11}
$$

since the expected value of any continuous CDF is $\frac{1}{2}$ by virtue of the probability integral transform. Using (II.9), (II.10), and (II.11), one can obtain the following for (II.8).

$$
\frac{f_{(i,j)}(x,x)}{2} + \gamma h \left[\sqrt{i}\frac{\partial}{\partial s} f_{(i,j)}(s,t)\bigg|_{s=t=x} - \sqrt{j}\frac{\partial}{\partial t} f_{(i,j)}(s,t)\bigg|_{s=t=x}\right] + o(h\sqrt{i} + \sqrt{j}) \tag{II.12}
$$

The double-sum of (II.12) is taken over $i < j$, j goes to n. Of particular interest is what happens to the joint order statistics density when this sum is taken. Work from Park et al.

2013, Theorem 3.5 gives the following result.

$$\frac{1}{n(n-1)}\sum_{j=2}^{n}\sum_{i=1}^{j-1} f_{(i,j)}(s,t) = f(s)f(t)u(t-s)$$

where

$$u(x) = \begin{cases} 1, & x > 0 \\ \frac{1}{2}, & x = 0 \\ 0, & x < 0 \end{cases}$$

Also, note that $\sum_{i=1}^{n}\sqrt{i} \leq n^{3/2}$, thus taking the sum of (II.12) one can obtain

$$\begin{aligned}
&\sum\sum_{i<j} \frac{2}{n^2.h^2\sqrt{ij}}\,\mathrm{E}\left[K\left(\frac{x-X_{(i)}}{\sqrt{i}h}\right)K\left(\frac{x-X_{(j)}}{\sqrt{j}h}\right)\right] \\
&= \sum\sum_{i<j}\frac{2}{n^2}\left(\frac{f_{(i,j)}(x,x)}{2} + \gamma h\left[\sqrt{i}\frac{\partial}{\partial s}f_{(i,j)}(s,t)\bigg|_{s=t=x} - \sqrt{j}\frac{\partial}{\partial t}f_{(i,j)}(s,t)\bigg|_{s=t=x}\right]\right. \\
&\qquad\qquad \left. + o(h\sqrt{i}+\sqrt{j})\right) \\
&= \frac{n-1}{n}f(x)^2 + o(\sqrt{n}h). \qquad\qquad \text{(II.13)}
\end{aligned}$$

Here, the assumption that $nh^2 \to 0$, which is established in (II.1), is used. This gives us the form for the second sum in (II.3). Lastly, the final sum will be considered.

Note that from (II.1) when calculating the bias, one can obtain

$$\begin{aligned}
&\sum_{i=1}^{n}\sum_{j=1}^{n}\frac{1}{n^2\sqrt{ij}}\,\mathrm{E}\left[K\left(\frac{x-X_{(i)}}{\sqrt{i}h}\right)\right]\mathrm{E}\left[K\left(\frac{x-X_{(j)}}{\sqrt{j}h}\right)\right] \\
&= \left(\frac{1}{nh}\sum_{i=1}^{n}\frac{1}{\sqrt{i}}\,\mathrm{E}\left[K\left(\frac{x-X_{(i)}}{\sqrt{i}h}\right)\right]\right)\left(\frac{1}{nh}\sum_{j=1}^{n}\frac{1}{\sqrt{j}}\,\mathrm{E}\left[K\left(\frac{x-X_{(j)}}{\sqrt{j}h}\right)\right]\right) \\
&= f(x)^2 + o(nh^2) \qquad\qquad \text{(II.14)}
\end{aligned}$$

Finally, a form for the variance can be given. Substituting (II.7), (II.13), and (II.14) into

(II.1) gives

$$
\begin{aligned}
Var(\hat{f}_h(x)) &= \frac{R(K)}{n^2 h}\frac{f(x)}{F(x)}\mathrm{E}\left[\sqrt{B_{n,F(x)}}\right] + \frac{n-1}{n} f(x)^2 - f(x)^2 + o(\sqrt{n}h) + o\left(n^{-1}h^{-1}\right) + o(nh^2) \\
&= \frac{R(K)}{n^2 h}\frac{f(x)}{F(x)}\mathrm{E}\left[\sqrt{B_{n,F(x)}}\right] - \frac{1}{n} f(x)^2 + o(\sqrt{n}h) \\
&= \frac{R(K)}{n^2 h}\frac{f(x)}{F(x)}\mathrm{E}\left[\sqrt{B_{n,F(x)}}\right] + o(\sqrt{n}h) \qquad \text{(II.15)} \\
&= O(n^{-2}h^{-1}) \\
&\to 0.
\end{aligned}
$$

$\square$

Using the results of Theorems 2.1.1 and 2.1.2 the following corollary results.

Corollary 2.1.1**.** *The MSE of* $\hat{f}_h(x)$ *is*

$$
MSE(\hat{f}_h(x)) = \frac{R(K)}{n^2 h}\frac{f(x)}{F(x)}\mathrm{E}\left[\sqrt{B_{n,F(x)}}\right] + \frac{\sigma_k^4 h^4}{2}\left[f^{''}(x) + (n-1)\left(f^{''}(x)F(x) + 3f(x)f^{'}(x)\right)\right]^2 + o(\sqrt{n}h)
$$

And the MISE of $\hat{f}_h(x)$ *is*

$$
\begin{aligned}
MISE(\hat{f}_h(\cdot)) = \frac{R(K)}{n^2 h}\int \frac{f(x)}{F(x)}\sum_{i=1}^{n}\sqrt{i}\binom{n}{i}[F(x)]^i[1-F(x)]^{n-i}\,dx \\
+ \frac{\sigma_k^4 h^4}{2} R\left(f^{''}(x) + (n-1)\left(f^{''}(x)F(x) + 3f(x)f^{'}(x)\right)\right) + o(\sqrt{n}h)
\end{aligned}
\qquad \text{(II.16)}
$$

Proof. Combine (II.21) and (II.15) to get an expression for the Mean Squared Error (MSE) of the estimator.

Then the MSE for the estimator at any given point x is

$$
\begin{aligned}
MSE(\hat{f}_h(x)) &= \mathrm{Var}(\hat{f}_h(x)) + Bias(\hat{f}_h(x))^2 \\
&= \frac{R(K)}{n^2 h}\frac{f(x)}{F(x)}\mathrm{E}\left[\sqrt{B_{n,F(x)}}\right] + \frac{\sigma_k^4 h^4}{2}\left[f^{''}(x) + (n-1)\left(f^{''}(x)F(x) + 3f(x)f^{'}(x)\right)\right]^2 + o(\sqrt{n}h)
\end{aligned}
$$

Integrating out x and expanding the expectation to obtain the Mean Integrated Squared Error(MISE) gives the expression

$$MISE(\hat{f}_h(\cdot)) = \frac{R(K)}{n^2h}\int \frac{f(x)}{F(x)} \sum_{i=1}^{n} \sqrt{i}\binom{n}{i}[F(x)]^i[1-F(x)]^{n-i}\,dx \\ + \frac{\sigma_k^4h^4}{2}R\left(f^{''}(x)+(n-1)\left(f^{''}(x)F(x)+3f(x)f^{'}(x)\right)\right)+o(\sqrt{n}h). \tag{II.17}$$

The integral in the first term can be simplified.

$$\begin{aligned}\int \frac{f(x)}{F(x)} \sum_{i=1}^{n} \sqrt{i}\binom{n}{i}[F(x)]^i[1-F(x)]^{n-i}\,dx &= \sum_{i=1}^{n}\sqrt{i}\int \frac{n!}{i!\,(n-i)!}[F(x)]^{i-1}[1-F(x)]^{n-i}f(x)\,dx \\ &= \sum_{i=1}^{n}\frac{1}{\sqrt{i}}\int \frac{n!}{(i-1)!\,(n-i)!}[F(x)]^{i-1}[1-F(x)]^{n-i}f(x)\,dx \\ &= \sum_{i=1}^{n}\frac{1}{\sqrt{i}}\int f_{(i)}(x)\,dx \\ &= \sum_{i=1}^{n}\frac{1}{\sqrt{i}}\end{aligned}$$

Then the expression for the Approximate MISE (AMISE) is

$$AMISE(\hat{f}_h(\cdot)) = \frac{R(K)}{n^2h}\sum_{i=1}^{n}\frac{1}{\sqrt{i}} + \frac{\sigma_k^4h^4}{2}R\left(f^{''}(x)+(n-1)\left(f^{''}(x)F(x)+3f(x)f^{'}(x)\right)\right) \tag{II.18}$$

□

In the same way that the classical optimal bandwidth is found in Wand and Jones (1994) (p. 22), one can find the AMISE optimal bandwidth via calculus. Taking the first derivative of (II.18) with respect to h,

$$\frac{d}{dh}AMISE(\hat{f}_h(\cdot)) = 2\sigma_k^4h^3R\left(f^{''}(x)+(n-1)\left(f^{''}(x)F(x)+3f(x)f^{'}(x)\right)\right) - \frac{R(K)}{n^2h^2}\sum_{i=1}^{n}\frac{1}{\sqrt{i}}.$$

Setting this equal to zero when evaluated at h_{opt} gives

$$0 = 2\sigma_k^4 h_{opt}^3 R\left(f^{''}(x) + (n-1)\left(f^{''}(x)F(x) + 3f(x)f^{'}(x)\right)\right) - \frac{R(K)}{n^2 h_{opt}^2} \sum_{i=1}^{n} \frac{1}{\sqrt{i}},$$

which leads to the solution

$$h_{opt} = \left[\frac{R(K)\sum_{i=1}^{n}\frac{1}{\sqrt{i}}}{2n^2\sigma_k^4 R\left(f''(x) + (n-1)\left(f''(x)F(x) + 3f(x)f'(x)\right)\right)}\right]^{1/5}. \tag{II.19}$$

2.2 Asymptotic Comparison to the Standard Density Estimator

The intention for this estimator is to reduce bias in estimating the density function of a continuous right skewed distribution. Note that

$$Bias(\hat{f}_{Std}(x)) = \frac{\sigma_K^2 h_{Std}^2}{2} f''(x) \tag{II.20}$$

for the standard estimator.

The rank weighted estimator's bias asymptotic form is given by

$$Bias(\hat{f}_{RW}(x)) = \frac{\sigma_K^2 h_{RW}^2}{2}\left[f^{''}(x) + (n-1)\left(f^{''}(x)F(x) + 3f(x)f^{'}(x)\right)\right]. \tag{II.21}$$

Define

$$S(f) = f^{''}(x) + (n-1)\left(f^{''}(x)F(x) + 3f(x)f^{'}(x)\right),$$

so that the bias for the rank weighted estimator can be represented in (II.21) in a similar form to that of the standard estimator in (II.20).

$$Bias(\hat{f}_{RW}(x)) = \frac{\sigma_K^2 h_{RW}^2}{2} S(f). \tag{II.22}$$

Note that the following ratio is of interest.

$$\begin{aligned}
\frac{Bias(\hat{f}_{Std}(x))}{Bias(\hat{f}_{RW}(x))} &= \frac{\frac{\sigma_K^2 h_{Std}^2}{2} f''(x)}{\frac{\sigma_K^2 h_{RW}^2}{2} S(f)} \\
&= \frac{h_{Std}^2 f''(x)}{h_{RW}^2 S(f)}.
\end{aligned}$$

If the bias of the rank weighted estimator is smaller, than this ratio should be greater than one.

This ratio has two important parts:

1. $\dfrac{h_{Std}^2}{h_{RW}^2}$
2. $\dfrac{f''(x)}{S(f)}$.

Both of these ratios actually represent quite some difficulty to analyze properly since f is typically unknown. For the first ratio, substitute in the form for the AMISE optimal bandwidth of the estimators,

$$\begin{aligned}
h_{Std} &= \left(\frac{R(K)}{nR(f'')}\right)^{1/5}, \text{ and} \\
h_{RW} &= \left(\frac{R(K)\sum_{i=1}^{n}\frac{1}{\sqrt{i}}}{2n^2 R(S(f))}\right)^{1/5}.
\end{aligned}$$

Then the ratio of the bandwidths is

$$\frac{h_{Std}^2}{h_{RW}^2} = \left(\frac{2nR(S(f))}{R(f'')\sum_{i=1}^{n}\frac{1}{\sqrt{i}}}\right)^{2/5}.$$

Now, $S(f) = O(n)$ which implies $R(S(f)) = O(n^2)$ since it is the integral of the squared function. Additionally, the sum of the inverse square roots of the ranks is of order $O(\sqrt{n})$

by integral approximation. Then

$$\frac{h_{Std}^2}{h_{RW}^2} = \left(O(\frac{n^3}{\sqrt{n}})\right)^{2/5}$$
$$= O(n),$$

and so,

$$\frac{Bias(\hat{f}_{Std}(x))}{Bias(\hat{f}_{RW}(x))} = O(n) \cdot \frac{f''(x)}{S(f)}$$
$$= O(n) \cdot O(n^{-1})$$
$$= O(1).$$

This indicates that the asymptotic order of the biases could be equivalent up to a constant as long bandwidth selectors are consistent for choosing a value close to the AMISE optimal h. However it does not indicate which estimator is better than the other.

Next, take a look at the second ratio, $f''(x)/S(f)$, and expand $S(f)$.

$$\frac{f''(x)}{S(f)} = \frac{f''(x)}{f''(x) + (n-1)\left(f''(x)F(x) + 3f(x)f'(x)\right)}$$

Consider a right skewed distribution like that depicted in Figure 2.1. At values of x beyond the mode of the distribution, some generalizations can be made about the structure of the distribution in terms of the signs of f, f', f'' and F.

The second derivative of f can be negative as the slope is getting steeper which makes $f''(x)F(x)$ negative. Also, the slope is negative, thus $f'(x)$ is negative, which indicates that $f'(x)f(x)$ is negative. This means that

$$|f''(x) + (n-1)\left(f''(x)F(x) + 3f(x)f'(x)\right)| > |f''(x)|.$$

This does not help as it points to a potentially larger bias for the rank weighted estimator, but this would depend on the value of the ratio of the squared bandwidths.

Further consideration on the structure of skewed right distributions may help the rank weighted estimator. They stereotypically have elongated right tails that asymptotically approach the horizontal axis. This means that while the slope is negative, it is slowly increasing, indicating a positive second derivative. It may be possible to characterize skewed distributions by when the following conditions are true:

1. $f''(x) > 0$,
2. $|(n-1)(f''(x)F(x)+3f(x)f'(x))| < f''(x)$, and
3. $(n-1)(f''(x)F(x)+3f(x)f'(x)) < 0$.

If these condtions are true, the result would be

$$|f''(x)+(n-1)(f''(x)F(x)+3f(x)f'(x))| < |f''(x)|,$$

which would give

$$\left|\frac{f''(x)}{f''(x)+(n-1)(f''(x)F(x)+3f(x)f'(x))}\right| > 1.$$

Also, in general $\frac{h^2_{Std}}{h^2_{RW}}$ should be greater than 1 since the ratio is of order n. Thus, the bias ratio

$$\frac{Bias(\hat{f}_{Std}(x))}{Bias(\hat{f}_{RW}(x))} = \frac{h^2_{Std} f''(x)}{h^2_{RW} S(f)}$$

should be greater than 1 in this special circumstance. Currently, it is unknown when this happens for skewed distributions, but at the very least there is an opportunity.

Unfortunately, this look does not take into account the fact that $h^2_{Std}/h^2_{RW} = O(n)$. Suppose that $h^2_{Std}/h^2_{RW} = cn$ for some positive value c. Then the result is

$$\frac{Bias(\hat{f}_{Std}(x))}{Bias(\hat{f}_{RW}(x))} = \frac{cnf''(x)}{f''(x)+(n-1)(f''(x)F(x)+f(x)f'(x))}.$$

By rearranging the denominator, the outcome is

$$\frac{cnf''(x)}{n\{f''(x)F(x)+f'(x)f(x)\}-f''(x)[1-F(x)]-f(x)f'(x)}$$
$$=\frac{cf''(x)}{f''(x)F(x)+f'(x)f(x)-n^{-1}\{f''(x)[1-F(x)]+f(x)f'(x)\}}.$$

The main interest is in the absolute value of this ratio, therefore examination of the following is in order,

$$\frac{c|f''(x)|}{|f''(x)F(x)+f'(x)f(x)-n^{-1}\{f''(x)[1-F(x)]+f(x)f'(x)\}|}.$$

Take n to be large such that the n^{-1} term is negligible in the denominator; this seems reasonable given that this is all taken from the asymptotic bias. The main issue is figuring out when the following is true:

$$c|f''(x)| > |f''(x)F(x)+f'(x)f(x)|.$$

Right skewed distributions are the assumed structure for $f(x)$, so $f'(x)f(x)$ should be negative for a large portion of the distribution, given that there should be a negative slope over much of the density. The sign of $f''(x)$ could be positive or negative.

Suppose $f''(x)$ is positive, i.e., past the right hand inflection point in the distribution. Then $|f''(x)F(x)+f'(x)f(x)| < |f''(x)F(x)|$ and so, to get the desired outcome a value of $F(x) < c$ is needed . This will occur at some point as long as the second derivative is positive at a small enough value of x, since c is positive. This seems plausible for a right skewed distribution with a fairly stretched out right tail. In fact, when the second derivative is positive and x is a value located in the right tail with a negative first derivative, there is room for favorable conditions because $|f''(x)F(x)+f'(x)f(x)| < |f''(x)F(x)|$.

2.3 Bandwidth Selection

Though a form for h is given by (II.19), one of the issues that needs to be dealt with is the fact that the denominator requires $F(x)$, $f(x)$, $f'(x)$, and $f''(x)$. This is a major issue given that $f(x)$ unknown and is the objective of estimation. This section will concentrate on how to approach this issue by adopting methods discussed for bandwidth selection in the standard estimator in chapter 1: Silverman's Rule of Thumb Silverman 1986, least squares cross validation (LSCV), and plug-in. Additionally, the possibility of an bandwidth selection method that serves as a bridge between the rule-of-thumb and plug-in estimators is discussed. Results of these methods will be examined via simulations.

The classic rule-of-thumb bandwidth selector, attributed to Silverman (1986), in the standard density estimation case is based on estimating $R(f'')$ by substituting the unknown second derivative of the target density with the second derivative of the normal distribution. The rule of thumb in the rank-weighted case will take a similar approach.

First, start with the AMISE optimal bandwidth,

$$h_{opt} = \left[\frac{R(K) \sum_{i=1}^{n} \frac{1}{\sqrt{i}}}{2n^2 \sigma_T^4(f)} \right]^{1/5}.$$

Then expand,

$$\begin{aligned} T(f) = R\left(f''(x)\right) + (n-1)^2 R\left(f''(x)F(x) + 3f(x)f'(x)\right) \\ + 2(n-1) \int f''(x)(f''(x)F(x) + 3f(x)f'(x))dx. \end{aligned} \tag{II.23}$$

Any normal distribution with a mean of 0, standard deviation σ, and density $f(x)$ can be rewritten in terms of the standard normal density $f(x) = (1/\sigma)\phi(z)$ where $z = x/\sigma$. Working this into the form for $T(f)$ and assuming f is a normal density with mean 0 results in

$$
\begin{aligned}
T(f) &= \frac{1}{\sigma^5} T(\phi) \\
&= \frac{1}{\sigma^5} \left\{ R\left(\phi^{''}(z)\right) + (n-1)^2 R\left(\phi^{''}(z)\Phi(z) + 3\phi(z)\phi^{'}(z)\right) \right. \\
&\qquad \left. + 2(n-1) \int \phi^{''}(z)(\phi^{''}(z)\Phi(z) + 3\phi(z)\phi^{'}(z))dz \right\}.
\end{aligned}
$$

The integrals needed for the R terms in $T(\phi)$ are solved numerically. The final step would be approximating σ in some manner such as using the sample standard deviation. Denote this approximation by $\hat{\sigma}$. Then, the rule-of-thumb bandwidth for the rank-weighted kernel is

$$
h_{ROT} = \hat{\sigma} \left[\frac{R(K) \sum_{i=1}^{n} \frac{1}{\sqrt{i}}}{2n^2 \sigma_k^4 T(\phi)} \right]^{1/5}. \tag{II.24}
$$

Least Squares Cross Validation as shown by Bowman (1984) revolves around the integrated squared error (ISE) criterion,

$$
\begin{aligned}
ISE(\hat{f}_h(x)) &= \int \left(\hat{f}_h(x) - f(x)\right)^2 dx \\
&= \int \hat{f}_h(x)^2 \, dx - \int \hat{f}_h(x) f(x) \, dx + \int f(x)^2 \, dx.
\end{aligned} \tag{II.25}
$$

The objective is to find an optimal h based on this so the right-hand term can be ignored. First, the left-hand term will be examined. This term can be simplified when assuming K is the Gaussian kernel. To see this, look at the expanded form:

$$
\begin{aligned}
\int \hat{f}_h(x)^2 \, dx &= \int \left(\frac{1}{nh} \sum_{i=1}^{n} \frac{1}{\sqrt{i}} K\left(\frac{x - X_{(i)}}{\sqrt{i}h} \right) \right)^2 dx \\
&= \int \frac{1}{n^2 h^2} \sum_{i=1}^{n} \sum_{j=1}^{n} \frac{1}{\sqrt{ij}} K\left(\frac{x - X_{(i)}}{\sqrt{i}h} \right) K\left(\frac{x - X_{(j)}}{\sqrt{j}h} \right) dx \\
&= \frac{1}{n^2} \sum_{i=1}^{n} \sum_{j=1}^{n} \int \frac{1}{\sqrt{i}h} K\left(\frac{x - X_{(i)}}{\sqrt{i}h} \right) \frac{1}{\sqrt{j}h} K\left(\frac{x - X_{(j)}}{\sqrt{j}h} \right) dx.
\end{aligned} \tag{II.26}
$$

Now, let K be the Gaussian kernel ϕ. The result is,

$$\int \hat{f}_h(x)^2 = \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\int \frac{1}{\sqrt{i}h}\phi\left(\frac{x-X_{(i)}}{\sqrt{i}h}\right)\frac{1}{\sqrt{j}h}\phi\left(\frac{x-X_{(j)}}{\sqrt{j}h}\right)dx. \tag{II.27}$$

Inside the integral is the product of two Gaussian pdfs. One with a mean of $X_{(i)}$ and standard deviation $\sqrt{i}h$, and the other with $X_{(j)}$ and standard deviation $\sqrt{j}h$.

The product of two Gaussian pdfs with means μ_1, μ_2 and standard deviations σ_1, σ_2 is

$$\frac{1}{\sqrt{\sigma_1^2+\sigma_2^2}}\phi\left(\frac{\mu_1-\mu_2}{\sqrt{\sigma_1^2+\sigma_2^2}}\right)\frac{1}{\sqrt{\frac{\sigma_1^2\sigma_2^2}{\sigma_1^2+\sigma_2^2}}}\phi\left(\frac{x-\frac{\mu_1\sigma_1^2+\mu_2\sigma_2^2}{\sigma_1^2+\sigma_2^2}}{\sqrt{\frac{\sigma_1^2\sigma_2^2}{\sigma_1^2+\sigma_2^2}}}\right).$$

The right side of this equation is a Gaussian pdf, so integrating the whole equation over x gives a value of

$$\frac{1}{\sqrt{\sigma_1^2+\sigma_2^2}}\phi\left(\frac{\mu_1-\mu_2}{\sqrt{\sigma_1^2+\sigma_2^2}}\right).$$

Therefore, the integral in (II.27) is,

$$\begin{aligned}\int \hat{f}_h(x)^2\,dx &= \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\int \frac{1}{\sqrt{i}h}\phi\left(\frac{x-X_{(i)}}{\sqrt{i}h}\right)\frac{1}{\sqrt{j}h}\phi\left(\frac{x-X_{(j)}}{\sqrt{j}h}\right)dx\\ &= \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\frac{1}{\sqrt{ih^2+jh^2}}\phi\left(\frac{X_{(i)}-X_{(j)}}{\sqrt{ih^2+jh^2}}\right). \end{aligned}\tag{II.28}$$

The next term in (II.25) is

$$2\int \hat{f}_h(x)f(x)\,dx = 2\,\mathrm{E}\left[\hat{f}_h(X)\right].$$

But, $f(x)$ is unknown so the expected value must be estimated. The estimator for $\mathrm{E}\left[\hat{f}_h(X)\right]$ is

$$\frac{1}{n}\sum_{i=1}^{n}\hat{f}_{-i}(X_i),$$

where

$$\hat{f}_{-i}(x) = \frac{1}{(n-1)h} \sum_{j=1, j\neq i}^{n} \frac{1}{\sqrt{i}} K\left(\frac{x - X_{(j)}}{h}\right),$$

which is an adaptation of the standard leave-one-out estimator for the rank-weighted case. Then, the estimate for $2\,\mathrm{E}\left[\hat{f}_h(X)\right]$ is

$$\frac{2}{n(n-1)h} \sum_{i=1}^{n} \sum_{j\neq i} \frac{1}{\sqrt{j}} K\left(\frac{X_{(i)} - X_{(j)}}{\sqrt{j}h}\right) = \frac{4}{n(n-1)h} \sum\sum_{i<j} \frac{1}{\sqrt{j}} K\left(\frac{X_{(i)} - X_{(j)}}{\sqrt{j}h}\right),$$

where the right-hand side comes from assuming a symmetric kernel. The above equation uses the analog of the leave-one-out estimator discussed by Bowman (1984). Here, the i^{th} rank is simply deleted but all other values retain their ranks from the whole sample. Bowman's original work created an unbiased estimator of $2\,\mathrm{E}\left[\int \hat{f}_h(x) f(x) dx\right]$. That result relied on the independence of the X_i values, which is not an available assumption here given the dependent structure of order statistics. Thus, an estimator the first two terms in (II.25) is

$$\begin{aligned} LSCV(h) = \frac{1}{n^2} \sum_{i=1}^{n} \sum_{j=1}^{n} \int \frac{1}{\sqrt{i}h} K\left(\frac{x - X_{(i)}}{\sqrt{i}h}\right) \frac{1}{\sqrt{j}h} K\left(\frac{x - X_{(j)}}{\sqrt{j}h}\right) \\ - \frac{4}{n(n-1)h} \sum\sum_{i<j} \frac{1}{\sqrt{j}} K\left(\frac{X_{(i)} - X_{(j)}}{\sqrt{j}h}\right). \end{aligned} \tag{II.29}$$

Then the optimum h is found by minimizing (II.29):

$$\begin{aligned} h_{LSCV} = \arg\min_h \Bigg\{ \frac{1}{n^2} \sum_{i=1}^{n} \sum_{j=1}^{n} \int \frac{1}{\sqrt{i}h} K\left(\frac{x - X_{(i)}}{\sqrt{i}h}\right) \frac{1}{\sqrt{j}h} K\left(\frac{x - X_{(j)}}{\sqrt{j}h}\right) \\ - \frac{4}{n(n-1)h} \sum\sum_{i<j} \frac{1}{\sqrt{j}} K\left(\frac{X_{(i)} - X_{(j)}}{\sqrt{j}h}\right) \Bigg\}. \end{aligned} \tag{II.30}$$

For a general kernel, not much can be done to simplify this. However, if one were to make a specific choice of kernel, one can show some simplification. For example, if a

Gaussian kernel were used, (II.30) would become the following,

$$h_{LSCV} = \arg\min_{h} \left\{ \frac{1}{n^2} \sum_{i=1}^{n} \sum_{j=1}^{n} \frac{1}{\sqrt{ih^2 + jh^2}} \phi \left(\frac{X_{(i)} - X_{(j)}}{\sqrt{ih^2 + jh^2}} \right) \right.$$
$$\left. - \frac{4}{n(n-1)h} \sum\sum_{i<j} \frac{1}{\sqrt{j}} \phi \left(\frac{X_{(i)} - X_{(j)}}{\sqrt{j}h} \right) \right\}. \tag{II.31}$$

The plug-in estimator in the standard density estimation case was centered around estimating $R(f'')$ which is in the denominator of the AMISE optimal h. This provided a basic framework for how to deal with the $T(f)$ term in the rank-weighted estimator. Again, the $T(f)$ term is:

$$T(f) = R\left(f^{''}(x) + (n-1)\left(f^{''}(x)F(x) + 3f(x)f^{'}(x)\right)\right).$$

There are special techniques for estimating $R(f'')$ that cannot be utilized in estimating $T(f)$. Estimation of T will be based on estimating all of the individual derivatives of the density f, and the CDF F,

$$\widehat{T(f)} = R\left(\widehat{f}''(x) + (n-1)\left(\widehat{f''(x)}\widehat{F}(x) + 3\widehat{f}(x)\widehat{f}'(x)\right)\right).$$

Work by Rao (2014) gives the following estimator form:

$$\widehat{f^{(p)}}(x) = \frac{1}{nh^{p+1}} \sum_{i=1}^{n} K^{(p)}\left(\frac{x - X_i}{h}\right). \tag{II.32}$$

The CDF, however, must also be estimated. A simple estimate is the classical empirical distribution function:

$$\widehat{F}(x) = \frac{1}{n} \sum_{i=1}^{n} \mathrm{I}(X_i \le x).$$

A major issue estimating $f^{(p)}$ for $p = 0, 1, 2$, is that for each value of p a different bandwidth would be considered optimal. Scott (1992) shows that the AMISE optimal bandwidth for a density derivative estimator of the form given in (II.32) is

$$h_{(p)} = \left[\frac{(2r+1)R\left(K^{(p)}\right)}{n\sigma_K^4 R\left(f^{(p+2)}\right)} \right]^{1/(2r+5)},$$

which is going to run into the issue of what to use for $f^{(p)}$ in $R\left(f^{(p+2)}\right)$ of the denominator. Unfortunately, the literature seems sparse on the estimation of density derivatives. The latest significant work seems to be from Stoker (1993). In the end, the most straightforward way of resolving the issue of finding $h_{(p)}$ is through cross validation techniques:

$$\hat{h}_{(p)} = \arg\min_h \left\{ \frac{R(K^{(p)})}{nh^{2r+1}} + \frac{(-1)^r}{n(n-1)h^{2r+1}} \sum\sum_{i<j} \left(K^{(p)} * K^{(p)} - 2K^{(2p)}\right)\left(\frac{X_i - X_j}{h}\right) \right\}.$$

where $*$ denotes the convolution of two functions. This method was proposed by Hardle, Marron, and Wand (1990).

For the rank-weighted density estimator, $T(f)$ is estimating the derivatives of f, and f itself with this technique. The CDF is estimated using the empirical distribution function. All of these estimates are substituted into $T(f)$. The rank-weighted plug-in bandwidth is

$$h_{RW:Plug} = \left[\frac{R(K)\sum_{i=1}^{n} \frac{1}{\sqrt{i}}}{2n^2\sigma_k^4 R\left(\widehat{f''}(x) + (n-1)\left(\widehat{f''}(x)\widehat{F}(x) + 3\widehat{f}(x)\widehat{f'}(x)\right)\right)} \right]^{1/5}. \tag{II.33}$$

The final bandwidth choice algorithm used is based on taking the rule-of-thumb approach given by (II.24) and changing the distribution used for calculating $T(\phi)$. To that end, the distribution used is the skew normal distribution proposed by Azzalini (1985). The form of the pdf for the skew normal distribution Azzalini 1985 is given by

$$\psi(x;\xi,\omega,\alpha) = \frac{2}{\omega}\phi\left(\frac{x-\xi}{\omega}\right)\Phi\left(\alpha\left(\frac{x-\xi}{\omega}\right)\right),$$

where ϕ and Φ represent the standard normal pdf and CDF, respectively. The parameter α controls the skewness. If α is postive, the distribution is right skewed, and it is left skewed when α is negative. The magnitude of the skewness increases with the magnitude of α.

Instead of using the normal distribution in the calculation of h_{ROT}, the skew normal distribution is used in its place. Then the bandwidth selection is

$$h^*_{SN} = \left[\frac{R(K) \sum_{i=1}^{n} \frac{1}{\sqrt{i}}}{2n^2 \sigma_k^4 T\left(\psi(\cdot;\xi,\omega,\alpha)\right)} \right]^{1/5}.$$

The parameters for the skew normal distribution must be given in order for this to be utilized. The parameters are therefore estimated using maximum likelihood estimation (MLE). Solutions to the MLE equations are not of closed form so numerical techniques must be used. These parameter estimates are then substituted into the above equation, giving the bandwidth selector:

$$h^*_{SN} = \left[\frac{R(K) \sum_{i=1}^{n} \frac{1}{\sqrt{i}}}{2n^2 \sigma_k^4 T\left(\psi(\cdot;\widehat{\xi},\widehat{\omega},\widehat{\alpha})\right)} \right]^{1/5}. \tag{II.34}$$

The motivation behind this methodology is that the rank-weighted estimator is intended to more accurately model skewed distributions. Therefore, a distribution with a skewness than can be selected seemed appropriate as a replacement for the normal distribution in the rule-of-thumb bandwidth selector.

2.4 Rank Weighted Transformation Kernel Density Estimator

For a density function with bounded support, e.g., $[0,\infty)$, the standard form for the density estimator can have issues with increased bias when estimating near the boundary. Work by Timsina (2015) seeks to correct this issue through a transformation kernel density estimator (TKDE):

$$\widehat{f}_T(x) = \frac{g'(x)}{nh} \sum_{i=1}^{n} K\left(\frac{g(x) - g(X_i)}{h}\right). \tag{II.35}$$

In this estimator, a function $g(x)$ is some known function used to transform data from a bounded interval to an unbounded interval so that it is more convenient to estimate the density of the transformed data.

As an adaptation of the rank weighting technique, the proposed estimator is the rank weighted transformation kernel density estimator (RWTKDE) and is given by

$$\widehat{f}_{RWT}(x) = \frac{g'(x)}{nh} \sum_{i=1}^{n} \frac{1}{\sqrt{i}} K\left(\frac{g(x) - g(X_i)}{h\sqrt{i}}\right) \tag{II.36}$$

An examination of asymptotic bias, variance, MSE, and MISE is currently unavailable. A preliminary simulation was done to assess the validity of this estimator. Given that asymptotic expressions are not available for the MISE, no form can be given for the optimal value for h. However, the techniques used by Timsina (2015) for bandwidth selection involve using least squares cross validation (LSCV) to find a bandwidth. These do not require a form for the optimal h and can therefore be used in its absence. Additionally, the work by Timsina concentrated specifically on the log transformation and using a Gaussian kernel. The comparison between TKDE and RWTKDE will be done using this transformation and kernel.

The LSCV bandwidth selector using $g(x) = log(x)$ and a Gaussian kernel given by Timsina is the following:

$$\begin{aligned} h_{TCV} = \arg\min_{h} \frac{1}{2n^2 h\sqrt{\pi}} \sum_{i=1}^{n} \sum_{j=1}^{n} \exp\left\{ \frac{1}{4h^2}\left(\log X_i + \log X_j - h^2\right) - \frac{(\log X_i)^2 + (\log X_j)^2}{2h^2} \right\} \\ - \frac{4}{n(n-1)h\sqrt{2\pi}} \sum\sum_{i<j} \frac{1}{X_i} \exp\left\{ - \frac{\left(\log X_i - \log X_j\right)^2}{2h^2} \right\}, \end{aligned}$$

which is the result of adapting the form of the standard LSCV selector given in (I.9).

In this same vein, the RWTKDE version of LSCV is given by

$$h_{RWTCV} = \arg\min_h \frac{1}{n^2 h\sqrt{2\pi}} \sum_{i=1}^{n}\sum_{j=1}^{n} \frac{1}{\sqrt{i+j}} \exp\left\{ -\frac{\left(d_{ij} + \frac{c_{ij} - ijh^2}{\sqrt{i+j}}\right)}{2ijh^2} \right\}$$
$$- \frac{4}{n(n-1)h\sqrt{2\pi}} \sum\sum_{i<j} \frac{1}{X_i\sqrt{j}} \exp\left\{ -\frac{\left(\log X_{(i)} - \log X_{(j)}\right)^2}{2jh^2} \right\},$$

where $c_{ij} = j\log X_{(i)} + i\log X_{(j)}$ and $d_{ij} = j(logX_{(i)})^2 + i(\log X_{(j)})^2$.

Using these forms for the bandwidth, a simulation study to determine the effectiveness of the RWTKD relative to the TKDE was conducted. Results are provided in chapter 4.

CHAPTER III

RANK WEIGHTED KERNEL REGRESSION

Next is a presentation of how to utilize both the observations and ranks in the kernel regression setting. The simplest case will be used where the bivariate relationship between two random variables, X and Y is examined. For a random sample $(X_1,Y_1),(X_2,Y_2),...,(X_n,Y_n)$ the general model being considered is

$$Y_i = g(X_i) + \varepsilon_i \tag{III.1}$$

where $g(\cdot)$ is a continuous function and the ε_i values are i.i.d. with $\mathrm{E}[\varepsilon_i] = 0$ and $Var(\varepsilon_i) = \sigma_\varepsilon^2$. Of interest is estimating the conditional expectation of Y given $X = x$. Denote the joint0 density of X and Y by $f(x,y)$ and the marginal density of X by $f_x(x)$.

$$\begin{aligned} g(x) &= \mathrm{E}[Y|X=x] \\ &= \frac{\int yf(x,y)dy}{\int f(x,y)dy} \\ &= \frac{\int yf(x,y)dy}{f_x(x)} \end{aligned} \tag{III.2}$$

3.1 Rank Weighted Nadaraya-Watson Estimator

Both Nadaraya (1964) and Watson (1964) proposed an estimator of the form

$$\hat{g}(x) = \frac{\sum_{i=1}^{n} K\left(\frac{X_i - x}{h}\right) Y_i}{\sum_{i=1}^{n} K\left(\frac{X_i - x}{h}\right)}.$$

The derivation of a rank weighted estimator will be similar to that of the Nadaraya-Watson estimator shown by Li and Racine (2007). The ranks of the X_i values will be denoted by R_i and the ranks of the Y_i values will be denoted by S_i. Then the bivariate rank weighted estimator is given by the following:

$$\hat{f}(x,y) = \frac{1}{nh^2} \sum_{i=1}^{n} \frac{1}{\sqrt{R_i S_i}} K\left(\frac{X_i - x}{\sqrt{R_i}h}\right) K\left(\frac{Y_i - y}{\sqrt{S_i}h}\right).$$

This leads to the rank weighted regression estimator, derived as follows:

$$\begin{aligned}
\hat{g}(x) &= \mathrm{E}\,[\widehat{Y|X = x}] \\
&= \frac{\displaystyle\int y \hat{f}(x,y) dy}{\displaystyle\int \hat{f}(x,y) dy} \\
&= \frac{\displaystyle\int y \frac{1}{nh^2} \sum_{i=1}^{n} \frac{1}{\sqrt{R_i S_i}} K\left(\frac{X_i - x}{\sqrt{R_i}h}\right) K\left(\frac{Y_i - y}{\sqrt{S_i}h}\right) dy}{\displaystyle\int \frac{1}{nh^2} \sum_{i=1}^{n} \frac{1}{\sqrt{R_i S_i}} K\left(\frac{X_i - x}{\sqrt{R_i}h}\right) K\left(\frac{Y_i - y}{\sqrt{S_i}h}\right) dy} \\
&= \frac{\displaystyle\sum_{i=1}^{n} \frac{1}{\sqrt{R_i}} K\left(\frac{X_i - x}{\sqrt{R_i}h}\right) \int y \frac{1}{\sqrt{S_i}} K\left(\frac{Y_i - y}{\sqrt{S_i}h}\right) dy}{\displaystyle\sum_{i=1}^{n} \frac{1}{\sqrt{R_i}} K\left(\frac{X_i - x}{\sqrt{R_i}h}\right) \int \frac{1}{\sqrt{S_i}} K\left(\frac{Y_i - y}{\sqrt{S_i}h}\right) dy} \\
&= \frac{\displaystyle\sum_{i=1}^{n} \frac{1}{\sqrt{R_i}} K\left(\frac{X_i - x}{\sqrt{R_i}h}\right) \int (Y_i + \sqrt{S_i} h v_i) K(v_i)\, dv_i}{\displaystyle\sum_{i=1}^{n} \frac{1}{\sqrt{R_i}} K\left(\frac{X_i - x}{\sqrt{R_i}h}\right) \int K(v_i)\, dv_i}
\end{aligned}$$

$$= \frac{\sum_{i=1}^{n} \frac{1}{\sqrt{R_i}} K\left(\frac{X_i - x}{\sqrt{R_i}h}\right) Y_i}{\sum_{i=1}^{n} \frac{1}{\sqrt{R_i}} K\left(\frac{X_i - x}{\sqrt{R_i}h}\right)}. \tag{III.3}$$

Another way of displaying this rank weighted Nadaraya-Watson estimator is in the following manner:

$$\hat{g}(x) = \frac{\sum_{i=1}^{n} \frac{1}{\sqrt{i}} K\left(\frac{X_{(i)} - x}{\sqrt{i}h}\right) Y_{[i]}}{\sum_{i=1}^{n} \frac{1}{\sqrt{i}} K\left(\frac{X_{(i)} - x}{\sqrt{i}h}\right)}. \tag{III.4}$$

Here, $X_{(i)}$ represents the reordered X values where i is the rank of these reordered values: $X_{(1)} \leq X_{(2)} \leq \ldots \leq X_{(n)}$. $Y_{[i]}$ denotes the resulting concomitant ordering of the Y by keeping the paired observations together: $(X_{(1)}, Y_{[1]}), (X_{(2)}, Y_{[2]}), \ldots, (X_{(n)}, Y_{[n]})$. This means that i represents the rank of $X_{(i)}$ of the reordered values but it does not necessarily give us information on the rank of $Y_{[i]}$.

Concomitant ordering is an important part of the literature in order statistics, see David and Nagaraja (2003). The main point is that depending on the structure of the relationship between X and Y, the order statistics of X may contain information about the order statistics of Y. The easiest scenario to visualize this would be in a linear relationship between X and Y, i.e., $Y = aX + b + \varepsilon$. If there is a positive association, the ranks of X correspond directly to the ranks of Y. If the association is negative, the ranks of X and Y should be opposing, i.e, as the rank i increases, $X_{(i)}$ increases while $Y_{[i]}$, the concomitant, decreases.

This leads to some intuition for why the rank weighted Nadaraya-Watson (RWNW) estimator should have situations where it performs better. At points where the regression function $g(x)$ is being estimated near low ranks of the $X_{(i)}$'s, the overall bandwidth $\sqrt{i}h$ will be smaller, allowing for more sensitivity to $g(x)$. At x near higher ranks of $X_{(i)}$, the value for $\sqrt{i}h$ will be larger, so the sensitivity will be decreased to changes in $g(x)$. Thus, the RWNW estimator may perform best when trying to estimate functions with fine structure at low values of x and less structure at larger values of x.

An extra point of discussion can be found by looking back at the ranks R_i and S_i used in deriving (III.3). R_i is the rank of X_i and S_i is the rank of Y_i. Perform a notation swap so that R_i is the rank of Y_i and S_i is the rank of X_i. The same steps in the derivation of (III.3) would apply, but now the rankings left are those of the response variable instead of those of the predictor variable. This could leave us with an estimator of the form

$$\hat{g}(x) = \frac{\sum_{i=1}^{n} \frac{1}{\sqrt{i}} K\left(\frac{X_{[i]} - x}{\sqrt{i}h}\right) Y_{(i)}}{\sum_{i=1}^{n} \frac{1}{\sqrt{i}} K\left(\frac{X_{[i]} - x}{\sqrt{i}h}\right)}, \tag{III.5}$$

where here i is the rank of the ordered Y values denoted by $Y_{(i)}$, and $X_{[i]}$ denotes the concomitant ordering of the X values.

There may be potential for this estimator, but an initial examination of its form may lead to the conclusion that its usefulness would be for a smaller subset of potential regression functions $g(x)$. The same idea applies as previously, but the now overall bandwidth $\sqrt{i}h$ is small at low values of y and larger at high values of y. This means that the fine structure would need to be mostly where there are low ranked values of Y. This is a different type of restriction from the rank by X case; as the estimator would be beneficial only if the shape of the function changed rapidly in the bottom half of the y-axis but had relatively calm behavior in the top half. Surely situations like this exist, but they represent a stronger restriction on the shape of $g(x)$. As of now, research into this idea is limited, and would be a strong consideration for future research.

3.2 Statistical Properties

As in the density estimation case, the asymptotic forms for the bias and variance will be given. The bias can be obtained in a form very similar to Li and Racine (2007). The derivation of the bias can be done by breaking the bias into a ratio,

$$\begin{aligned}\hat{g}(x) - g(x) &= \frac{(\hat{g}(x) - g(x))\hat{f}_h(x)}{\hat{f}_h(x)} \\ &=: \frac{\hat{m}(x)}{\hat{f}_h(x)}. \end{aligned} \tag{III.6}$$

Here, $\hat{f}_h(x)$ is the density estimate of $f_x(x)$ as shown in Chapter II. The reason the bias is given this way is show it in the form of an isolated ratio. The following lemma will be useful in the approximation of the bias. It is a fairly common result, but its proof is simple so it is presented here.

Lemma 3.2.1**.** *Let X, and Y be random variables where Y cannot be 0 and* $\mathrm{E}[X], \mathrm{E}[Y] < \infty$. *Let $h(x,y)$ be some continuous function with first order derivatives. Then*

$$\mathrm{E}[h(X,Y)] \approx h(\mathrm{E}[X], \mathrm{E}[Y]).$$

Proof. Denote the partial derivatives of h with respect to x and y by h_x and h_y, respectively. The first order Taylor expansion of h about $\theta = (\theta_1, \theta_2)$ is

$$h(x,y) = h(\theta) + h_x(\theta)(x - \theta_1) + h_y(\theta)(y - \theta_2) + R_{xy},$$

where R_{xy} is the remainder. Thus we get

$$\begin{aligned}\mathrm{E}[h(X,Y)] &= \mathrm{E}[h(\theta) + h_x(\theta)(X - \theta_1) + h_y(\theta)(Y - \theta_2) + R_{xy}] \\ &= h(\theta) + h_x(\theta)(\mathrm{E}[X] - \theta_1) + h_y(\theta)(\mathrm{E}[Y] - \theta_2) + R_{xy} \\ &\approx h(\theta) + h_x(\theta)(\mathrm{E}[X] - \theta_1) + h_y(\theta)(\mathrm{E}[Y] - \theta_2).\end{aligned}$$

Set $\theta = (\mathrm{E}[X], \mathrm{E}[Y])$, and then the second and third terms are 0. Therefore, we are left with

$$\mathrm{E}\left[h(X,Y)\right] \approx h(\mathrm{E}\left[X\right],\mathrm{E}\left[Y\right]).$$

□

The function that will be used with Lemma 3.2.1 is $h(x,y) = x/y$. Next, the bias will be derived.

Theorem 3.2.2**.** *Let* $\widehat{g}$ *be defined as in* (III.4)*. And assume that the kernel function K has the same properties as discussed previously. Properties:*

1. $K(u) \geq 0, -\infty < u < \infty$
2. $\int K(u)\,du = 1$
3. $\int uK(u)\,du = 0$
4. $\sigma_K^2 := \int u^2 K(u)\,du < \infty.$

Next, assume that the regression function $g(x)$ *is second order continuous with respect to its derivative. Lastly, let* $h \to 0$ *such that* $nh^2 \to 0$*, but* $nh \to \infty$*.*

Then $Bias(\widehat{g}(x)) \to 0$ *as* $n \to \infty$*.*

Proof. The function $\hat{m}$ in (II.21) can be further broken down in the following manner:

$$\begin{aligned}
\hat{m}(x) &= (\hat{g}(x) - g(x))\hat{f}_h(x) \\
&= \frac{1}{nh}\sum_{i=1}^{n}\frac{1}{\sqrt{i}}K\left(\frac{X_{(i)}-x}{\sqrt{i}h}\right)Y_{[i]} - \frac{1}{nh}\sum_{i=1}^{n}\frac{1}{\sqrt{i}}K\left(\frac{X_{(i)}-x}{\sqrt{i}h}\right)g(x) \\
&= \frac{1}{nh}\sum_{i=1}^{n}\frac{1}{\sqrt{i}}K\left(\frac{X_{(i)}-x}{\sqrt{i}h}\right)(g(X_{(i)})+\varepsilon_{[i]}) - \frac{1}{nh}\sum_{i=1}^{n}\frac{1}{\sqrt{i}}K\left(\frac{X_{(i)}-x}{\sqrt{i}h}\right)g(x) \\
&= \frac{1}{nh}\sum_{i=1}^{n}\frac{1}{\sqrt{i}}K\left(\frac{X_{(i)}-x}{\sqrt{i}h}\right)(g(X_{(i)})-g(x)) + \frac{1}{nh}\sum_{i=1}^{n}\frac{1}{\sqrt{i}}K\left(\frac{X_{(i)}-x}{\sqrt{i}h}\right)\varepsilon_{[i]} \\
&=: \hat{m}_1(x) + \hat{m}_2(x),
\end{aligned}$$

where $\hat{m}_1$ is the left hand summation term and $\hat{m}_2$ is the right hand summation term.

To get the bias, the expectations of $\hat{m}_1$ and $\hat{m}_2$ are needed. The expectation of $\hat{m}_2$ is simply zero since the ε values are generated independently of the X values and $\mathrm{E}[\varepsilon_i] = 0$. This means that only the expectation of $\hat{m}_1$ need be derived. This can be done in the following manner:

$$\begin{aligned}
\mathrm{E}[\hat{m}_1(x)] &= \mathrm{E}\left[\frac{1}{nh}\sum_{i=1}^{n}\frac{1}{\sqrt{i}}K\left(\frac{X_{(i)}-x}{\sqrt{i}h}\right)(g(X_{(i)})-g(x))\right] \\
&= \int\sum_{i=1}^{n}\frac{1}{nh}\frac{1}{\sqrt{i}}K\left(\frac{t-x}{\sqrt{i}h}\right)(g(t)-g(x))f_{(i)}(t)\,dt \\
&= \frac{1}{n}\sum_{i=1}^{n}\int K(z)(g(x+\sqrt{i}hz)-g(x))f_{(i)}(x+\sqrt{i}hz)\,dz \\
&= \frac{1}{n}\sum_{i=1}^{n}\int K(z)\left(g(x)+\sqrt{i}hzg'(x)+\frac{ih^2z^2}{2}g''(x)+o(ih^2z^2)-g(x)\right) \\
&\qquad\qquad \cdot\left(f_{(i)}(x)+\sqrt{i}hzf'_{(i)}(x)+\frac{ih^2z^2}{2}f''_{(i)}(x)+o(ih^2z^2)\right)dz \\
&= \frac{h^2}{2n}\sum_{i=1}^{n}i\left(2g'(x)f'_{(i)}(x)+g''(x)f_{(i)}(x)\right)\int z^2K(z)\,dz+o(nh^2) \\
&= \frac{\sigma_k^2h^2}{2n}\sum_{i=1}^{n}i\left(2g'(x)f'_{(i)}(x)+g''(x)f_{(i)}(x)\right)+o(nh^2) \\
&= \frac{\sigma_k^2h^2}{2n}\left(2g'(x)\sum_{i=1}^{n}if'_{(i)}(x)+g''(x)\sum_{i=1}^{n}if_{(i)}(x)\right)+o(nh^2).
\end{aligned}$$

In the derivation of (II.1), it was found that

$$\sum_{i=1}^{n}if_{(i)}(x) = nf(x)[1+(n-1)F(x)],$$

which leads to the following:

$$\begin{aligned}
\mathrm{E}[\hat{m}_1(x)] = \frac{\sigma_k^2h^2}{2}&\big(2g'(x)f(x)[1+(n-1)F(x)] \\
&+g''(x)\left\{f(x)[1+(n-1)f(x)]+f'(x)[1+(n-1)F(x)]\right\}\big)+o(nh^2). \qquad \text{(III.7)}
\end{aligned}$$

Denote the coefficient of $\sigma_K^2 h^2/2$ by $A_n(x)$. The expectation can then be written as the following:

$$\begin{aligned}\mathrm{E}\left[\hat{m}(x)\right] &= \hat{m}_1(x) \\ &= \frac{\sigma_K^2 h^2}{2} A_n(x) + o(nh^2). \end{aligned} \tag{III.8}$$

Next, (III.8) can be combined with (II.1) from Chapter 2 to give an approximate bias for $\hat{g}(x)$, as follows:

$$\begin{aligned} Bias(\hat{g}_{RNW}(x)) &= \mathrm{E}\left[\frac{m(x)}{\hat{f}_h(x)}\right] \\ &\approx \frac{\mathrm{E}\left[m(x)\right]}{\mathrm{E}\left[\hat{f}_h(x)\right]} \\ &= \frac{\frac{\sigma_K^2 h^2}{2} A_n(x) + o(nh^2)}{f(x) + \frac{\sigma_K^2 h^2}{2}\left[f''(x) + (n-1)\left(f''(x)F(x) + 3f(x)f'(x)\right)\right] + o(nh^2)} \\ &= O(nh^2). \end{aligned}$$

□

The next step will be to try to get an asymptotic form for the variance. This can be done using the following approximation for two random variables X and Y (Stuart and Ord 1987, Section 10.5):

$$V\left(\frac{X}{Y}\right) \approx \left(\frac{\mathrm{E}\left[X\right]}{\mathrm{E}\left[Y\right]}\right)^2 \left[\frac{V(X)}{\mathrm{E}\left[X\right]^2} + \frac{V(Y)}{\mathrm{E}\left[Y\right]^2} - \frac{2 \cdot Cov(X,Y)}{\mathrm{E}\left[X\right]\mathrm{E}\left[Y\right]}\right]. \tag{III.9}$$

Theorem 3.2.3**.** *Under the same assumptions of Theorem* 3.2.2*.* $Var(\widehat{g}(x)) \to 0$, with the additional restriction that $h \to 0$ such that $n^2 h \to \infty$.

Proof. The approximation (III.9) is applied to the ratio given in (III.4) while noting that

$$\mathrm{Var}(\hat{g}(x)) = \mathrm{Var}(\hat{g}(x) - g(x))$$

$$= \operatorname{Var}\left(\frac{\hat{m}(x)}{\hat{f}_h(x)}\right).$$

Then, $\operatorname{Var}(\hat{m}(x))$ and $Cov(\hat{m}(x), \hat{f}_h(x))$ are needed. For $\operatorname{Var}(\hat{m}(x))$,

$$\operatorname{Var}(\hat{m}(x)) = \operatorname{Var}(\hat{m}_1(x)) + \operatorname{Var}(\hat{m}_2(x)) + 2\operatorname{Cov}(\hat{m}_1(x), \hat{m}_2(x)). \tag{III.10}$$

For $\operatorname{Var}(\hat{m}_1(x))$, $\operatorname{E}\left[\hat{m}_1(x)^2\right]$ is the only component needing to be derived as $\operatorname{E}\left[\hat{m}_1(x)\right]$ is given in (III.8). Note,

$$\begin{aligned}
\operatorname{E}\left[\hat{m}_1(x)^2\right] &= \frac{1}{n^2h^2}\sum_{i=1}^{n}\frac{1}{i}\int K\left(\frac{y-x}{\sqrt{i}h}\right)^2 (g(y)^2 - 2g(y)g(x) + g(x)^2) f_{(i)}(y)dy \\
&\quad + \frac{1}{n^2h^2}\sum\sum_{i\neq j}\frac{1}{\sqrt{ij}}\int\int_{-\infty}^{t} K\left(\frac{s-x}{\sqrt{i}h}\right)K\left(\frac{t-x}{\sqrt{j}h}\right) \\
&\qquad \cdot [g(s)g(t) - g(x)(g(s)+g(t))g(x)^2] f_{(i,j)}(s,t)\,ds\,dt \\
&= \frac{1}{n^2h}\sum_{i=1}^{n}\frac{1}{\sqrt{i}}K(u)^2 (g(x+\sqrt{i}hu)^2 - 2g(x+\sqrt{i}hu)g(x) + g(x)^2) f_{(i)}(x+\sqrt{i}hu)dy \\
&\quad + \frac{1}{n^2}\sum\sum_{i\neq j}\int\int_{-\infty}^{v} K(u)K(v) \\
&\qquad \cdot [g(x+\sqrt{j}hu)g(x+\sqrt{j}hv) - g(x)(g(x+\sqrt{j}hu) + g(x+\sqrt{j}hv)) + g(x)^2] \\
&\qquad \cdot f_{(i,j)}(x+\sqrt{j}hu, x+\sqrt{j}hv)\,ds\,dt.
\end{aligned} \tag{III.11}$$

The two summations will be worked on separately. First,

$$\begin{aligned}
&\frac{1}{n^2h}\sum_{i=1}^{n}\frac{1}{\sqrt{i}}\int K(u)^2 (g(x+\sqrt{i}hu)^2 - 2g(x+\sqrt{i}hu)g(x) + g(x)^2) f_{(i)}(x+\sqrt{i}hu)dy \\
&= \frac{1}{n^2h}\sum_{i=1}^{n}\frac{1}{\sqrt{i}}K\left(u^2\right)\int (g(x)^2 + 2\sqrt{i}hug(x)g'(x) - 2g(x)^2 - 2\sqrt{i}hug(x)g'(x) + g(x)^2 + o(\sqrt{i}hu)) \\
&\qquad \cdot (f_{(i)}(x) + o(1))\,du \\
&= \frac{1}{n^2h}\sum_{i=1}^{n}\frac{1}{\sqrt{i}}\int K(u)^2 (g(x)^2 + 2\sqrt{i}hug(x)g'(x) - 2g(x)^2 - 2\sqrt{i}hug(x)g'(x) + g(x)^2 + o(\sqrt{i}hu))(f_{(i)}(x) + \\
&\qquad \cdot (f_{(i)}(x) + o(1))\,du
\end{aligned}$$

$$= \frac{1}{n^2 h} \sum_{i=1}^{n} \frac{1}{\sqrt{i}} \int K(u)^2 \, o(\sqrt{i} hu) \, du$$
$$= o(n^{-1}). \tag{III.12}$$

Secondly,

$$\frac{1}{n^2} \sum\sum_{i \neq j} \int \int_{-\infty}^{v} K(u) K(v) [g(x + \sqrt{j}hu) g(x + \sqrt{j}hv) - g(x)(g(x + \sqrt{j}hu) + g(x + \sqrt{j}hv)) + g(x)^2]$$
$$\cdot f_{(i,j)}(x + \sqrt{j}hu, x + \sqrt{j}hv) \, ds \, dt$$
$$= \frac{1}{n^2} \sum\sum_{i \neq j} \int \int_{-\infty}^{v} K(u) K(v) \left[\left(g(x) + \sqrt{i}hug'(x) + o(\sqrt{i}hu) \right) \left(g(x) + \sqrt{j}hvg'(x) + o(\sqrt{j}hv) \right) \right.$$
$$\left. - g(x) \left(g(x) + \sqrt{i}hug'(x) + g(x) + \sqrt{j}hvg'(x) + o(h(\sqrt{i}u + \sqrt{j}v) \right) + g(x)^2 \right]$$
$$\left(f_{(i,j)}(x,x) + o(1) \right) dudv$$
$$= \frac{1}{n^2} \sum\sum_{i \neq j} \int \int_{-\infty}^{v} K(u) K(v) \left[g(x)^2 + (\sqrt{i}u + \sqrt{j}v) h g(x) g'(x) + \sqrt{ij} h^2 uv g'(x)^2 \right.$$
$$\left. - 2g(x)^2 - (\sqrt{i}u + \sqrt{j}v) h g(x) g'(x) + g(x)^2 + o(h(\sqrt{i}u + \sqrt{j}v)) \right]$$
$$= \frac{1}{n^2} \sum\sum_{i \neq j} \int \int_{-\infty}^{v} K(u) K(v) \left[\sqrt{ij} h^2 uv g'(x)^2 + o(h(\sqrt{i}u + \sqrt{}j)) \right] \left(f_{(i,j)}(x,x) + o(1) \right) dudv.$$

Using integral approximations, it can be seen that $\sum\sum_{i \neq j} \sqrt{ij} = o(n^3)$, and just as in the density estimation case $\sum\sum_{i \neq j} (\sqrt{i} + \sqrt{j}) = o(n^{5/2})$. Thus, one can get

$$\frac{1}{n^2} \sum\sum_{i \neq j} \int \int_{-\infty}^{v} K(u) K(v) \left[\sqrt{ij} h^2 uv g'(x)^2 + o(h(\sqrt{i}u + \sqrt{}j)) \right] \left(f_{(i,j)}(x,x) + o(1) \right) dudv$$
$$= o(nh^2) + o(\sqrt{n}h)$$
$$= o(\sqrt{n}h). \tag{III.13}$$

Then using (III.12) and (III.13),

$$\mathrm{E}\left[\hat{m}_1(x)^2 \right] = o(n^{-1} + \sqrt{n}h), \tag{III.14}$$

which gives the asymptotic order of the variance of $\hat{m}_1(x)$ when combined with (III.8). The result is

$$
\begin{aligned}
\operatorname{Var}(\hat{m}_1(x)) &= \mathrm{E}\left[\hat{m}_1(x)^2\right] - \mathrm{E}\left[\hat{m}_1(x)\right]^2 \\
&= o(n^{-1} + \sqrt{n}h) \\
&\quad - \left(\frac{\sigma_k^2 h^2}{2} \left(2g'(x)f(x)[1+(n-1)F(x)] \right.\right. \\
&\qquad \left.\left. + g''(x)\left\{f(x)[1+(n-1)f(x)] + f'(x)[1+(n-1)F(x)]\right\}\right) + o(nh^2) \right)^2 \\
&= o(n^{-1} + \sqrt{n}h). \qquad \text{(III.15)}
\end{aligned}
$$

For $\operatorname{Var}(\hat{m}_2(x))$ one needs only to find $\mathrm{E}\left[\hat{m}_2(x)^2\right]$ since $\mathrm{E}\left[\hat{m}_2(x)\right] = 0$. Thus,

$$
\begin{aligned}
\operatorname{Var}(\hat{m}_2(x)) &= \mathrm{E}\left[\hat{m}_2(x)^2\right] \\
&= \frac{1}{n^2h^2} \mathrm{E}\left[\left(\sum_{i=1}^n \frac{1}{\sqrt{i}} \varepsilon_{[i]} K\left(\frac{X_{(i)} - x}{\sqrt{i}h}\right)\right)^2 \right] \\
&= \frac{1}{n^2h^2} \sum_{i=1}^n \frac{1}{i} \mathrm{E}\left[\varepsilon_{[i]}^2 K\left(\frac{X_{(i)} - x}{\sqrt{i}h}\right)^2 \right] \\
&\quad + \frac{1}{n^2h^2} \sum\sum_{i \neq j} \frac{1}{\sqrt{ij}} \mathrm{E}\left[\varepsilon_{[i]}\varepsilon_{[j]} K\left(\frac{X_{(i)} - x}{\sqrt{i}h}\right) K\left(\frac{X_{(j)} - x}{\sqrt{j}h}\right)\right] \\
&= \frac{\sigma_\varepsilon^2}{n^2h^2} \sum_{i=1}^n \frac{1}{i} \int K\left(\frac{y-x}{\sqrt{i}h}\right)^2 f_{(i)}(y)\,dy + 0 \\
&= \frac{\sigma_\varepsilon^2 R(K)}{n^2 h} \sum_{i=1}^n \frac{1}{\sqrt{i}} f_{(i)}(x) + o\left(\frac{1}{nh}\right) \qquad \text{(III.16)}
\end{aligned}
$$

Lastly, for the covariance of $\hat{m}_1$ and $\hat{m}_2$ (the single and double sum of $\mathrm{E}\left[\hat{m}_1(x)^2\right]$) one can obtain

$$
\begin{aligned}
\operatorname{Cov}(\hat{m}_1(x), \hat{m}_2(x)) &= \mathrm{E}\left[\hat{m}_1(x)\hat{m}_2(x)\right] + \mathrm{E}\left[\hat{m}_1(x)\right]\mathrm{E}\left[\hat{m}_2(x)\right] \\
&= \mathrm{E}\left[\left(\sum_{i=1}^n \frac{1}{\sqrt{i}} K\left(\frac{X_{(i)} - x}{\sqrt{i}h}\right)(g(X_{(i)}) - g(x)) \right) \left(\sum_{i=1}^n \frac{1}{\sqrt{i}} K\left(\frac{X_{(i)} - x}{\sqrt{i}h}\right) \varepsilon_{[i]} \right) \right] + 0
\end{aligned}
$$

$$
\begin{aligned}
&= \frac{1}{n^2h^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\frac{1}{\sqrt{ij}}\,\mathrm{E}\left[K\left(\frac{X_{(i)}-x}{\sqrt{i}h}\right)(g(X_{(i)})-g(x))K\left(\frac{X_{(j)}-x}{\sqrt{j}h}\right)\varepsilon_{[j]}\right] \\
&= \frac{1}{n^2h^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\frac{1}{\sqrt{ij}}\,\mathrm{E}\left[K\left(\frac{X_{(i)}-x}{\sqrt{i}h}\right)(g(X_{(i)})-g(x))K\left(\frac{X_{(j)}-x}{\sqrt{j}h}\right)\right]\mathrm{E}\left[\varepsilon_{[j]}\right] \\
&= 0. \qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad \text{(III.17)}
\end{aligned}
$$

Taking the result above, (III.17) , and combining this with (III.15) and (III.16), one can get an expression for the variance of $\hat{m}(x)$. This results in

$$
\begin{aligned}
\mathrm{Var}(\hat{m}(x)) &= \mathrm{Var}(\hat{m}_1(x)) + \mathrm{Var}(\hat{m}_2(x)) + 2\,\mathrm{Cov}(\hat{m}_1(x), \hat{m}_2(x)) \\
&= o(n^{-1} + \sqrt{n}h) + \frac{\sigma_\varepsilon^2 R(K)}{n^2h}\sum_{i=1}^{n}\frac{1}{\sqrt{i}}f_{(i)}(x) + o\left(n^{-3/2}h^{-1}\right) \\
&= \frac{\sigma_\varepsilon^2 R(K)}{n^2h}\sum_{i=1}^{n}\frac{1}{\sqrt{i}}f_{(i)}(x) + o\left(n^{-1} + \sqrt{n}h + n^{-3/2}h^{-1}\right).
\end{aligned}
$$

Lastly, $\mathrm{Cov}(\hat{m}(x), \hat{f}_h(x))$ is needed. Note that

$$
\mathrm{Cov}(\hat{m}(x), \hat{f}_h(x)) = \mathrm{E}\left[\hat{m}(x)\hat{f}_h(x)\right] - \mathrm{E}\left[\hat{m}(x)\right]\mathrm{E}\left[\hat{f}_h(x)\right].
$$

The right hand term is known since $\mathrm{E}\left[\hat{m}(x)\right]$ and $\mathrm{E}\left[\hat{f}_h(x)\right]$ have already been derived. Thus, work on the left hand term gives the following:

$$
\begin{aligned}
&\mathrm{E}\left[\hat{m}(x)\hat{f}_h(x)\right] \\
&= \mathrm{E}\left[\left(\frac{1}{nh}\sum_{i=1}^{n}\frac{1}{\sqrt{i}}K\left(\frac{X_{(i)}-x}{\sqrt{i}h}\right)(g(X_{(i)})-g(x)) + \frac{1}{nh}\sum_{i=1}^{n}\frac{1}{\sqrt{i}}K\left(\frac{X_{(i)}-x}{\sqrt{i}h}\right)\varepsilon_{[i]}\right)\right. \\
&\qquad \left.\cdot\left(\frac{1}{nh}\sum_{i=1}^{n}\frac{1}{\sqrt{i}}K\left(\frac{x-X_{(i)}}{\sqrt{i}h}\right)\right)\right] \\
&= \mathrm{E}\left[\left(\frac{1}{nh}\sum_{i=1}^{n}\frac{1}{\sqrt{i}}K\left(\frac{X_{(i)}-x}{\sqrt{i}h}\right)(g(X_{(i)})-g(x))\right)\left(\frac{1}{nh}\sum_{i=1}^{n}\frac{1}{\sqrt{i}}K\left(\frac{x-X_{(i)}}{\sqrt{i}h}\right)\right)\right] \\
&\quad + \frac{1}{n^2h^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\frac{1}{\sqrt{ij}}\,\mathrm{E}\left[K\left(\frac{x-X_{(i)}}{\sqrt{i}h}\right)K\left(\frac{x-X_{(j)}}{\sqrt{j}h}\right)(g(X_{(i)}-g(x)))\right]
\end{aligned}
$$

$$= \frac{1}{n^2h^2}\sum_{i=1}^{n}\frac{1}{i}\,\mathrm{E}\left[K\left(\frac{x-X_{(i)}}{\sqrt{i}h}\right)^2(g(X_{(i)}-g(x)))\right]$$
$$+\frac{2}{n^2h^2}\sum\sum_{i<j}\frac{1}{\sqrt{ij}}\,\mathrm{E}\left[K\left(\frac{x-X_{(i)}}{\sqrt{i}h}\right)K\left(\frac{x-X_{(j)}}{\sqrt{j}h}\right)(g(X_{(i)}-g(x)))\right]$$
$$= \frac{1}{n^2h}\sum_{i=1}^{n}\frac{1}{\sqrt{i}}\int K(z_i)^2\,(g(x+\sqrt{i}hz_i)-g(x))f_{(i)}(x+\sqrt{i}hz_i)\,dz_i$$
$$+\frac{2}{n^2}\sum\sum_{i<j}\int\int_{-\infty}^{v_j}K(u_i)K(v_j)\,(g(x+\sqrt{i}hu_i)-g(x))f_{(i,j)}(x+\sqrt{i}hu_i,x+\sqrt{j}hv_j)\,du_idv_j$$
$$= \frac{1}{n^2h}\sum_{i=1}^{n}\frac{1}{\sqrt{i}}\int K(z_i)^2\,(\sqrt{i}hz_ig'(x)+o(\sqrt{i}h))(f_{(i)}(x)+\sqrt{i}hz_if'_{(i)}(x)+o(\sqrt{i}h))$$
$$+\frac{2}{n^2}\sum\sum_{i<j}\int\int_{-\infty}^{v_j}K(u_i)K(v_j)\,(\sqrt{i}hz_ig'(x)+o(\sqrt{i}h))$$
$$\left(f_{(i,j)}(x,x)+uh\sqrt{i}\frac{\partial}{\partial s}f_{(i,j)}(s,t)\bigg|_{t=x_+}^{s=x}+vh\sqrt{j}\frac{\partial}{\partial t}f_{(i,j)}(s,t)\bigg|_{t=x_+}^{s=x}+o\left(h(\sqrt{i}+\sqrt{j})\right)\right)du\,dv$$
$$= \frac{1}{n^2h}\sum_{i=1}^{n}\frac{1}{\sqrt{i}}\int K(z_i)^2\,(\sqrt{i}hz_if_{(i)}(x)g'(x)+ih^2z^2f'_{(i)}(x)g'(x)+o(\sqrt{i}h))dz_i$$
$$+\frac{2}{n^2}\sum\sum_{i<j}\int\int_{-\infty}^{v_j}K(u_i)K(v_j)$$
$$\left(\sqrt{i}hu_if_{(i,j)}(x,x_+)g'(x)+h^2g'(x)\left(u\sqrt{i}\frac{\partial}{\partial s}f_{(i,j)}(s,t)\bigg|_{t=x_+}^{s=x}+v\sqrt{j}\frac{\partial}{\partial t}f_{(i,j)}(s,t)\bigg|_{t=x_+}^{s=x}\right)\right.$$
$$\left.+o\left(h(\sqrt{i}+\sqrt{j})\right)\right)du\,dv$$
$$= o(\sqrt{n}h).$$

This allows for the completion of the approximation of the variance of $\hat{g}(x)$. Let $A_n(x) = \sigma_K^2h^2/2$, and $B_n(x) = \frac{\sigma_K^2h^2}{2}\left[f''(x)+(n-1)\left(f''(x)F(x)+3f(x)f'(x)\right)\right]$. Then we obtain the following form for the variance:

$$V(\hat{g}(x)) \approx \left(\frac{\mathrm{E}[\hat{m}(x)]}{\mathrm{E}[\hat{f}_h(x)]}\right)^2\left[\frac{V(\hat{m}(x))}{\mathrm{E}[\hat{m}(x)]^2}+\frac{V(\hat{f}_h(x))}{\mathrm{E}[\hat{f}_h(x)]^2}-\frac{2\cdot Cov(\hat{m}(x),\hat{f}_h(x))}{\mathrm{E}[\hat{m}(x)]\,\mathrm{E}[\hat{f}_h(x)]}\right]$$
$$= \left(\frac{\frac{\sigma_K^2h^2}{2}A_n(x)+o(nh^2)}{f(x)+\frac{\sigma_K^2h^2}{2}B_n(x)+o(nh^2)}\right)^2$$

$$
\begin{aligned}
&\cdot \left[\frac{o(n^{-1}+\sqrt{n}h)}{\left(\frac{\sigma_K^2 h^2}{2}A_n(x)+o(nh^2)\right)^2} + \frac{\frac{R(K)}{n^2h}\sum_{i=1}^n \frac{1}{\sqrt{i}} f_{(i)}(x)+o(\sqrt{n}h)}{\left(f(x)+\frac{\sigma_K^2 h^2}{2}B_n(x)+o(nh^2)\right)^2} \right. \\
&+ \left. \frac{o(\sqrt{n}h)}{\left(\frac{\sigma_K^2 h^2}{2}A_n(x)+o(nh^2)\right)\left(f(x)+\frac{\sigma_K^2 h^2}{2}B_n(x)+o(nh^2)\right)} \right] \\
&= \left(\frac{1}{f(x)+\frac{\sigma_K^2 h^2}{2}B_n(x)+o(nh^2)} \right)^2 \\
&\cdot \left[o(n^{-1}+\sqrt{n}h) + \frac{\left(\frac{\sigma_K^2 h^2}{2}A_n(x)+o(nh^2)\right)^2 \left(\frac{R(K)}{n^2h}\sum_{i=1}^n \frac{1}{\sqrt{i}} f_{(i)}(x)+o(\sqrt{n}h)\right)}{\left(f(x)+\frac{\sigma_K^2 h^2}{2}B_n(x)+o(nh^2)\right)^2} \right. \\
&+ \left. \frac{\left(\frac{\sigma_K^2 h^2}{2}A_n(x)+o(nh^2)\right) o(\sqrt{n}h)}{f(x)+\frac{\sigma_K^2 h^2}{2}B_n(x)+o(nh^2)} \right]. \qquad \text{(III.18)}
\end{aligned}
$$

Note that

$$f(x)+\frac{\sigma_K^2 h^2}{2}B_n(x)+o(nh^2) \to f(x)$$

for any x in the distribution support. Additionally, the following converge to 0:

1. $n^{-1}+\sqrt{n}h$
2. $\frac{\sigma_K^2 h^2}{2}A_n(x)+o(nh^2)$
3. $\frac{R(K)}{n^2h}\sum_{i=1}^n \frac{1}{\sqrt{i}} f_{(i)}(x)+o(\sqrt{n}h)$

Therefore, the variance converges to 0 under the assumptions for the rate of convergence of h. $\square$

Trying to manipulate the form of the variance given in (III.18) into something more tractable or convenient has not produced any results. The convenient form of the variance for the standard NW estimator in (I.15) is a result of its simpler overall structure combined with the use of algebraic approximations.

3.3 **Bandwidth Selection**

As mentioned in the review of bandwidth methods for the Nadaraya-Watson (NW) regression estimator, there are two main types of bandwidth selection methods for h in $\hat{g}_{NW}(x)$: cross-validation and penalized residual sum of squares. Currently, both types are being investigated for the rank based Nadaraya-Watson (RNW) estimate. Denote the rank based Nadaraya-Watson estimate by $\hat{g}_{RNW}(x)$. A similar form for both types of bandwidth methods will be used.

The cross-validation method will be quite similar to that for the standard NW estimator in (I.17). The objective function being used for the RNW estimator is

$$CV(h) = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{g}_{-i,h}(x))^2,$$

where

$$\hat{g}_{-i,h}(x) = \frac{\sum_{j\neq i}^{n} \frac{1}{\sqrt{j}} K\left(\frac{X_{(j)} - x}{\sqrt{j}h}\right) Y_{[j]}}{\sum_{j\neq i}^{n} \frac{1}{\sqrt{j}} K\left(\frac{X_{(j)} - x}{\sqrt{j}h}\right)},$$

the leave-one-out estimator that does not re-rank the data.

The other main bandwidth selection method being used is that of the penalized residual sum of squares given by

$$G(h) = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{g}(X_i))^2 \Xi\left(\frac{1}{n}W_{hi}(X_i)\right),$$

where

$$W_{hi} = \frac{\frac{1}{\sqrt{i}} K\left(\frac{X_{(i)} - x}{\sqrt{i}h}\right) Y_{[i]}}{\sum_{i=1}^{n} \frac{1}{\sqrt{i}} K\left(\frac{X_{(i)} - x}{\sqrt{i}h}\right)}.$$

There are several functions that can be tried for Ξ. Additionally, usage of the *AIC* given by (I.18) is also easily implemented given the generality of its form. Simulations presented within this work are done using the CV, Shibata, and AIC bandwidth selection methods presented within the literature review.

CHAPTER IV

SIMULATIONS

This chapter concentrates on some numerical results derived from simulations that are meant to compare the standard density and Nadaraya-Watson estimators given in the introductory chapter with their rank weighted counterparts given in the previous two chapters.

A series of simulations was performed to see how the rank-weighted estimator compares to the standard density estimator. First, the bandwidth choices in the simulations will be discussed, and then the distributions simulated will be listed.

For the rank weighted density estimator, all four discussed bandwidth selectors are used: Rule-of-Thumb (II.24), LSCV with a Gaussian kernel (II.31), the plug-in estimator (II.33), and the hybrid rule-of-thumb estimator (II.34). For the standard density estimator, the bandwidth choices are made by the rule-of-thumb approach, LSCV, and Sheather-Jones plug-in estimator.

A series of skewed distributions and normal mixtures were simulated. Under each distribution, 3000 samples were taken under each of the following sample size settings: $n = 20$, 50, 100, 200, and 400. For each sample, the density estimators were calculated using the different bandwidth selection methods discussed. In each distributional setting, the bias and variance were estimated at 512 equidistant points across the support. These are combined in MSE estimates at the points. Using numerical integration, these MSE values are used to approximate the MISE. These MISE estimates are used to evaluate which estimator performs in a superior manner.

The choice of distributions fall into two main categories: unbounded skewed distributions, and normal distribution mixtures. Both of the two unbounded skewed distributions used are referred to as ”skew normal” distributions; both of these skew normal distributions allow for a control of a ”skewness parameter” that increases the skewness of the distribution. These skew normal distributions have distinct shapes as this skewness parameter increases. They allow for the ability to see under what level of skewness, if any, would the expected superiority of the RWKDE be achieved.

The first skew normal distribution is the skew normal distribution from Azzalini (1985). This skew normal distribution will be represented by the notation $SkewNormalAZ(\mu, \sigma, \alpha)$, with parameters for location μ, scale σ, and shape α. This skew normal distribution is characterized by the following density function using ϕ and Φ to represent the standard normal density and cumulative distribution functions, respectively.

$$f(x) = \frac{1}{\sigma}\phi\left(\frac{x-\mu}{\sigma}\right)\Phi\left(\alpha\frac{x-\mu}{\sigma}\right).$$

As α increases, this distribution becomes more and more skewed. This can be seen in Figure 4.1. For the simulations, results for when $\alpha = 2, 6,$ and 10 will be discussed.

The second skew normal distribution is given by Fernández and Steel (1998). The notation for this skew normal will be $SkewNormalFG(\mu, \sigma, \xi)$. Similarly, the parameters are for location μ, scale σ, and shape ξ. The density function for the skew normal FG distribution is given by the following:

$$f(x) = \frac{2}{\xi + \frac{1}{\xi}}\left\{\phi\left(\frac{x-\mu}{\xi\sigma}\right)\mathrm{I}(x \geq \mu) + \phi\left(\frac{\xi(x-\mu)}{\sigma}\right)\mathrm{I}(x < \mu)\right\}.$$

Similar to the previous skew normal distribution, the skew normal FG distribution becomes more and more skewed as ξ increases. This is depicted in Figure 4.2. For the simulations, results for when $\xi = 2, 6, 10$ will be discussed.

A variety of mixtures of normal distributions were also considered. The following table gives the mixtures. The notation here shows the standard deviation as the second parameter.

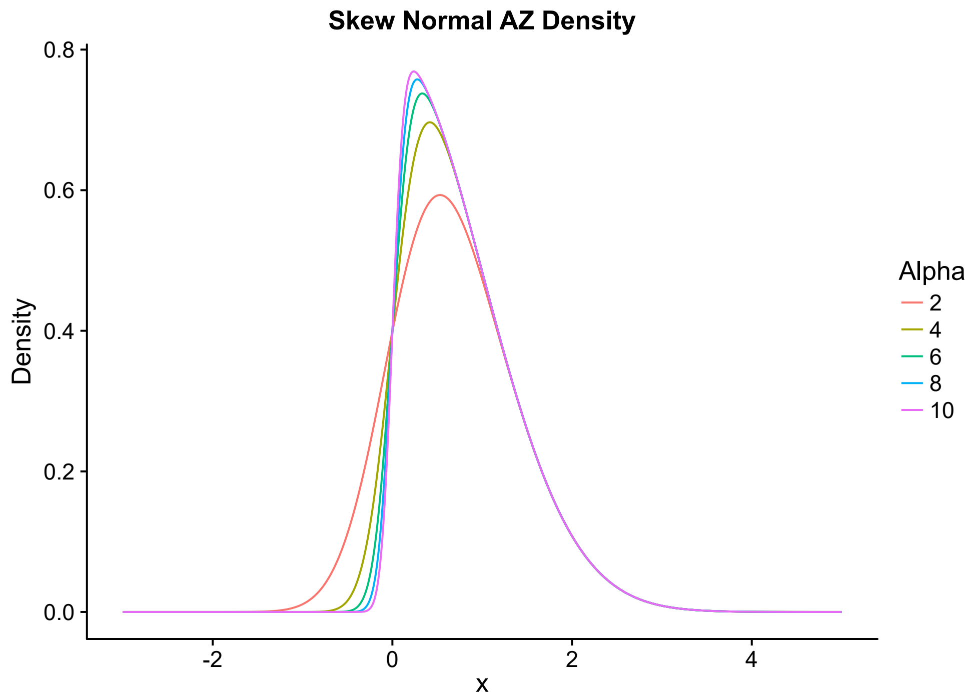


Figure 4.1: Depiction of skew normal AZ distribution as α increases.

A depiction of these mixtures is available in Figure 4.3.

	Mixture
1	$0.6 \cdot N(-1.5, 1) + 0.4 \cdot N(1.5, 1)$
2	$1/3 \cdot N(-1, 0.4) + 1/3 \cdot N(0, 0.3) + 1/3 \cdot N(2, 1)$
3	$1/3 \cdot N(-1, 0.3) + 1/3 \cdot N(0, 0.5) + 1/3 \cdot N(2, 1)$
4	$1/9 \cdot N(-2, \sqrt{0.05}) + 8/9 \cdot N(0, 4)$

Table 4.1: Listing of the Normal Mixtures used in density simulations.

Another set of kernel density estimation simulations were ran for testing the transformation kernel density estimator (TKDE) versus the rank weight transformation kernel density estimator. The simulation were ran under the $Lognormal(\mu, \sigma)$ distribution with $\mu = 0$ and $\sigma = 0.5, 1, 2$. Additionally, results were obtained from the $Weibull(1, \kappa)$ distribution where κ is the shape parameter and the values chosen were $\kappa = 0.5, 1$, and 2. Under each distributional setting, 3000 replications were considered. The MISE was calculated in the

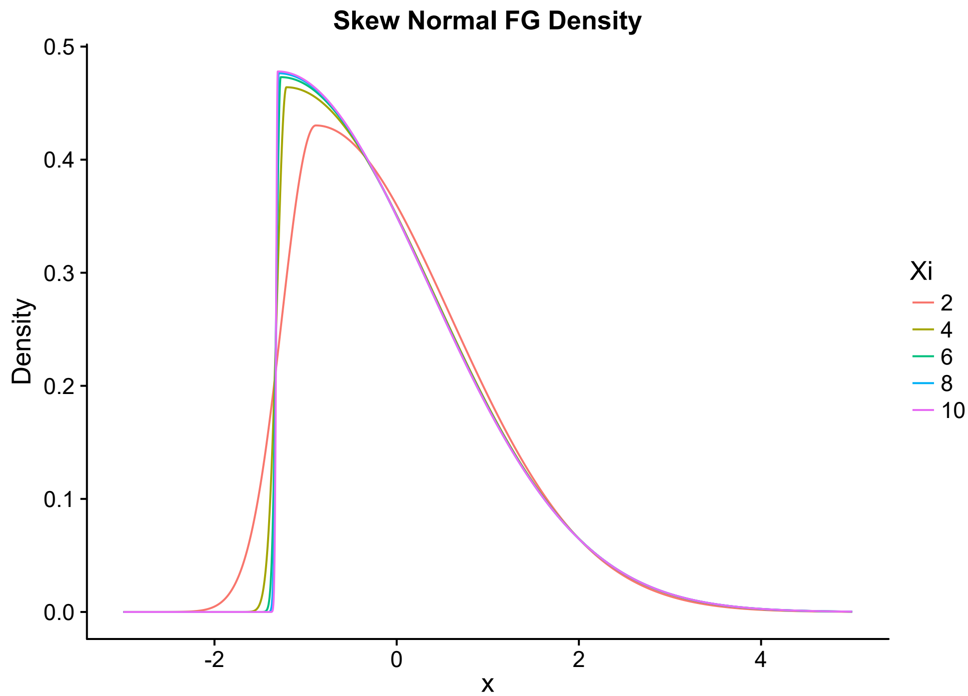


Figure 4.2: Depiction of skew normal FG distribution as ξ increases.

same manner as in the non-transformation kernel simulations. An investigation of bandwidth behavior is not provided due to there being only one bandwidth method under TKDE and RWTKDE.

The last set of simulations will test the NW estimator versus the RWNW estimator in a regression setting. Five model functions $g(x)$ were considered for regression simulations. The form for the functions is given in Table 4.2, and a visualization of these functions is given in Figure IV.

	Function: $g(x)$
1	$1-48x+218x^2-315x^3+145x^4$
2	$\log(x+.01)$
3	$\sin(\pi(1-x))\cdot\cos(4\pi(1-x)^5)$
4	e^{-5x}
5	$1.5\sin(4x)$

Table 4.2: The five regression model functions considered for simulations.

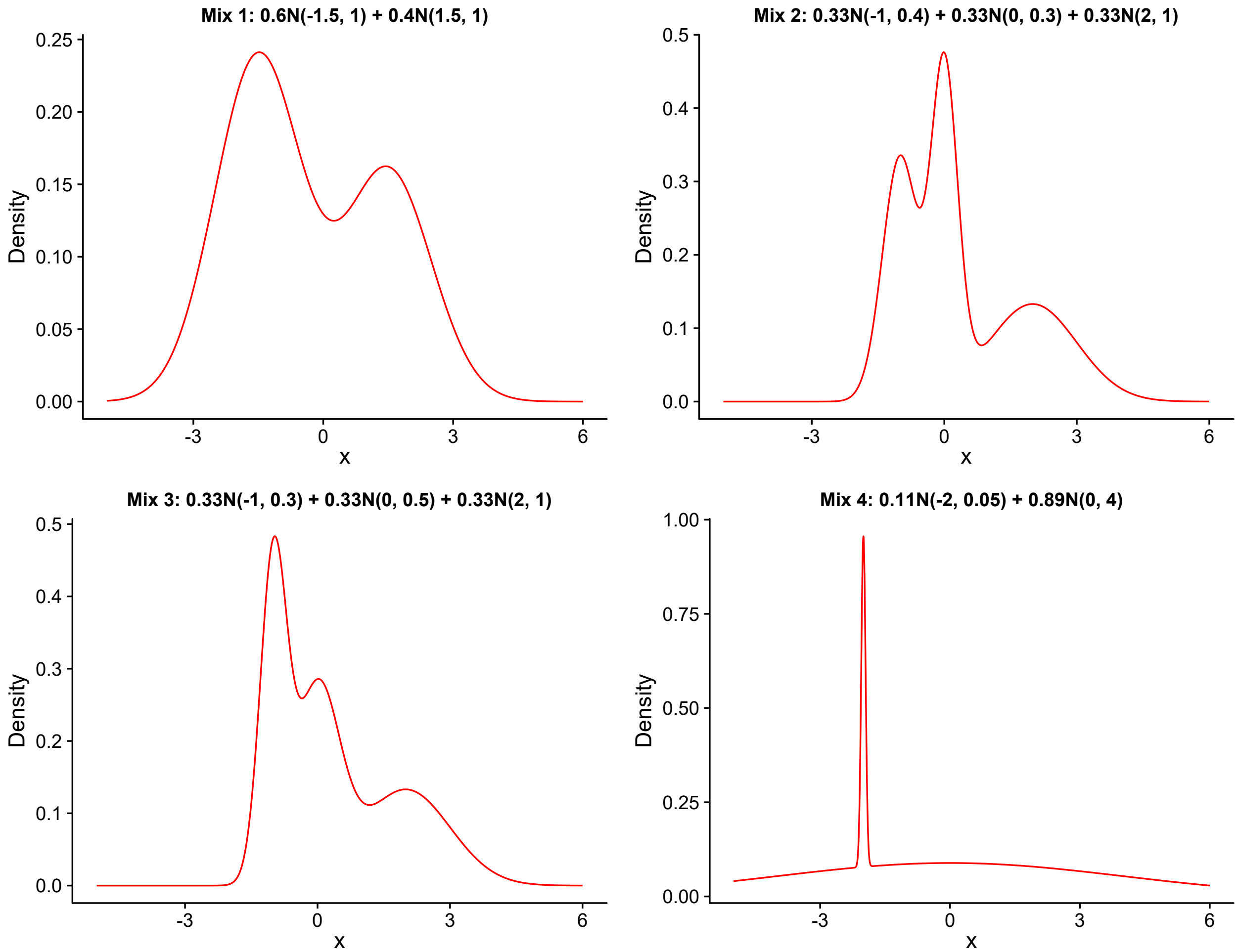


Figure 4.3: Depiction of normal mixture distributions.

Beta Distribution

1	$Beta(1,1)$
2	$Beta(1,3)$
3	$Beta(1,5)$
4	$Beta(1,7)$

Mixture Distribution

1	$.5 \cdot Beta(1,7) + .5 \cdot Beta(1,1)$
2	$.8 \cdot Beta(1,5) + .2 \cdot Beta(5,2)$

Table 4.3: Table listing the distributions for the x variable in the regression simulations.

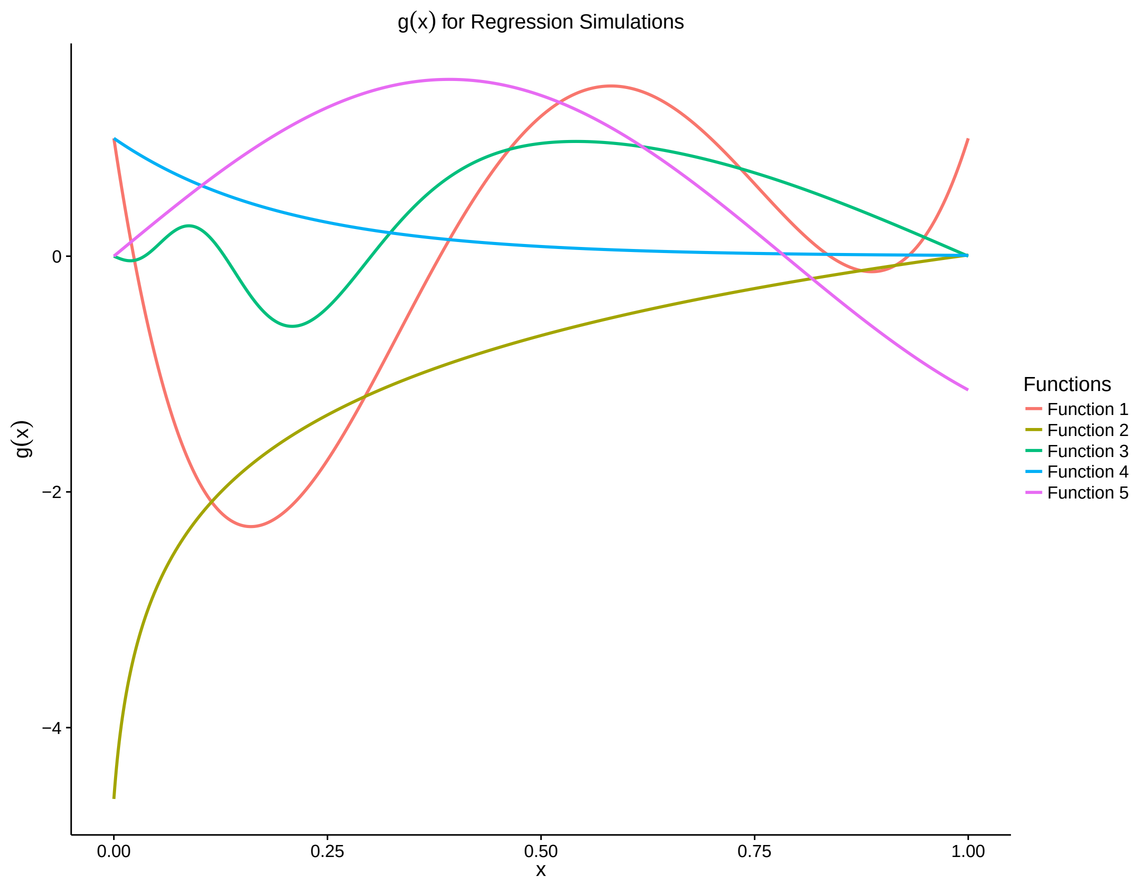


Figure 4.4: Depiction of the different functions chosen for regression simulations.

Function 1 was also used by Hurvich, Simonoff, and Tsai (1998) due to it being a function with ”less fine structure and a trend, in some sense ’typical’ of many regression situations.” Function 2 was chosen due to it having even less fine structure, but still a moderate trend. The 0.01 was added to avoid values approaching negative infinity. Function 3 has a mixture of both fine structure and steady structure, so it is a good test for the adaptability of an estimator. Functions 4 and 5 are chosen as fairly moderate shapes in terms of complexity.

Under each function, several different beta distributions were used to simulate the x values. The beta distribution was selected because it gave quite a bit of variability for how the x values could distributed. Mixtures of beta distributions were also used. The specific distributions can be found in Table 4.3. For a visualization of these densities and mixture densities, see Figure **??**. The unmixed beta densities were chosen due to the increasing right skew nature when the second shape parameter is increased. The mixtures were chosen to see how the estimators handle different structures of skewness. In mixture one, the density does not converge to 0, but in mixture two, the density only goes to 0 at the very far right. This is somewhat different behavior than in the unmixed beta densities.

For each set of x values simulated the y values were generated using $y = g(x) + \varepsilon$ where $\varepsilon \sim N(0, 0.1^2)$ and $N(0, 0.5^2)$ which represent high and moderate correlation between the x and y observations. In a similar manner to the density estimation case, the bias, variance, and MSE were estimated pointwise using 512 points across the horizontal axis. The MISE was then approximated using the 512 MSE calculations and applying a trapezoidal approximation to numerically integrate the MSE values. These MISE approximations are used as the comparison tool between the rank-weighted and standard estimator.

The regression simulations were done under several different bandwidth settings for both the NW and RWNW estimator. Both estimators used the following bandwidth method discussed previously: Shibata, AIC, and LSCV. Since each of the methods is applied nearly identically in both estimators, this allows a very direct comparison between the NW and RWNW estimators.

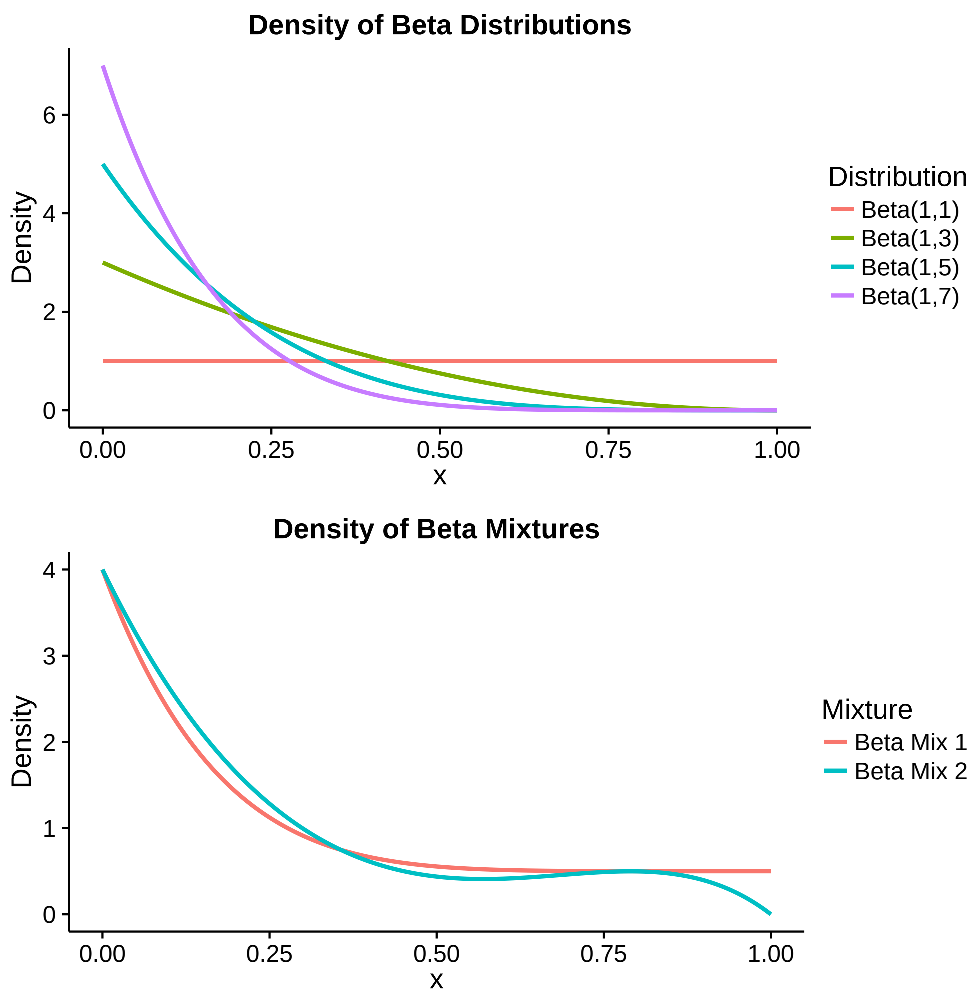


Figure 4.5: Depiction of the beta and beta mixture distributions used for simulating the x variable in regression simulations.

4.1 Kernel Density Estimation Simulations

To review the performance of the RWKDE, an investigation of how the bandwidth selection methods vary in terms of bandwidth choice is presented. The best case would be a stable, unimodal distribution of bandwidth choices with minimal variability. If there is too much spread in the possible choices for bandwidth, the method is less reliable. As a reference to excellent bandwidth behavior, look to the bandwidth behavior of the rule-of-thumb method depicted in Figure 4.6. Note that at each sample size, it is unimodal and the least varied of the other selection methods. The issue with the rule-of-thumb method is that is assumes a normal distribution, so it may be centered at values that are too high or low depending on the underlying distribution of the sample generated.

The following Figures show density plots of the bandwidth choice methods for the rank-weighted estimator under all distributions discussed so far: Rule-of-Thumb (II.24), LSCV with a Gaussian kernel (II.31), the plug-in estimator (II.33), and the hybrid rule-of-thumb estimator(II.34). First, the *SkewNormalAZ* with $\alpha = 2, 6, 10$ results will be discussed. Then, the *SkewNormalFG* results with $\xi = 2, 6, 10$ will be considered. Finally, the bandwidth plots for the normal mixtures will be examined. The plots will display a density plot of the bandwidth choices at sample sizes of 20, 100, and 400.

In Figures 4.6 and 4.7, a summarization of the skew normal AZ results can be seen. Results for the other skew normal AZ distribution settings were similar and can be seen in the appendix. The rule-of-thumb bandwidth seems to be the most stable/least varied among the various bandwidth selection methods. The hybrid method seems to have the same stability as the rule-of-thumb method at the largest sample sizes but the bandwidth choice seems to be bi-modal. This indicates unstable parameter estimation in the MLE estimates of the skew normal AZ distribution. The most intriguing part about this is that, this instability in the bandwidth choice gets worse as the skewness parameter increases. The plug-in estimator and LSCV estimator are the most spread out in bandwidth choice with the plug-ing estimator regularly selecting the smallest bandwidths as sample size increases. An interesting result is that the LSCV selector seems to have a mean value that is higher than the other selectors whenever the sample size increases. This is most likely because, as mentioned, the LSCV algorithm given may no longer be an unbiased estimator of the ISE given in (II.25). It should be noted that across all distributions, the behavior of the rule-of-thumb selector was relatively stable, i.e., not changing much between distributions.

For the skew normal FG distribution, there were some similarities and some differences in bandwidth selection when compared to the AZ version. First, when the skewness parameter of the FG version is set to $\xi = 2$, the spread of bandwidth choice was almost identical to when bandwidths were selected for the skew normal AZ distribution with skewness parameter of $\alpha = 10$ (see Figure 4.7). The difference between the skew normal AZ and FG distributions in terms of bandwidth choice occurred when $\xi = 6$ and 10 for the skew normal FG distribution. The results for $\xi = 10$ are pictured in Figure 4.8, for $\xi = 6$ the results were similar. The spike in the hybrid bandwidth method seems much more stable when compared to that in the skew normal AZ distribution. The spike in hybrid bandwidth choice occurs near the lowest of bandwidth choices, but there is little bimodality even at the smallest sample sizes. In the AZ version, the hybrid bandwidth choice showed strong bimodality in some cases. In both distributions, the LSCV and plug-in selectors are fairly spread out, showing little to no signs of converging. The LSCV selector seems to have a mean that is above all other bandwidths, just as in the AZ case.

Bandwidth choice among the normal mixtures did not yield results that were very different from those previously discussed. Mixture 1 had results that were nearly identical to the skew normal AZ with $\alpha = 2$ (see Figure 4.6). Mixtures 2 and 3 were also similar to this. The results for mixture 4 were quite different and are pictured in Figure 4.9. It can be seen that the behavior of the LSCV and plug-in bandwidth estimators changes a great deal as the sample size increases. The two bandwidth selection methods are centered at much lower values than the rule-of-thumb and hybrid methods at the largest sample size. Additionally, at the largest sample size some rather extreme behavior from the LSCV selector can be seen in the right tail of the graph. There seems to be a small number of bandwidths chosen by the LSCV selector that were still far above any other bandwidth choice. This behavior from the LSCV is inexplicably extreme.

In general, it seems that at smaller sample sizes, say up to $n = 100$, the rule-of-thumb selector may be best given its stability. But at larger sample sizes, the hybrid selector may be a better choice given its similar stability at those sizes. If the main concern is to avoid undersmoothing, the LSCV seems the proper choice given its higher mean value.

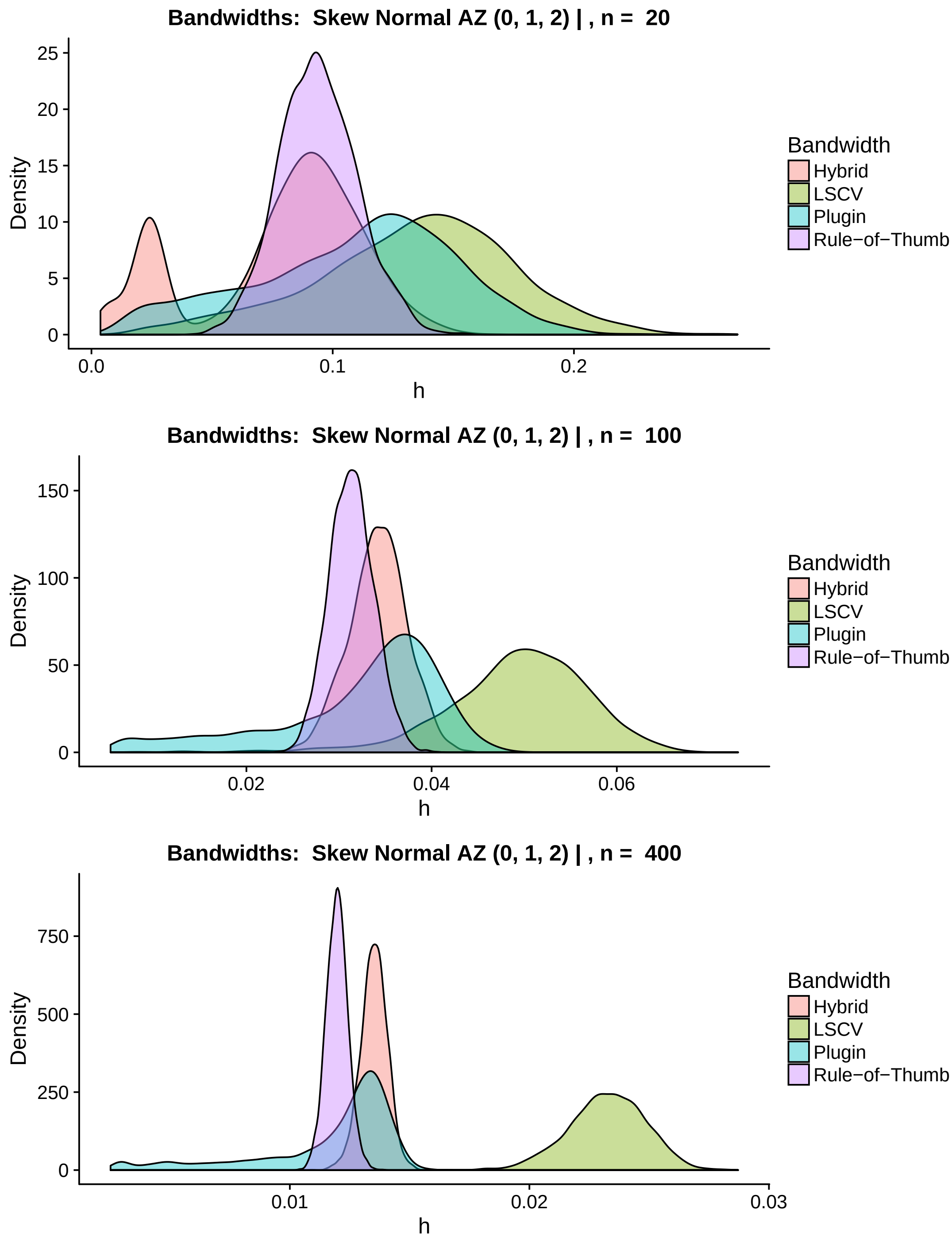


Figure 4.6: Plot of bandwidths for each of the four bandwidth selection methods discussed for the rank weighted estimator at sample sizes 20, 100, 200, and 400 of the $SkewNormAZ(0,1,2)$.

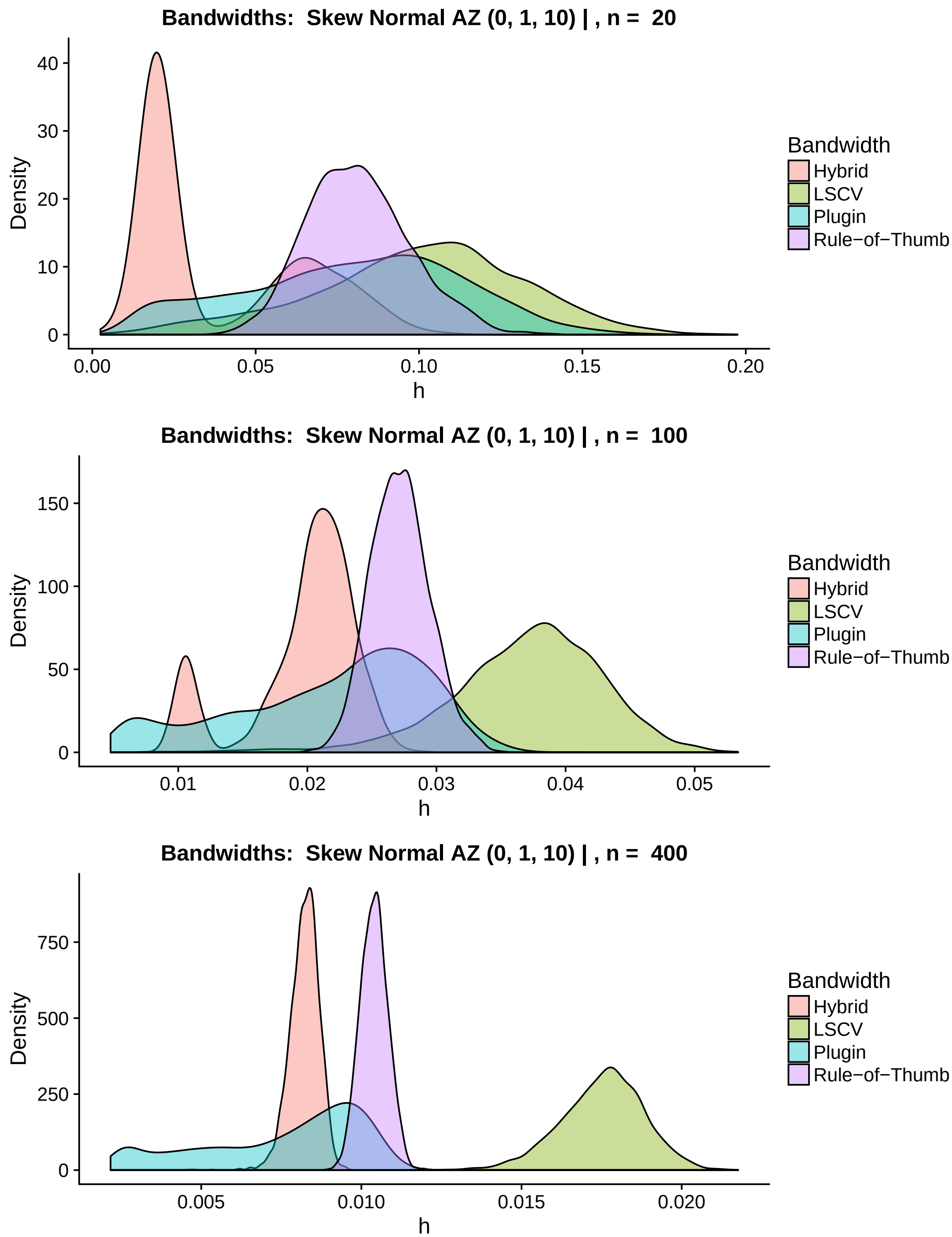


Figure 4.7: Plot of bandwidths for each of the four bandwidth selection methods discussed for the rank weighted estimator at sample sizes 20, 100, 200, and 400 of the $SkewNormAZ(0,1,10)$.

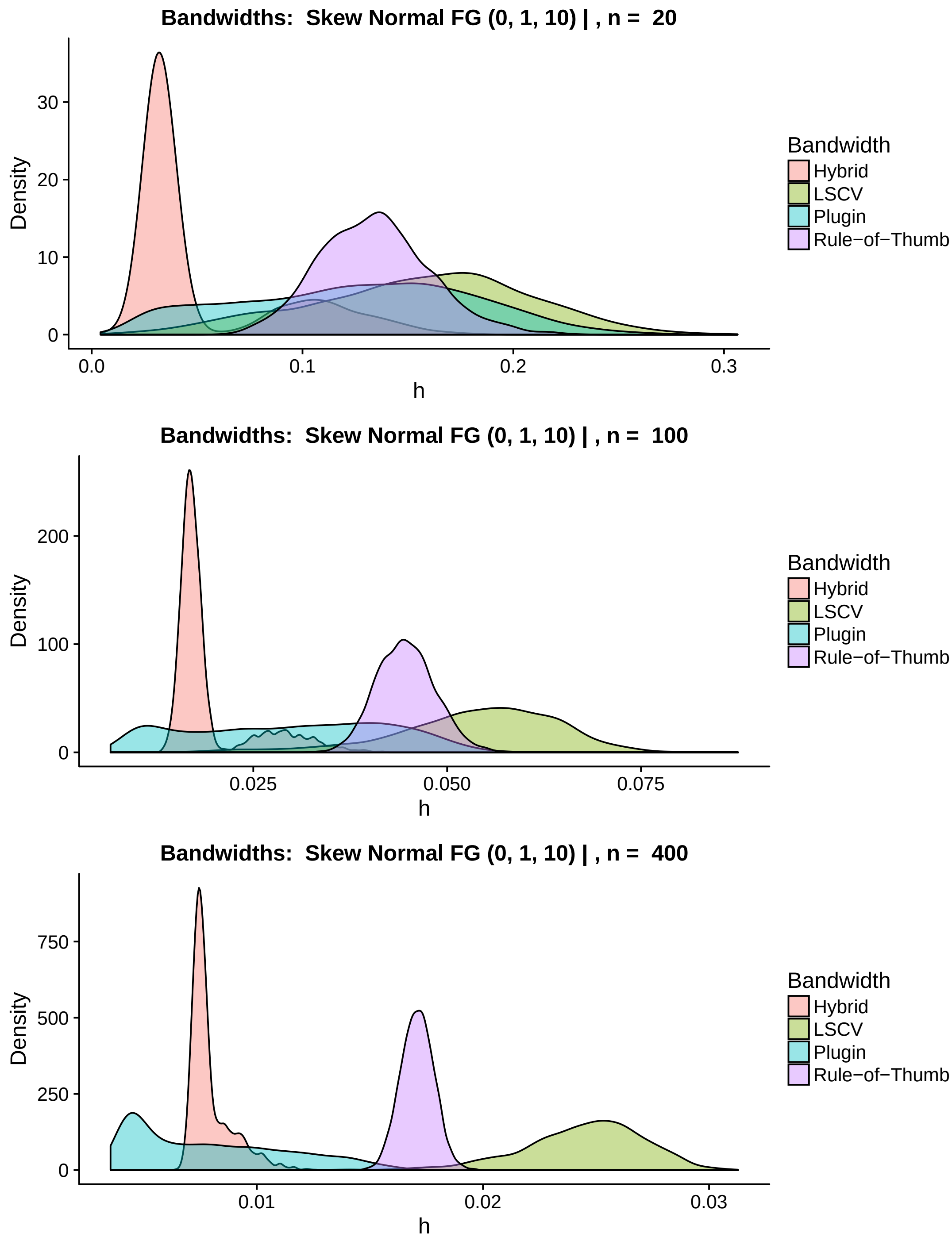


Figure 4.8: Plot of bandwidths for each of the four bandwidth selection methods discussed for the rank weighted estimator at sample sizes 20, 100, 200, and 400 of the $SkewNormFG(0,1,10)$.

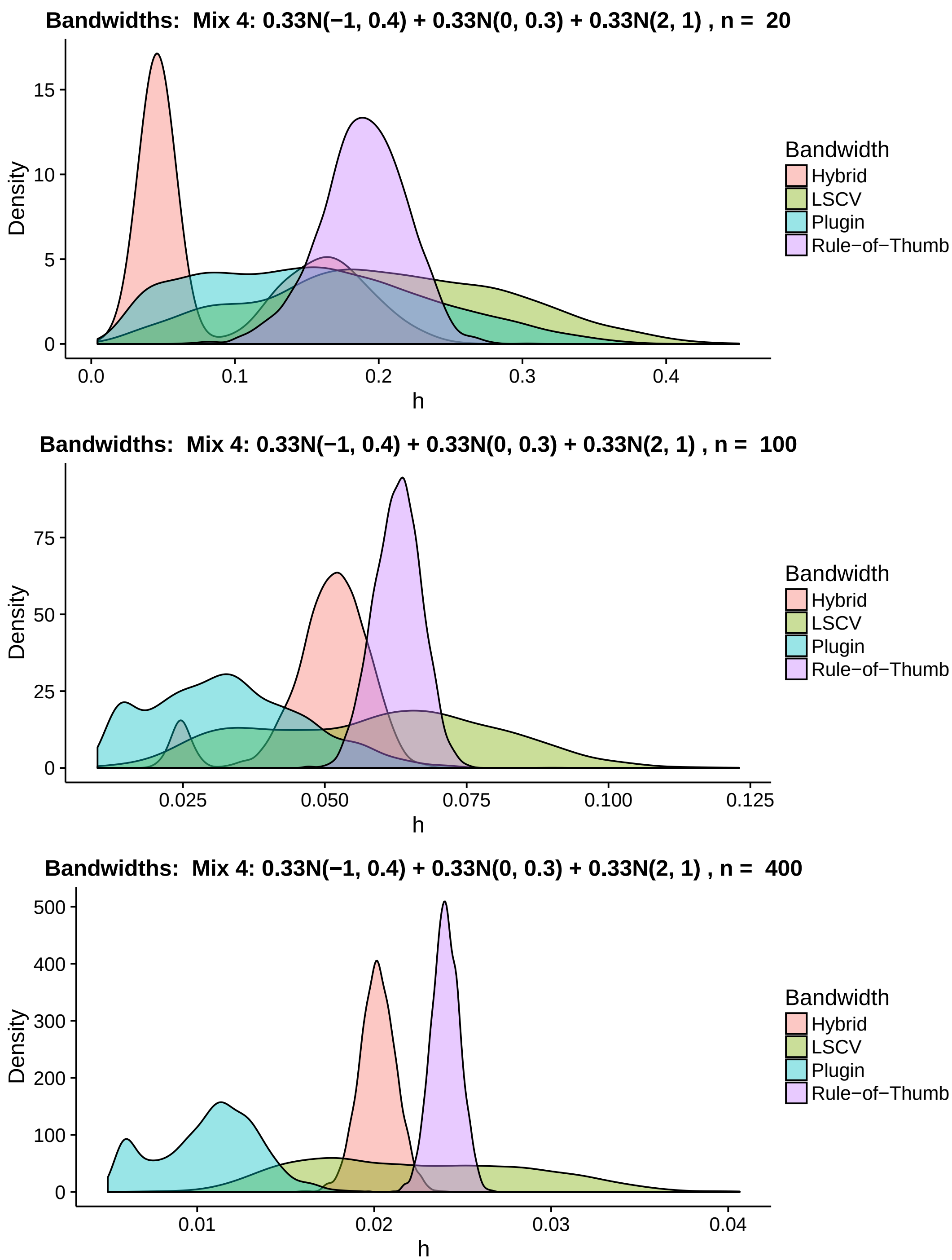


Figure 4.9: Plot of bandwidths for each of the four bandwidth selection methods discussed for the rank weighted estimator at sample sizes 20, 100, 200, and 400 of the normal mixture 4.

Next, a discussion of the MISE results for the density estimators is presented. For the same simulation data, rank-weighted density estimates and standard density estimates will be compared. Again, the rank-weighted MISE results will be displayed for the discussed bandwidth selection methods. These will be displayed along with the MISE results for the same data using the standard density estimator with the following bandwidth selectors: rule-of-thumb, LSCV, and plug-in (Sheather-Jones). To summarize, the following table lists the type of estimator and the bandwidth selection methods considered.

	Estimator	
	Rank-Weighted	**Standard**
Bandwidth Method	Rule-of-Thumb	Rule-of-Thumb
	LSCV	LSCV
	Plug-in	Plug-in (Sheather-Jones)
	Hybrid	

Table 4.4: Summary of density estimation methods and bandwidth selection methods.

Figures 4.10 and 4.11 display the MISEs across the various estimation techniques when simulating from the skew normal AZ distribution. These graphs are for skewness parameters of $\alpha = 2$, and 10. Results for remaining skewness parameters are available in the appendix as skewness parameters between 2 and 10 yielded similar results. Blue colors denote MISE results for the standard estimator. Red colors are the rank-weighted results. At $\alpha = 2$, it can be seen that standard techniques work best, as would be expected since this is very close to a symmetric distribution. When the skewness parameter is tuned to the highest, $\alpha = 10$, the rule-of-thumb method for the rank-weighted estimator takes the lead slightly. Among the various bandwidth selection methods for the rank-weighted estimator, the rule-of-thumb is usually performing best. Under the skew normal AZ distributional setting, this result is very interesting. It would be expected that the hybrid method performs best for the RWKDE. This is because the hybrid method calculates the bandwidth based off estimating the parameters of the skew normal AZ distribution using the sample data.

Figures 4.12 and 4.13 display the MISEs for the skew normal FG distribution with skewness parameter $\xi = 6$, and 10. Results are more favorable for the rank-weighted estimator in this case. At $\xi = 2$, the results are very similar to what occurred with the skew normal AZ distribution at $\alpha = 2$ pictured in Figure 4.10; the standard methods have better MISEs for the most part. However, for $\xi > 2$ the rank-weighted rule-of-thumb starts performing best at all sample sizes, and once the sample size is 50 or above, the rank-weighted LSCV estimator is working better than any of the standard methods as well. At the highest sample sizes, the rank-weighted hybrid estimator is also performing better than any standard estimators.

For the normal mixtures, again, refer to Table 4.1 for the corresponding mixtures and Figure 4.3 to see the shapes of the mixtures, if needed. For mixtures 1 and 2, results are similar so a display of the results for mixture 1 are given in Figure 4.14. It can be seen that for these mixtures, the standard estimator works best. These mixtures were chosen because they do seem to have some factor of right skewness, but it presents in odd ways given the multi-peaked natures. The most favorable mixture for the rank-weighted estimator was mixture 3 where at lower sample sizes of 100 or below, the rank-weighted rule-of-thumb estimator performs best. This can be seen in Figure 4.15. The fact that the RWKDE performs best here is logical given this mixture has a shape that is very similar to that of skew normal distributions that have a larger skewness parameter. Mixture 4 yielded interesting results given its anomalous spiked shape; results are visible in 4.16. The MISEs have strange behaviors with some converging at higher values and others continuously shrinking. The rank-weighted LSCV estimator seems to perform best at the highest sample size.

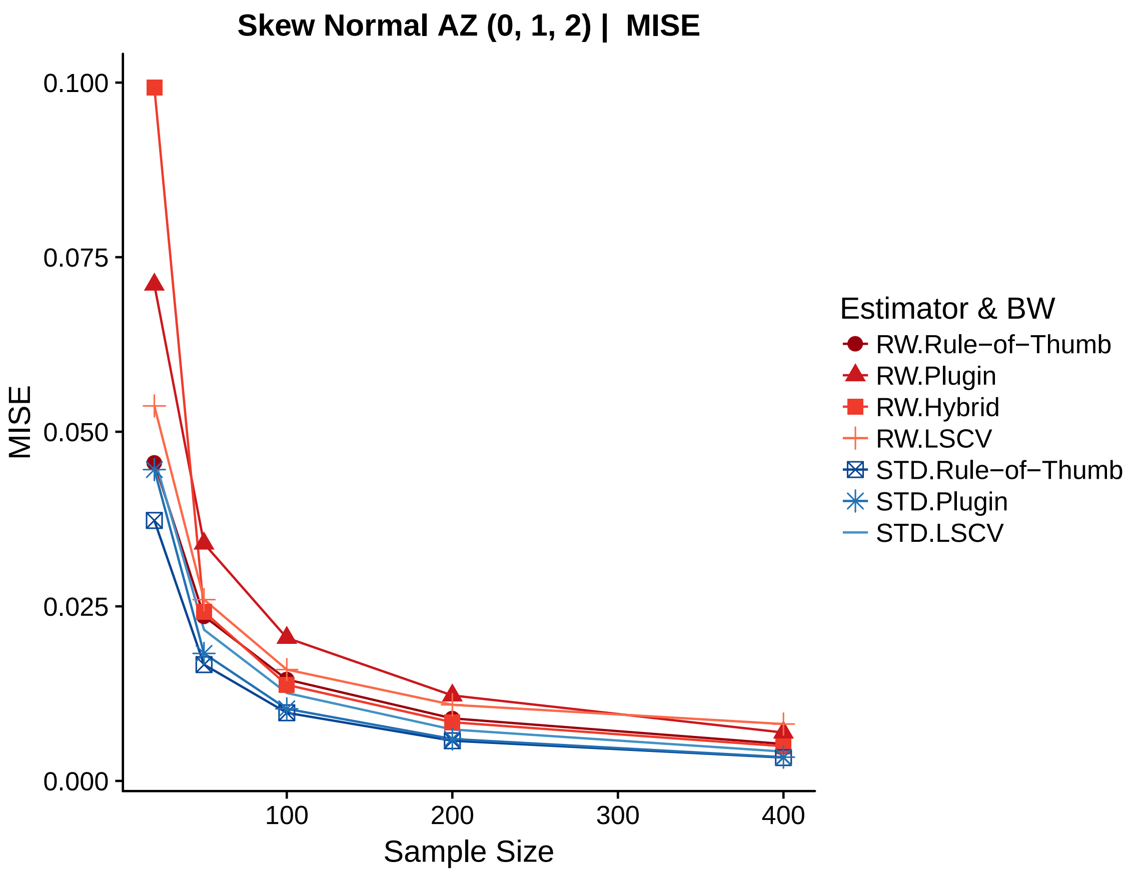


Figure 4.10: MISEs for the $SkewNormAZ(0,1,2)$.

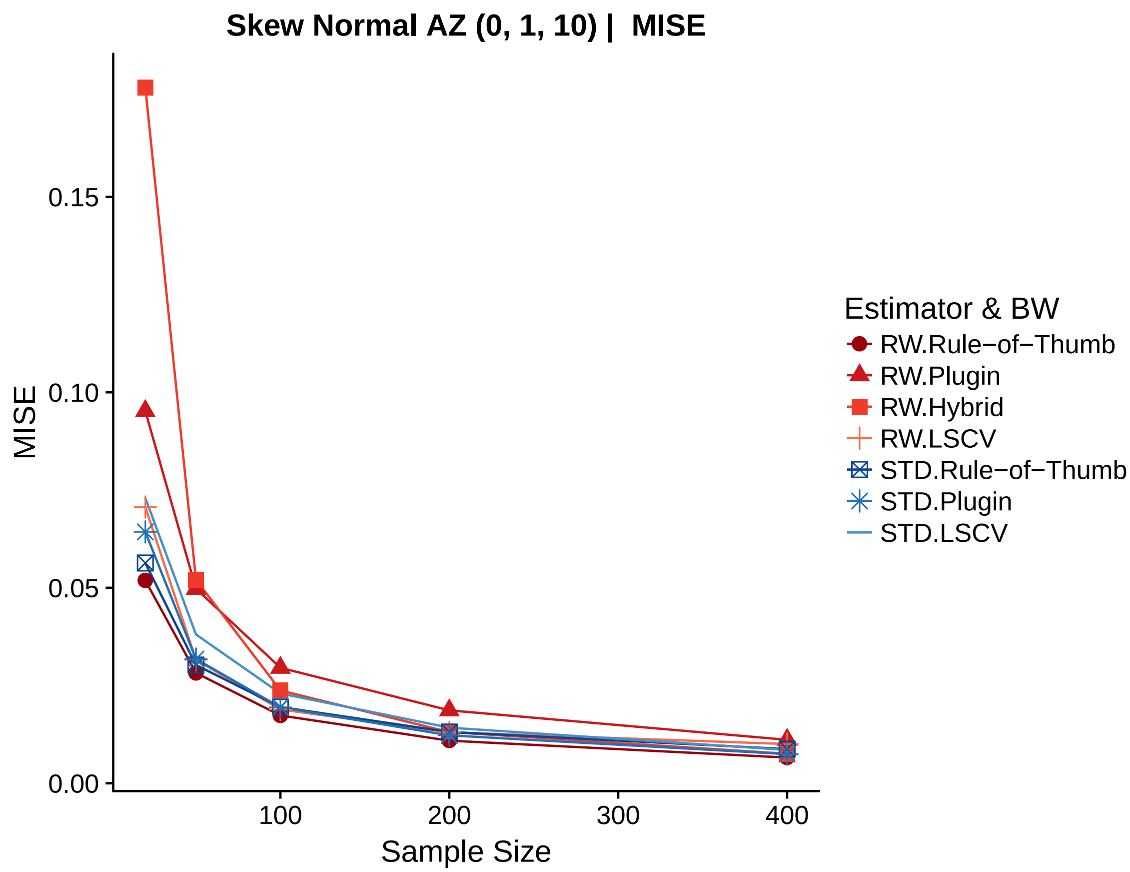


Figure 4.11: MISEs for the $SkewNormAZ(0,1,10)$.

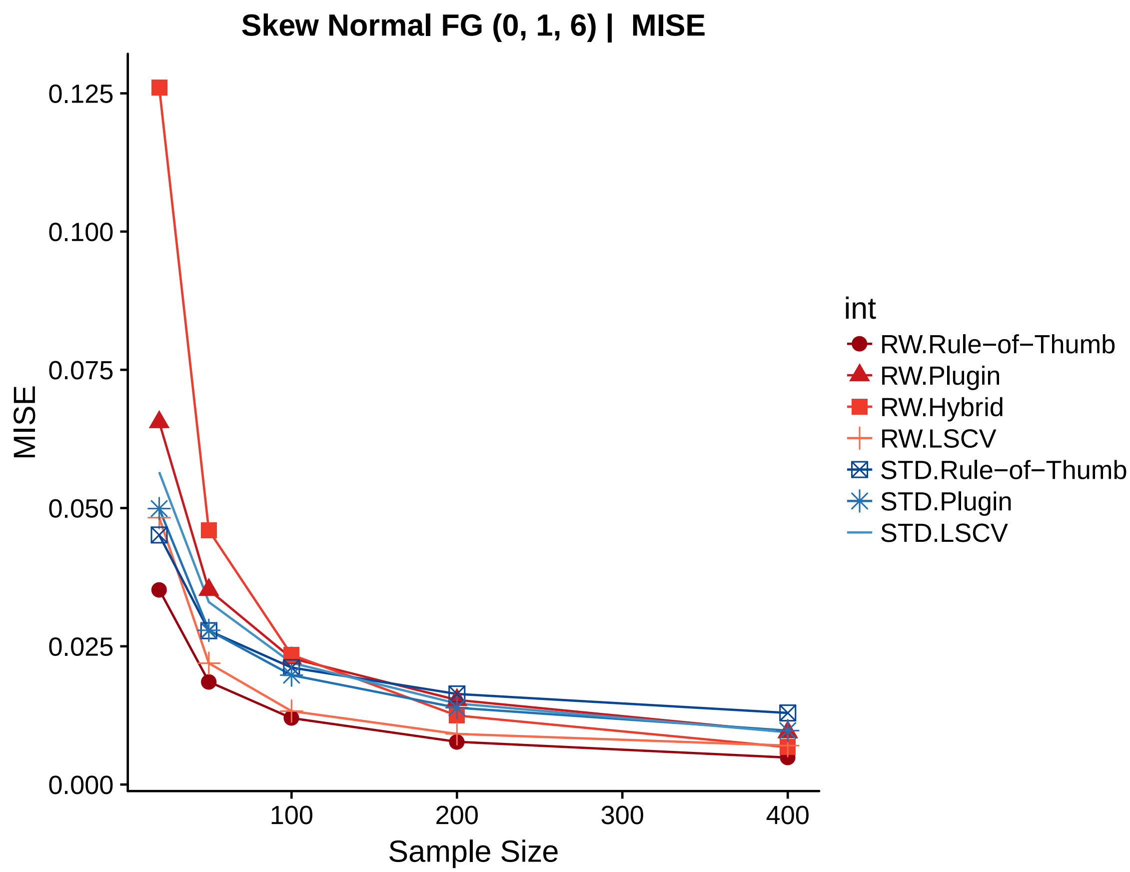


Figure 4.12: MISEs for the $SkewNormFG(0,1,6)$.

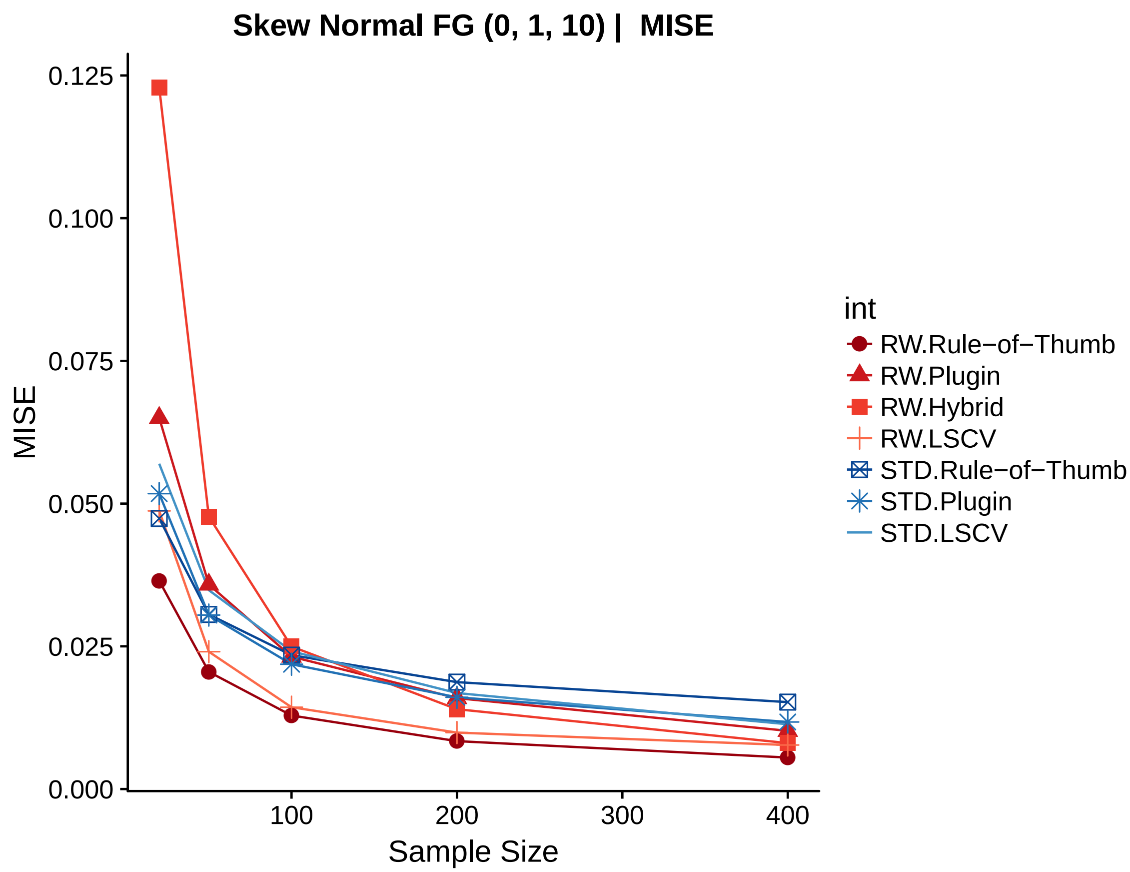


Figure 4.13: MISEs for the $SkewNormFG(0, 1, 10)$.

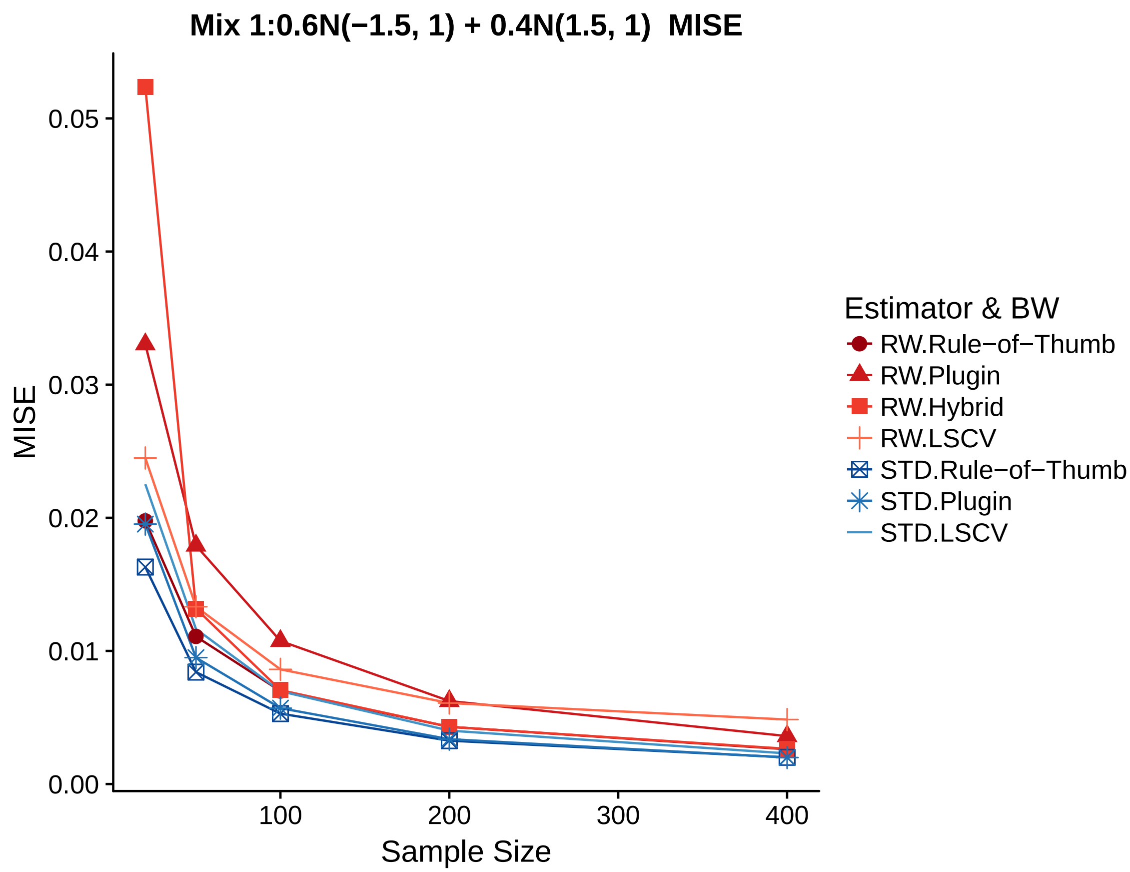


Figure 4.14: MISEs for the normal mixture 1.

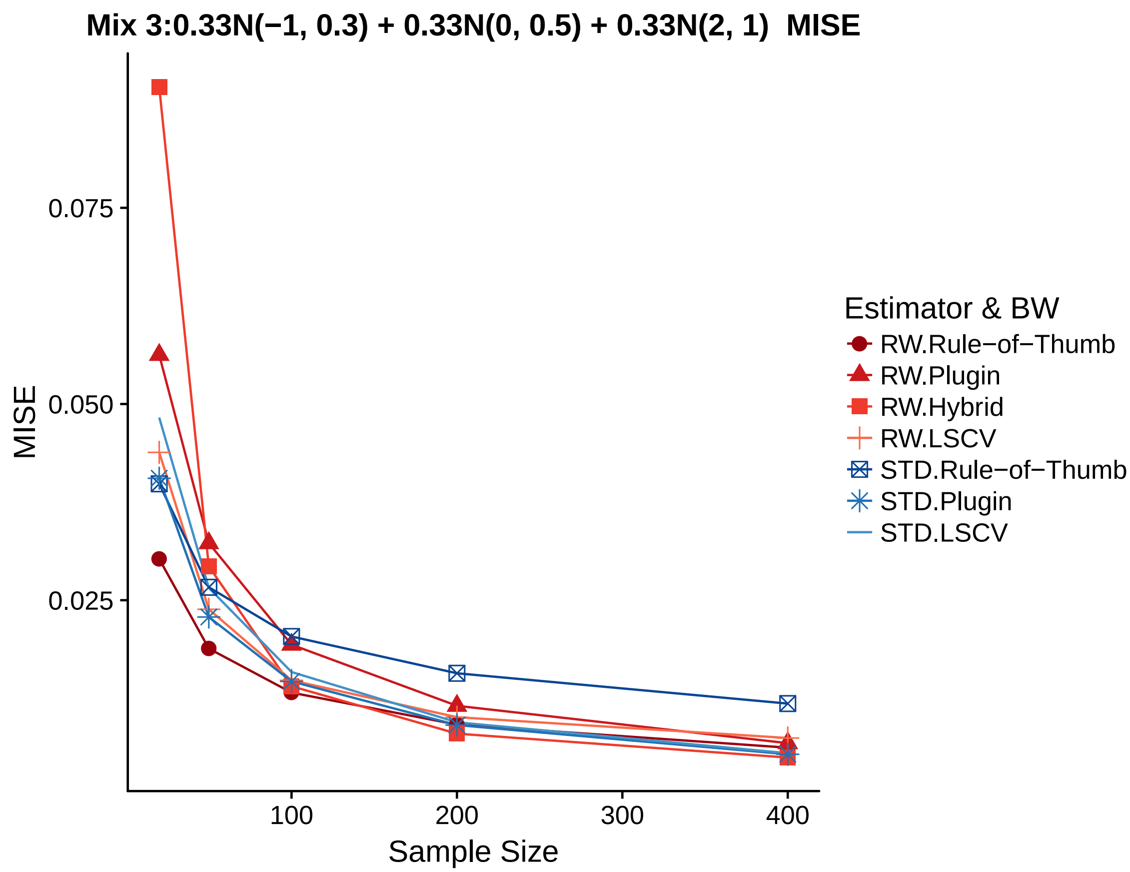


Figure 4.15: MISEs for the normal mixture 3.

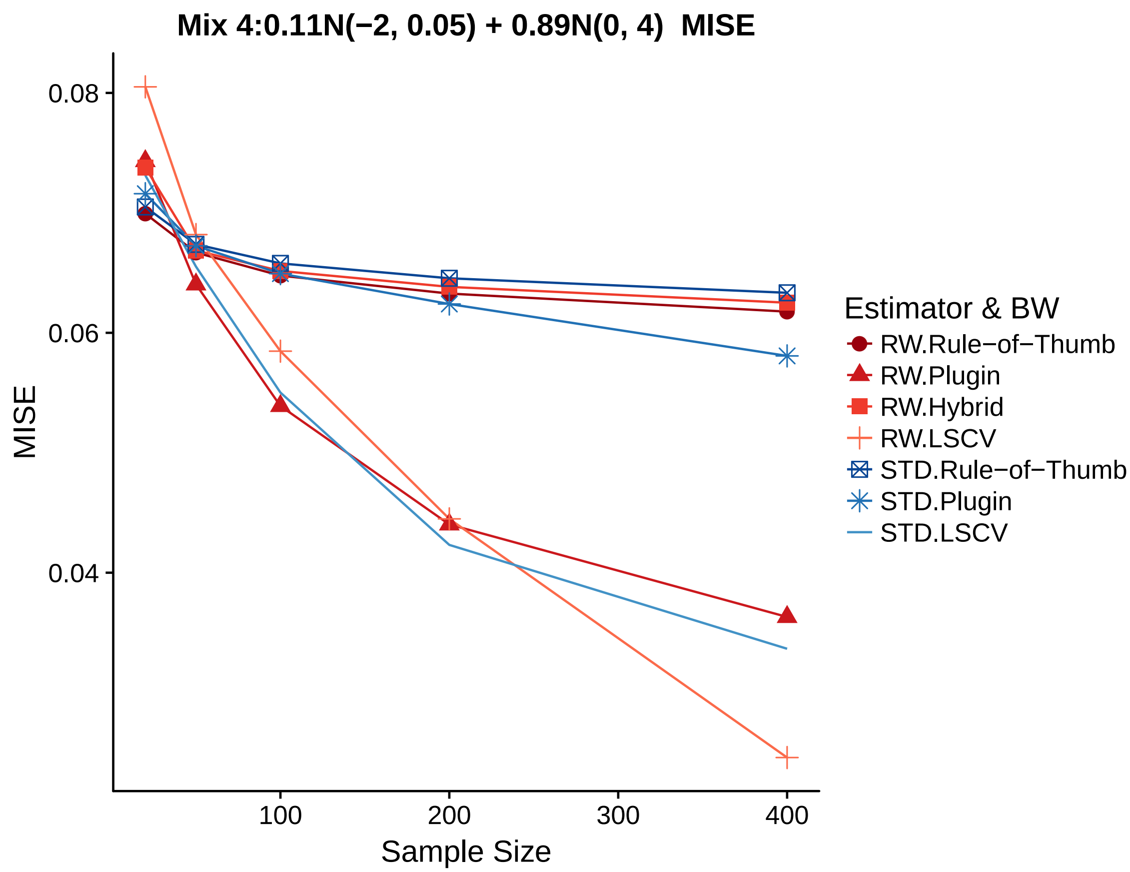


Figure 4.16: MISEs for the normal mixture 4.

4.2 Transformation Kernel Density Estimation Simulations

Next, the results when using the TKDE and RWTKDE will be discussed. Figures 4.17 and 4.18 display the estimated MISEs under the *Lognormal*$(0,0.5)$ and *Weibull*$(1,2)$ settings. It can be seen that the TKDE method of Timsina (2015) out performs the rank weighted version. Results across all other distributional settings were similar for the transformation estimator.

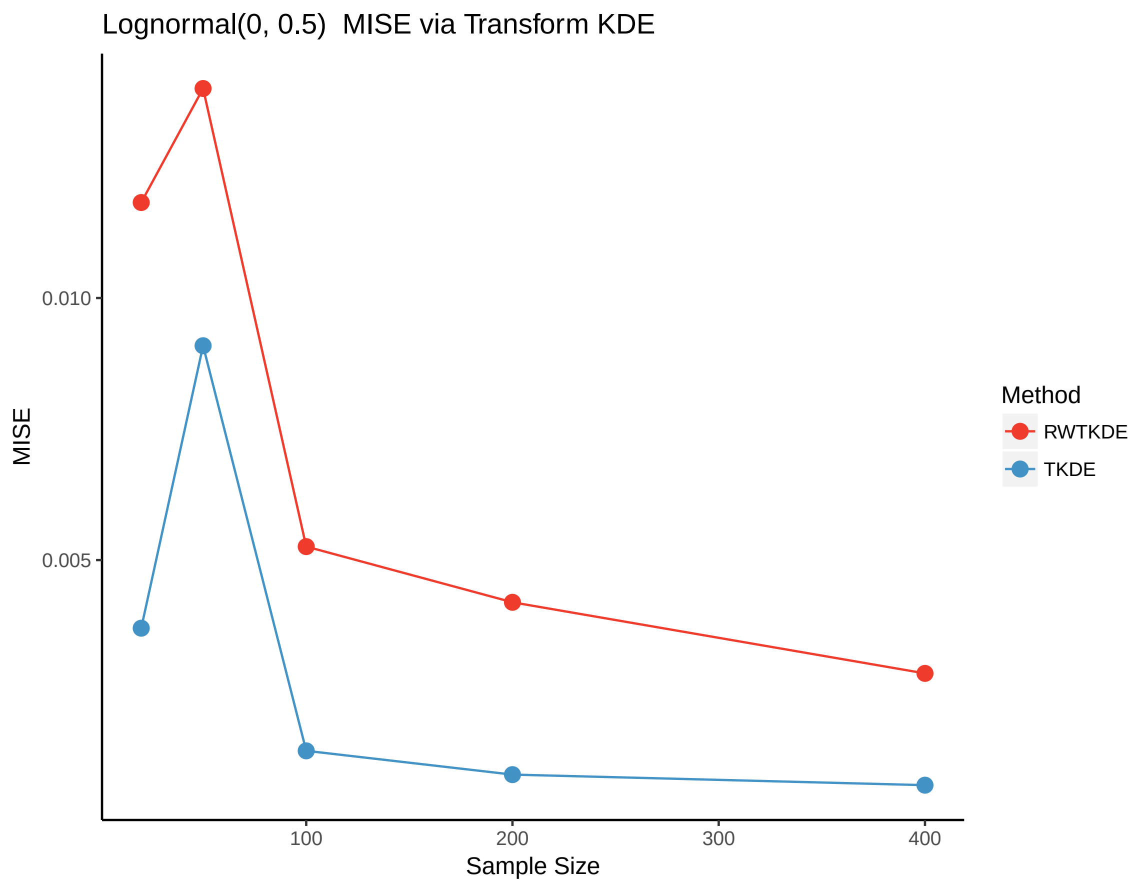


Figure 4.17: Plot of MISEs for the TKDE and RWTKDE estimators under the *Lognormal*$(0,0.5)$ distribution.

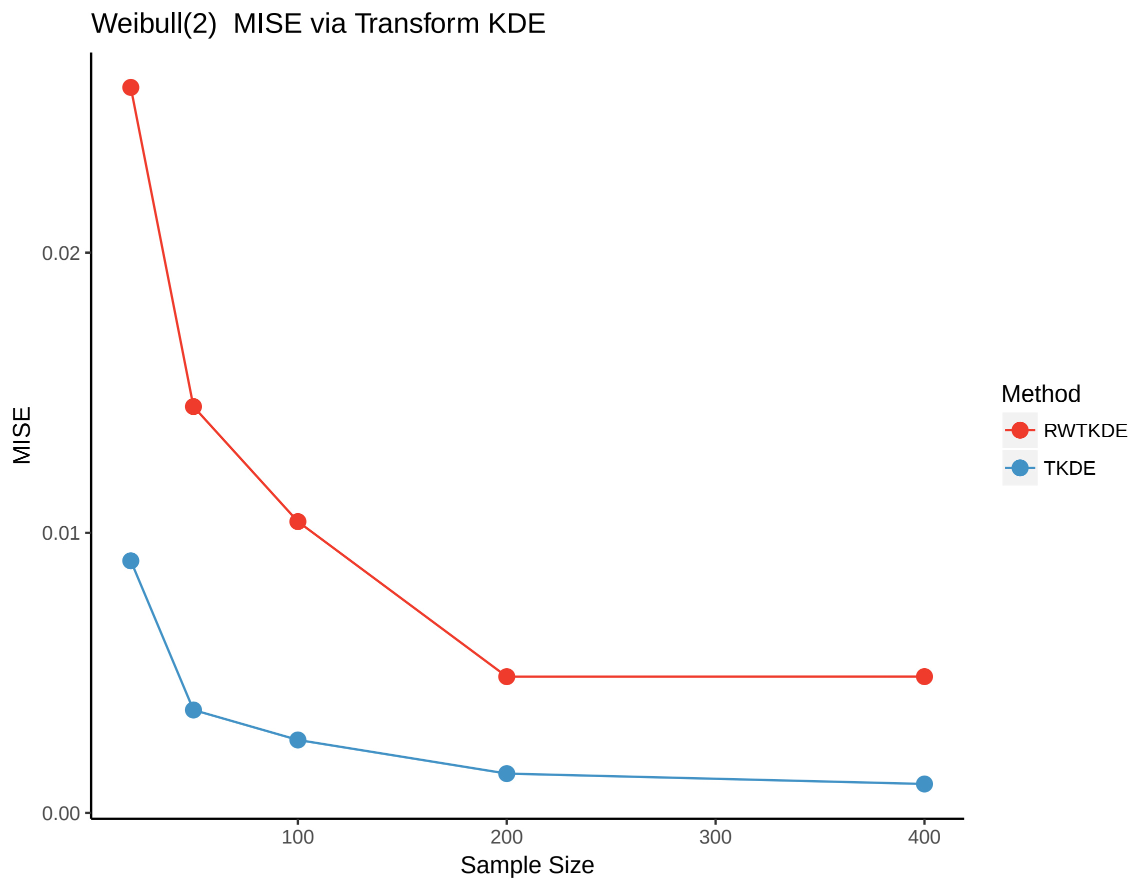


Figure 4.18: Plot of MISEs for the TKDE and RWTKDE estimators under the $Weibull(1,2)$ distribution.

4.3 **Kernel Regression Simulations**

The last set of simulations discussed will be for the rank weighted Nadaraya-Watson estimator (RWNW) compared to the standard Nadaraya-Watson (NW) estimator. Results will be discussed in the order that the functions are listed in the introduction to this chapter. Results not directly discussed can be viewed in the appendix.

For function 1, results for the $Beta(1,1)$, and $Beta(1,7)$ are viewable in Figures 4.19, and 4.20, respectively. Under $Beta(1,1)$, the results are similar for both error settings and no version of either estimator separates from the others. As the second shape parameter increases to 7, it becomes apparent that the standard NW estimator performs better in both error settings; specifically, the LSCV bandwidth is best. Under the first mixture, results in Figure 4.21 show that the rank weighted estimator performs best when the error term 0.5 under all three bandwidth methods once the sample size is greater than fifty. It seems that when the correlation between variables is lower (larger error term), there is more room for the RWNW estimator to perform better. The results under the beta mixture distributions were similar. This pattern will be seen in the results for the remaining functions.

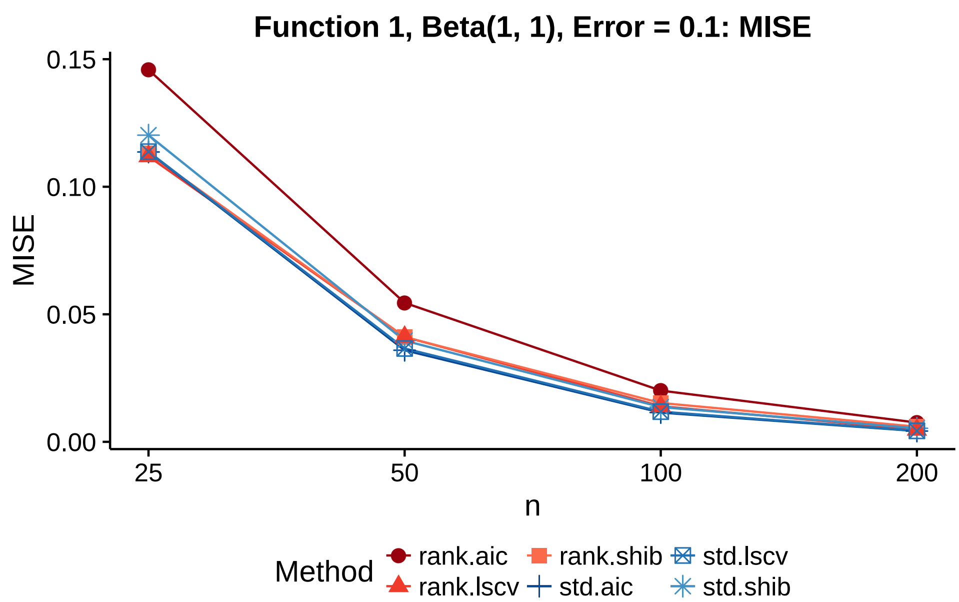
Next, for function 2, the most distinguishing results are visible in Figure 4.22 where the RWNW estimator outperforms the NW estimator under all three bandwidth methods when there is a larger error term. A similar pattern occurs in all other beta distributions including the mixtures. Consistently, the RWNW has a lower MISE using at least two and sometimes three of the bandwidth selection methods.

For function 3, the main results are visible in Figures 4.24, 4.25, and 4.26. Under the $Beta(1,1)$ distribution, all estimators seem to perform equivalently. but under the $Beta(1,3)$ distribution, all RWNW bandwidth methods perform better than the NW methods once the sample size incerases to fifty. Results are very similar under the first mixture distribution. Other distributions showed in general that the RWNW estimator performs best using the LSCV bandwidth method.

Function 4 results were some of the most clear. All can be summarized by the results shown in Figure 4.27. Across all sample sizes in the larger error term case, it can be seen that in the simulations the RWNW estimator was better under all bandwidth methods than the NW estimator. This is the case for all distributions considered including mixtures.

The last function, function 5, has results visible in Figures 4.28, 4.29, 4.30. Under the $Beta(1,1)$ distribution for the x variable, the results favor the standard NW estimator, specifically the LSCV bandwidth choice under both error terms. Results are rather anomalous under the $Beta(1,7)$ distribution where the MISE curve is concave instead of convex like in most other cases; when the error term is large, the RWNW performs best under all bandwidth choices until the largest sample size where the NW estimator has the lower MISE. In the first mixture, the RWNW estimator seems to perform about the same as the LSCV NW estimator, but all RWNW bandwidth choices have a lower MISE than the Shibata and AIC bandwidth choices of the NW estimator.

A summarization of results obtained on both the RWKDE and and RWNW estimators will be given in the next chapter. In general, there are cases favorable to the rank weighting. An attempted analysis on what is causing the rank weighted or standard estimators to be better will be given along with potential ideas for future research.



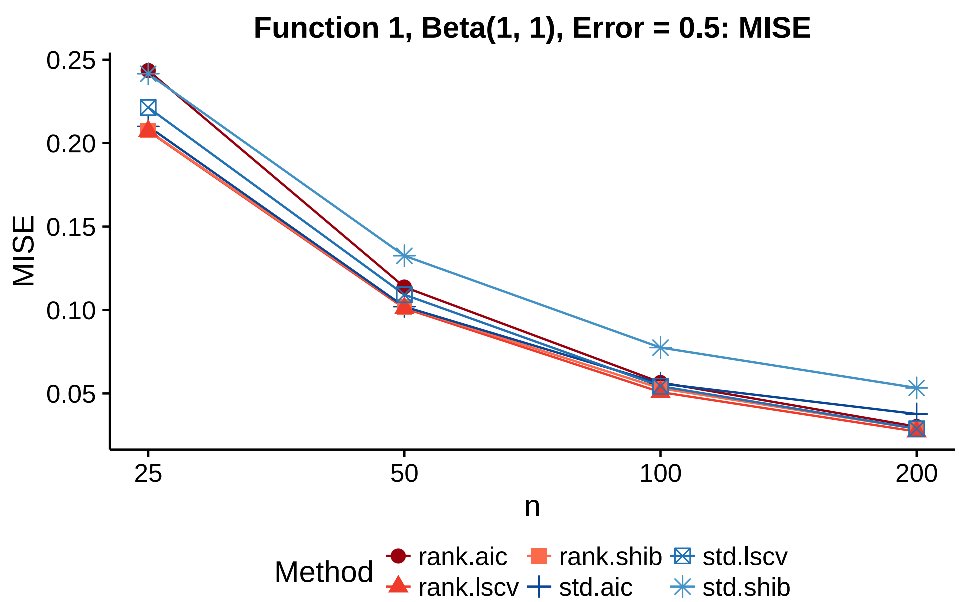


Figure 4.19: MISEs comparing the estimation abilities of the NW estimator and RWNW estimator under the indicated bandwidth methods in the legend. Results are for function 1 under the Beta(1, 1) distribution.

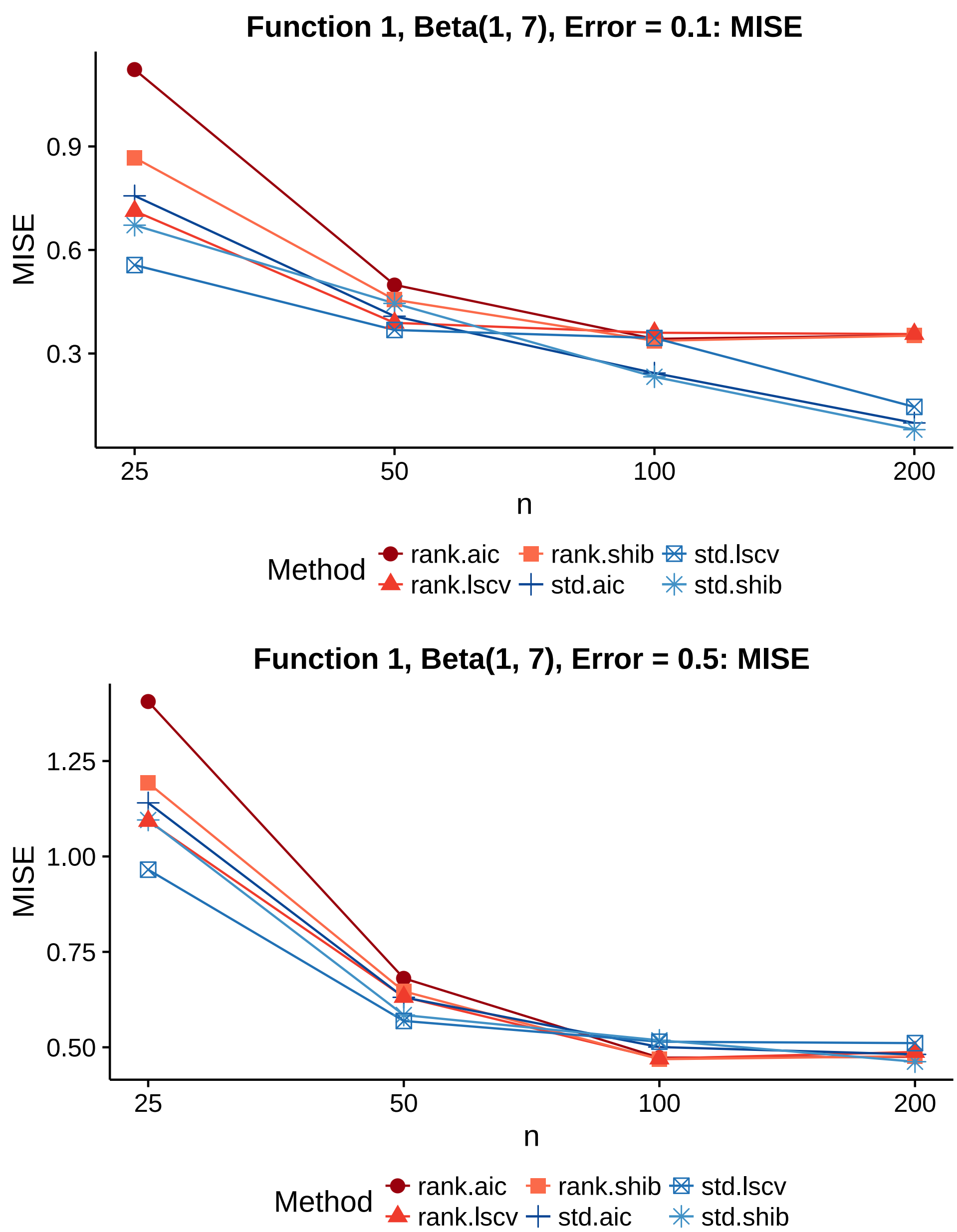
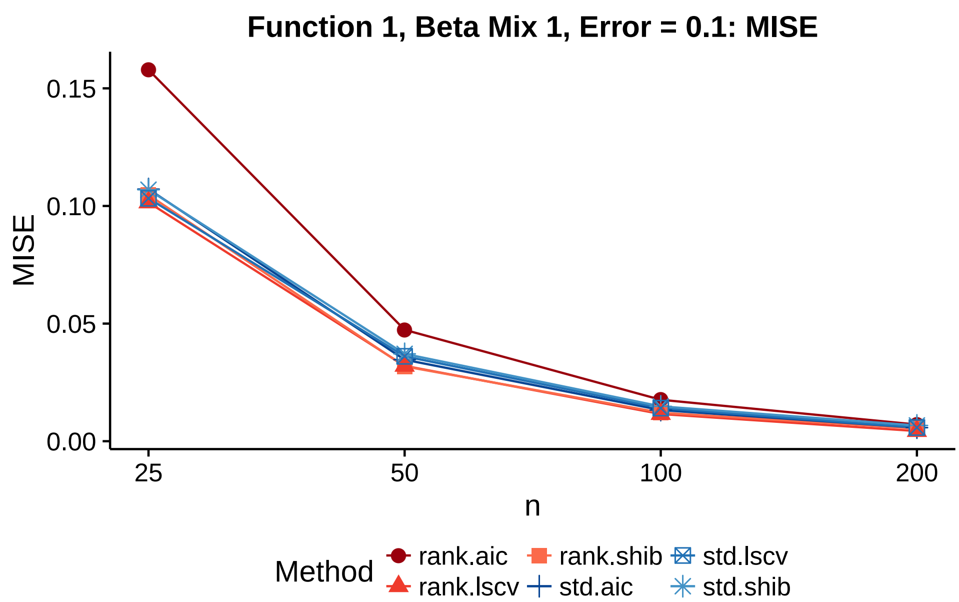


Figure 4.20: MISEs comparing the estimation abilities of the NW estimator and RWNW estimator under the indicated bandwidth methods in the legend. Results are for function 1 under the Beta(1, 7) distribution.



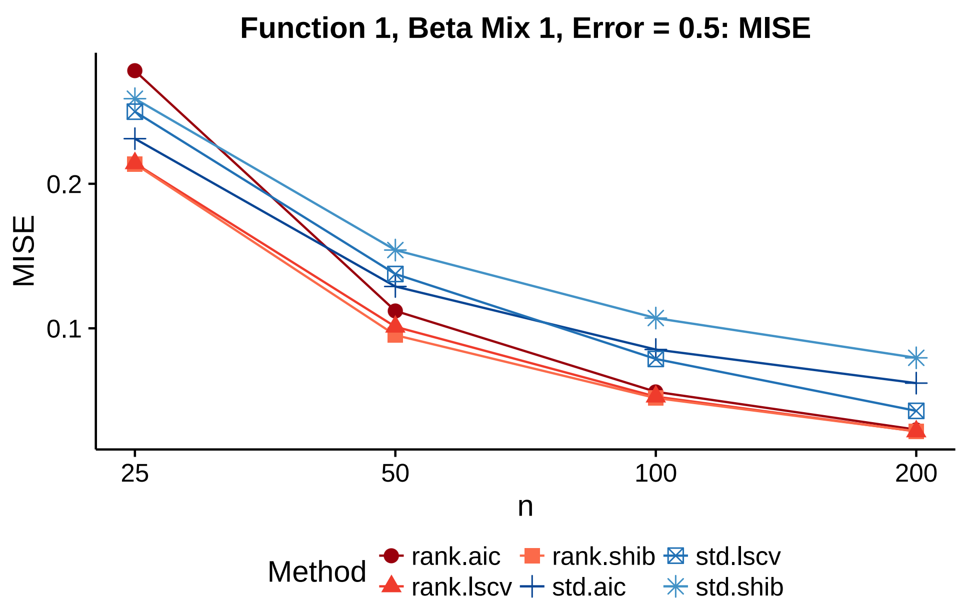
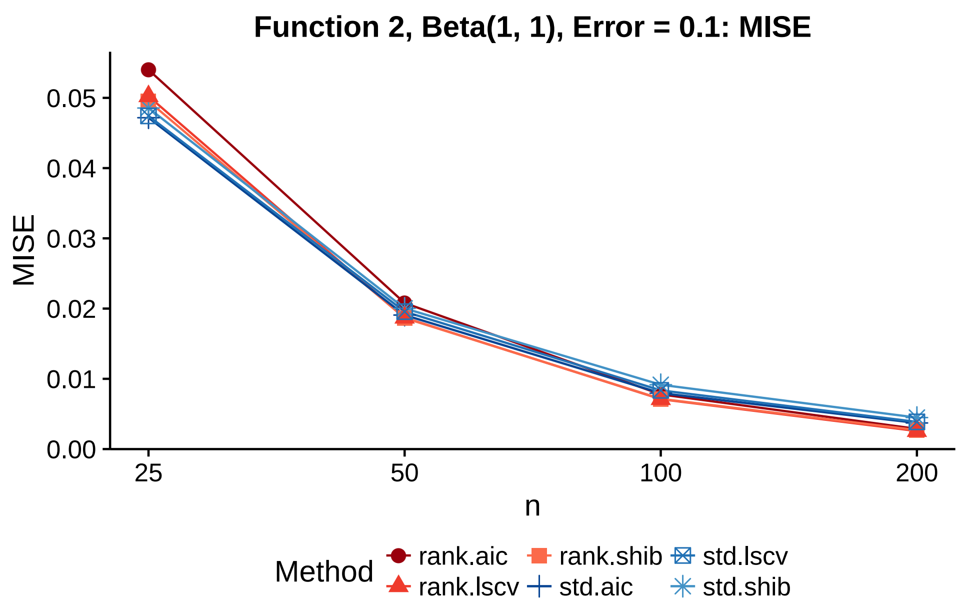


Figure 4.21: MISEs comparing the estimation abilities of the NW estimator and RWNW estimator under the indicated bandwidth methods in the legend. Results are for function 1 under the beta mixture 1.



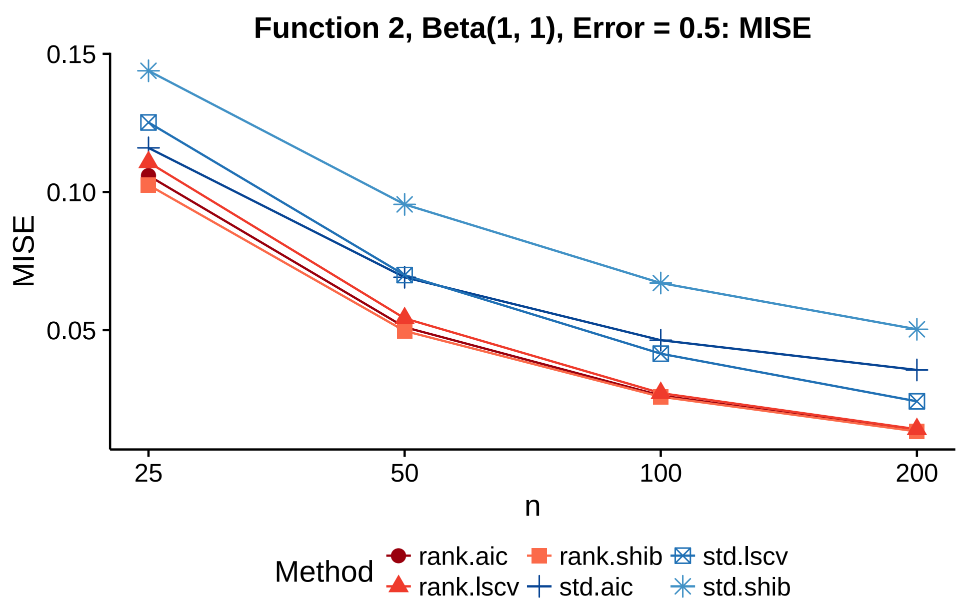
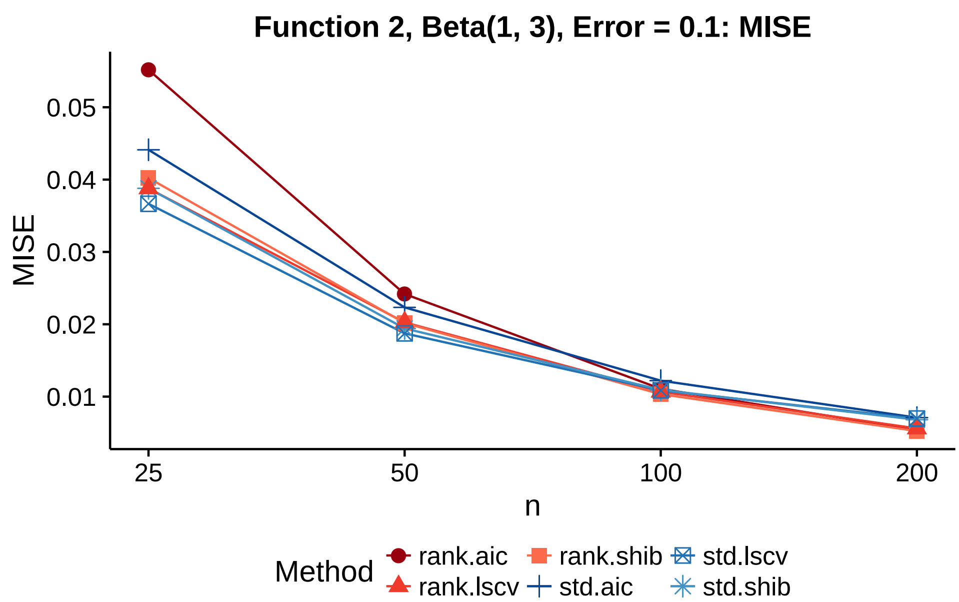


Figure 4.22: MISEs comparing the estimation abilities of the NW estimator and RWNW estimator under the indicated bandwidth methods in the legend. Results are for function 2 under the Beta(1, 1) distribution.



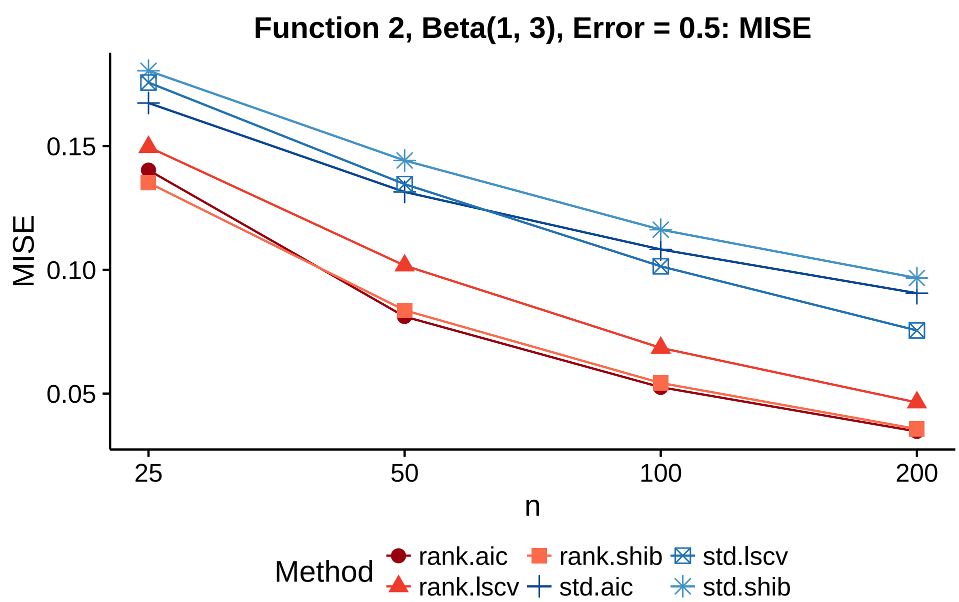
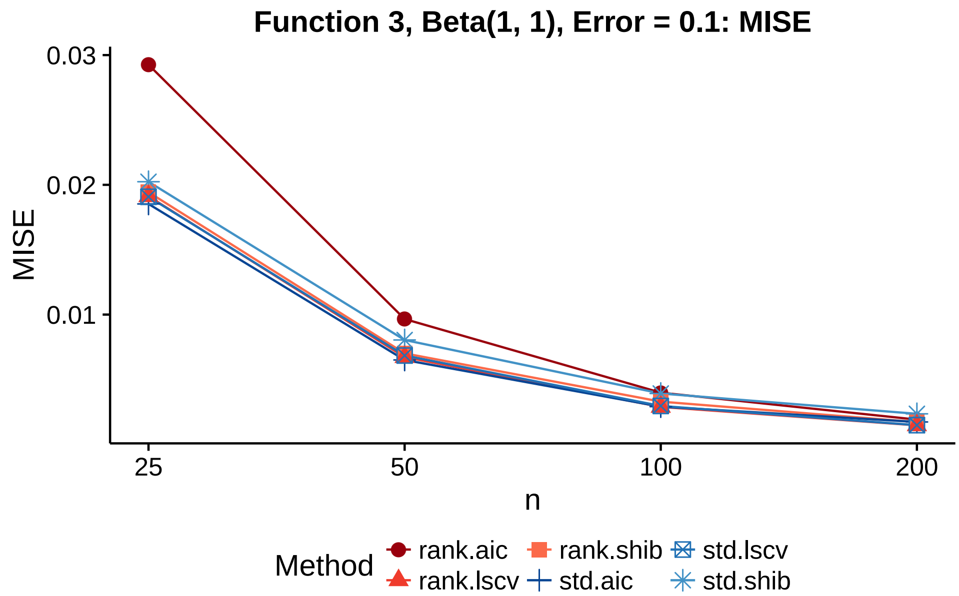


Figure 4.23: MISEs comparing the estimation abilities of the NW estimator and RWNW estimator under the indicated bandwidth methods in the legend. Results are for function 2 under the Beta(1, 3) distribution.



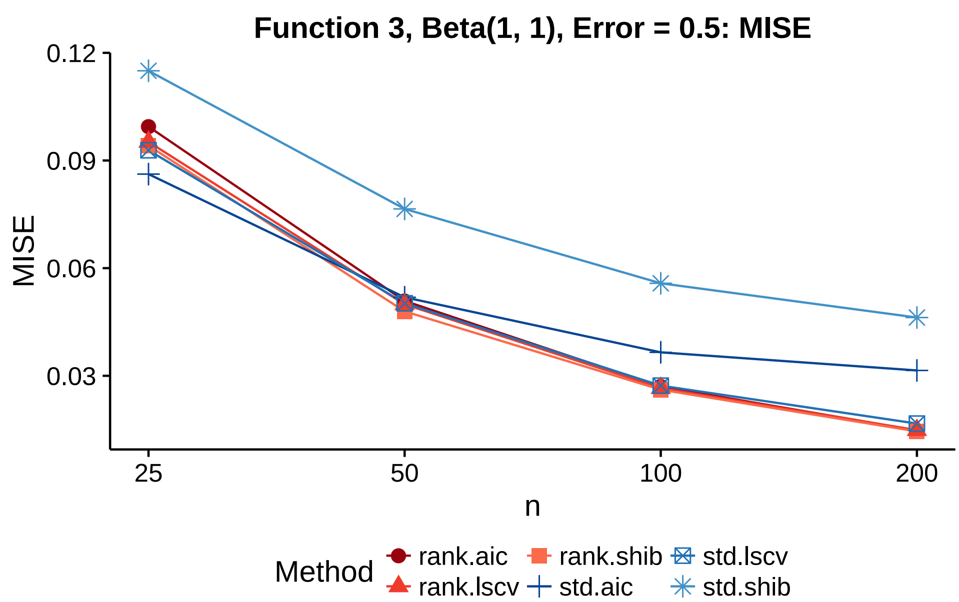
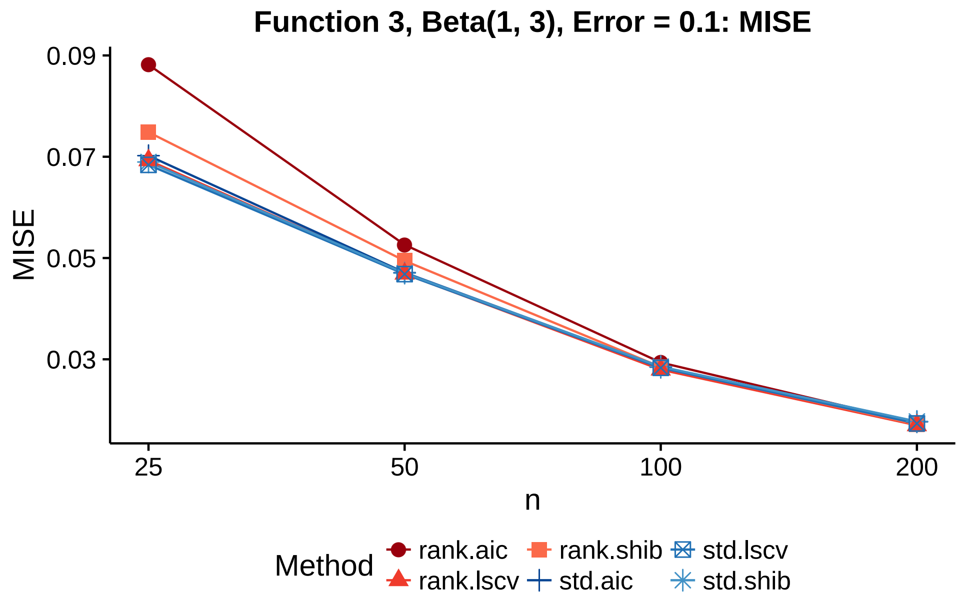


Figure 4.24: MISEs comparing the estimation abilities of the NW estimator and RWNW estimator under the indicated bandwidth methods in the legend. Results are for function 3 under the Beta(1, 1) distribution.



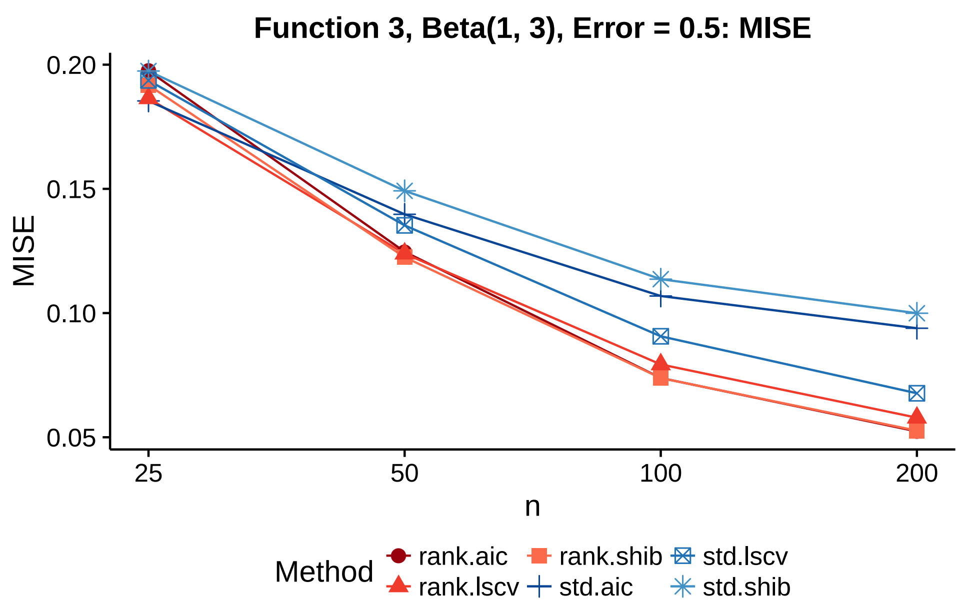


Figure 4.25: MISEs comparing the estimation abilities of the NW estimator and RWNW estimator under the indicated bandwidth methods in the legend. Results are for function 3 under the Beta(1, 3) distribution.

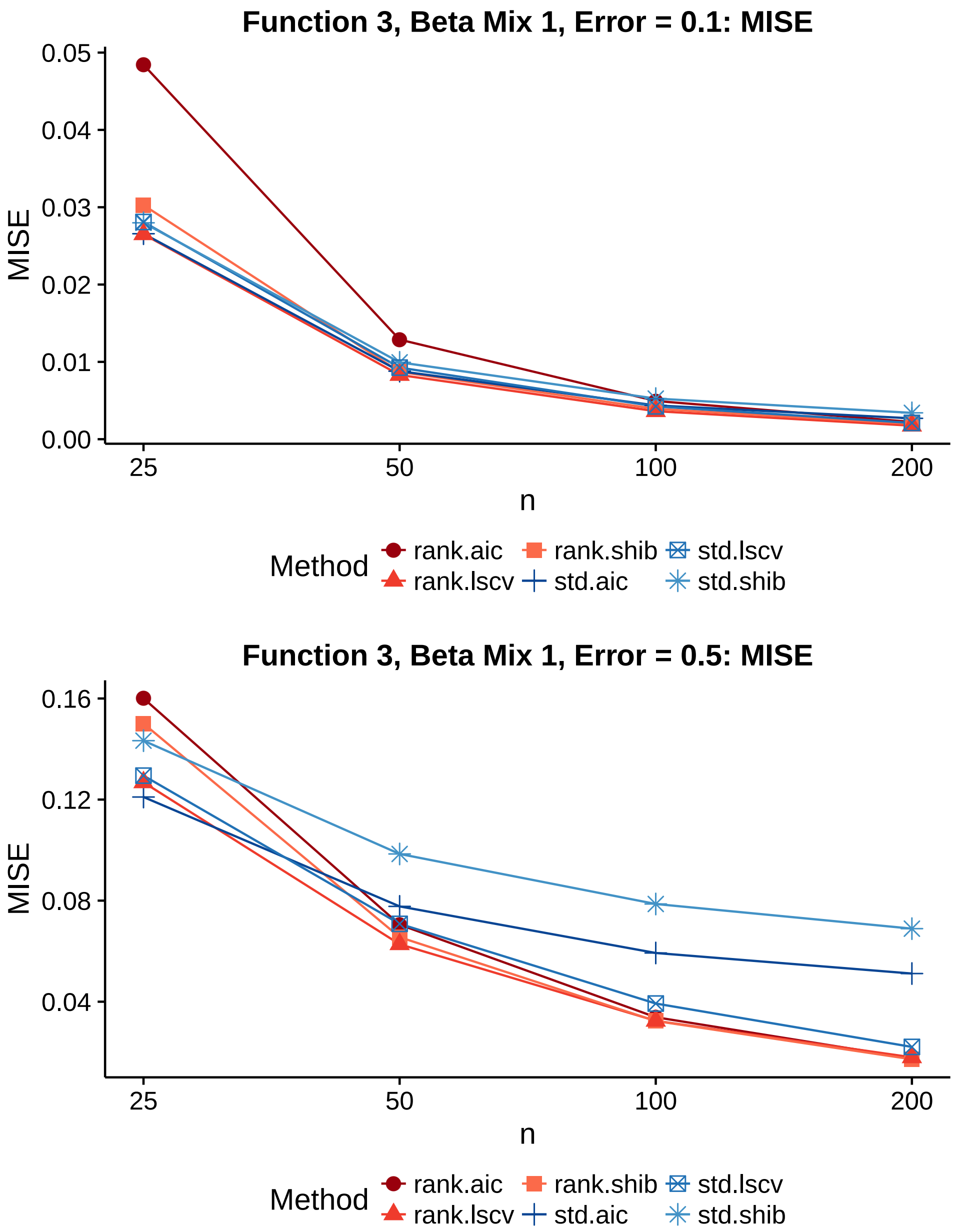


Figure 4.26: MISEs comparing the estimation abilities of the NW estimator and RWNW estimator under the indicated bandwidth methods in the legend. Results are for function 3 under the beta mixture 1.

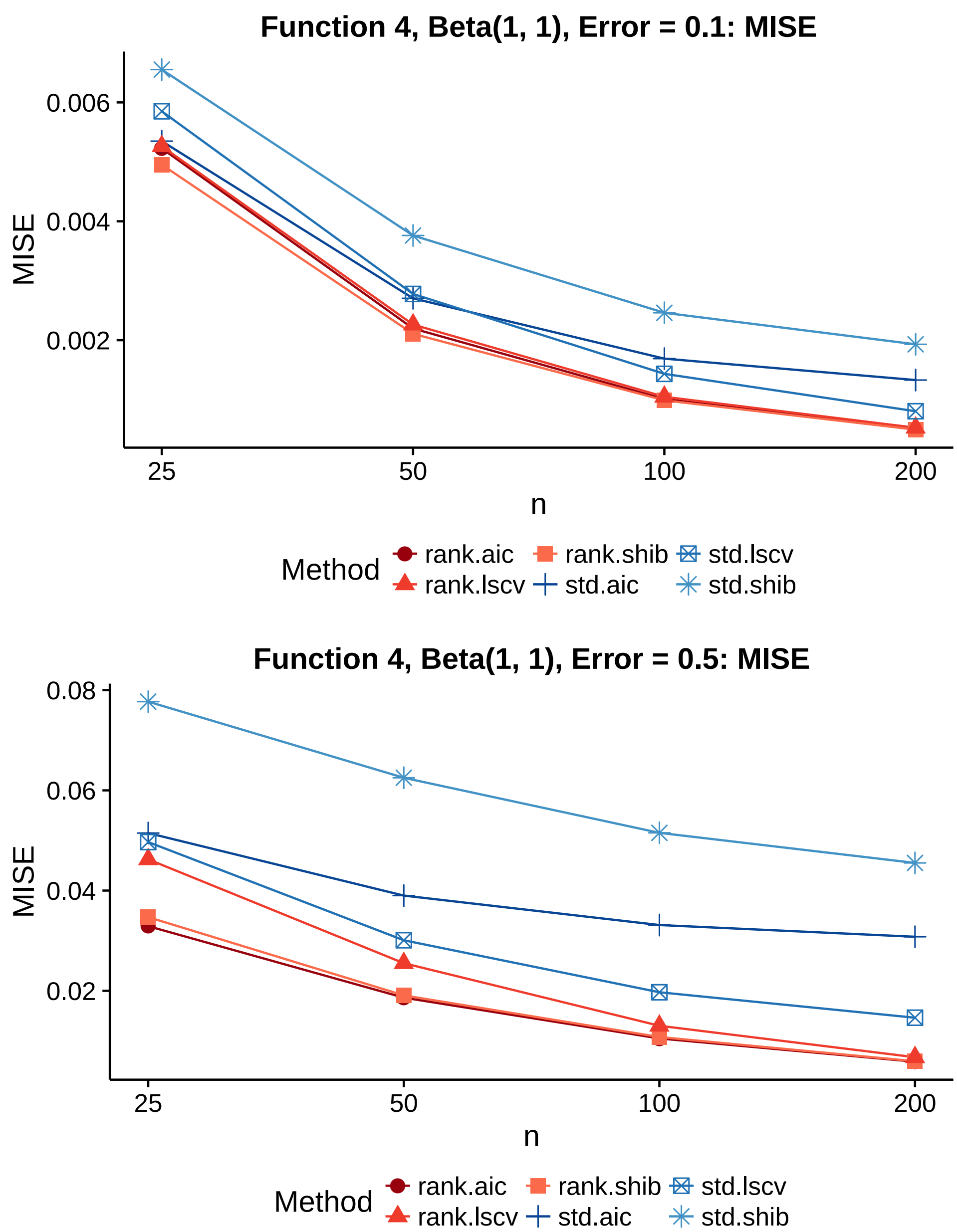
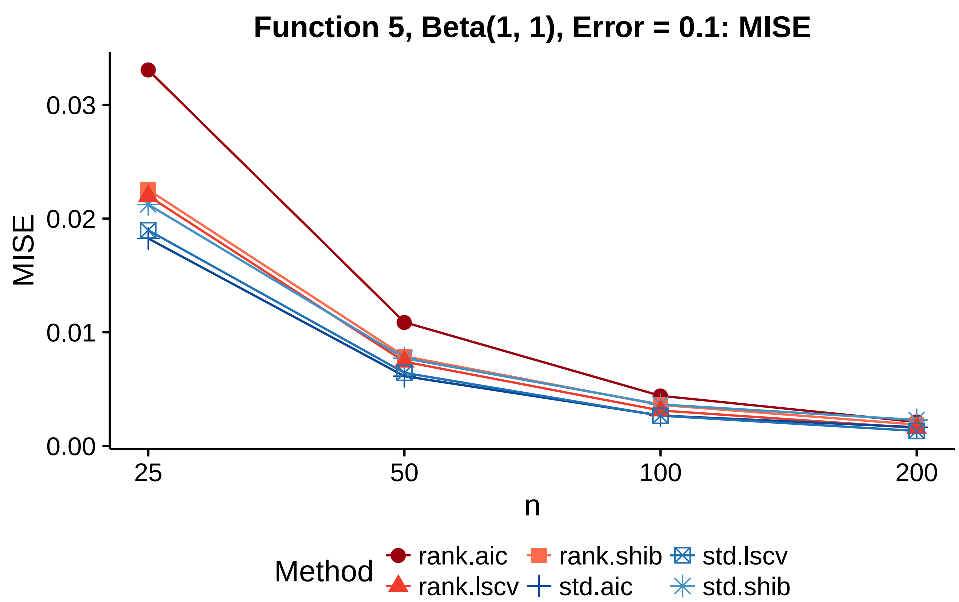


Figure 4.27: MISEs comparing the estimation abilities of the NW estimator and RWNW estimator under the indicated bandwidth methods in the legend. Results are for function 4 under the Beta(1, 1) distribution.



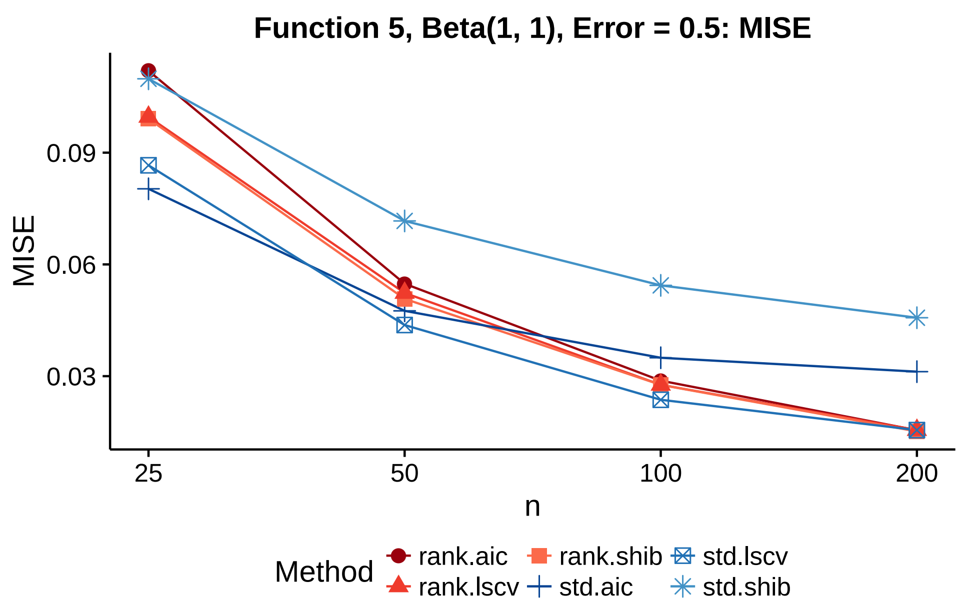
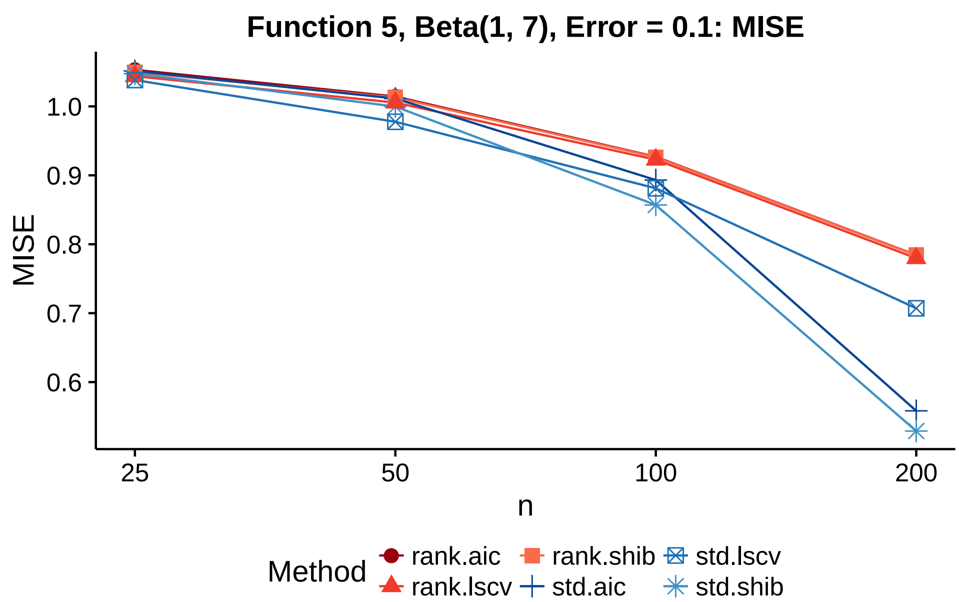


Figure 4.28: MISEs comparing the estimation abilities of the NW estimator and RWNW estimator under the indicated bandwidth methods in the legend. Results are for function 5 under the Beta(1, 1) distribution.



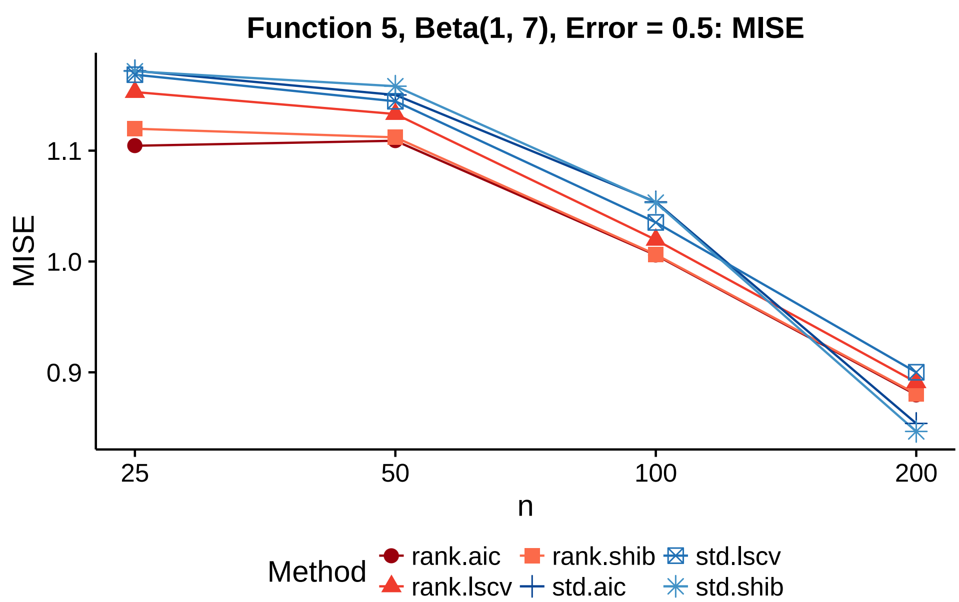


Figure 4.29: MISEs comparing the estimation abilities of the NW estimator and RWNW estimator under the indicated bandwidth methods in the legend. Results are for function 5 under the Beta(1, 7) distribution.

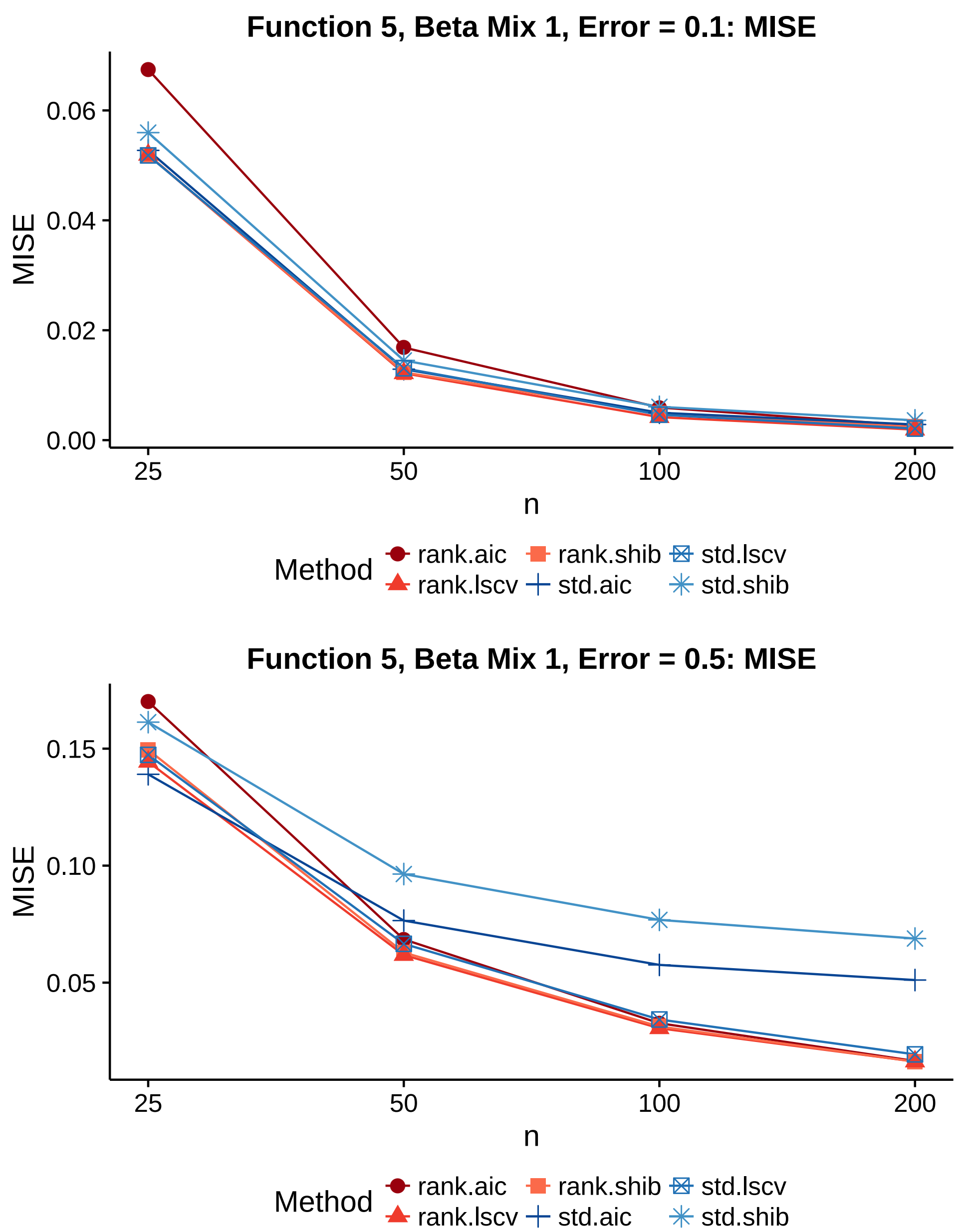


Figure 4.30: MISEs comparing the estimation abilities of the NW estimator and RWNW estimator under the indicated bandwidth methods in the legend. Results are for function 5 under the beta mixture 1.

CHAPTER V

CONCLUSIONS AND FUTURE WORK

The purpose of this work is to try to extract extra information from the data by including the use of ranks in estimation of densities and regression functions. In this paper, the utilization of ranks simultaneously with observations was done under two estimation settings in nonparametric statistics: kernel density estimation and kernel regression. A set of conclusions summarizing the results from simulations will be given, then a set of ideas for future research will be presented.

5.1 **Conclusions**

There were situations where rank-weighting performed better than the standard estimation techniques for both density estimation and kernel regression. First, a discussion of the performance of the RWKDE bandwidth selection methods will be discussed. A variety of bandwidth selection methods were examined for the RWKDE setting. The most stable bandwidth choice was that of the rule-of-thumb. The rule-of-thumb bandwidth method had a similar spread of bandwidths across all simulated distributions for samples. In contrast to this is the hybrid method. Typically, the hybrid method seemed to have a bimodal structure in the distribution of potential bandwidth choices for lower sample sizes. This meant the bandwidth choice was highly variable, which inflates the MISE. The LSCV and plug-in bandwidth methods had much higher variability in bandwidth choice than initially hoped. These results indicate that the rule-of-thumb estimator may perform best in terms of MISE in simulations, which is discussed next.

The most important issue with the RWKDE is to establish when this estimator is superior to the standard kernel estimator. Obviously, the RWKDE did not always perform best. In the case of both skew normal distributions, when the skewness parameter of each distribution was in the lowest setting, the standard KDE performed best. This would be because for low values of the skewness parameters, the skew normal distributions are very close in structure to the normal distribution, and thus, nearly symmetric. At higher values for the skewness parameters in both skew normal distributions, the RWKDE using the rule-of-thumb bandwidth selection technique performed better than any bandwidth techniques in the standard KDE. The LSCV bandwidth estimator in the RWKDE case performed similarly, but slightly worse than the rule of thumb. A general trend in the simulations was that if any bandwidth selection method was performing best in the RWKDE setting, it was mostly the rule-of-thumb method. The main exception to this was in the case of the fourth mixture of normal distributions where plug-in and LSCV methods performed better for the RWKDE, but this is attributed to the shape of the fourth mixture. Overall, simulations show that as a distribution appears more skewed, it is the RWKDE that performs best using the rule-of-thumb bandwidth method. It seems that the instability and increased variability of the other bandwidth methods prevent them from working as well as the rule-of-thumb.

As briefly discussed, a transformation technique explored by Marron and Ruppert (1994) and Timsina (2015) was considered for use along with the rank-weighting in the density estimation scenario. Results indicated that the non-rank-weighted TKDE performed better than the RWTKDE. A possible explanation for this involves the structure of the TKDE:

$$\widehat{f}_T(x) = \frac{g'(x)}{nh} \sum_{i=1}^{n} K\left(\frac{g(x) - g(X_i)}{h}\right).$$

The transformation explored in simulations was the log function. Its derivative is $1/x$.This mean that the applied bandwidth when estimating the density at a point x is $\frac{1}{xh}$. As x increases, this bandwidth increases. In the rank-weighted setting, the ranks further increase the size of the bandwidth: $\frac{1}{\sqrt{ixh}}$. This results in extremely small bandwidths at low ranks of the sample and in extremely large bandwidths at high ranks, resulting in poorer performance when compared to the non-rank-weighted estimator. Therefore, the recommendation is to use the TKDE instead of RWTKDE. It may be that if other bandwidth methods are explored, this conclusion may change, but initial results do not merit further exploration.

Next, is a brief summary of the kernel regression results. One of the most clear results is that when the error term is larger, the rank-weighted estimator will usually perform better or equivalently to the standard estimator. The main exception to this was when the y values were generated using function 5, which had a very mild structure. Initially it was assumed that when the x values are generated using a $Beta(1,1)$ distribution, the standard NW estimator would perform best and as the second shape parameter increased, the RWNW estimator would perform better. In functions 2 and 4, the RWNW estimator performed better even in the $Beta(1,1)$ setting for the x values. Overall, if the distribution of the x values is skewed and there seems to be more noise in the data, the RWNW estimator was performing better. No bandwidth selection method really separated itself in terms of performance from others, although, the AIC bandwidth method seemed to have the worst performance out of the three when samples sizes were at their smallest. An initial suggestion may be to use the Shibata method for kernel regression bandwidths in the rank-weighted setting.

Overall, there is potential for the rank-weighted estimator in both the density estimation case and the regression case, although there is definitely room for improvements and innovations.

5.2 Future Work

There are a few clear options for future research for the rank-weighted density estimator. One is further exploration of bandwidth techniques. It seems unsatisfactory that a rule-of-thumb method that simply assumes a normal distribution and calculates an optimal bandwidth for the normal distribution performs best. Just as the Sheather-Jones plug-in method for bandwidth selection is considered a superior choice in the standard KDE, it was hoped that the plug-in method in the RWKDE would be a better option. One possible issue is that the plug-in method for the RWKDE used LSCV to estimate the derivatives of the underlying density $f(x)$. The LSCV method is often criticized for its high variability which negatively impacts the MISE. Other methods for estimating the functional given by (II.23) need to be explored in order to improve the plug-in bandwidth selection methods. Another option for bandwidth selection is to stabilize the hybrid bandwidth selector by alternative parameter estimation techniques.

For regression estimation, the main avenue for potential research is investigating applying rank-weighting to the local-linear kernel regression estimator described by Li and Racine (2007). In the standard case, the local-linear estimator fixes some issues with the NW estimator. The main issue is that the bias for the NW estimator can increase in magnitude by a large amount near boundaries of the regression function. Additionally, the local-linear estimator usually performs better. Hopefully, this will extend to rank-weighting.

There is another topic of interest for future research that applies to both the rank-weighted density and regression estimators. There may be potential improvements gained from examining the rank-weights $\frac{1}{\sqrt{i}}$. In the density estimation case where only one variable requires ranking, it may be beneficial to generalize the rank-weighted kernel density estimator to the form

$$\frac{1}{nh}\sum_{i=1}^{n}\psi(i)K\left(\psi(i)\frac{x-X_{(i)}}{h}\right),$$

where ψ is some function. In the current case, $\psi(i) = 1/\sqrt{i}$. For the rank-weighted Nadaraya-Watson estimator, it may be possible to take this idea a step further and use some sort of weighting function on R_i and S_i, which are defined to be the ranks of both the

X_i and Y_i observations, respectively. The form of the estimator would then be

$$\frac{\sum_{i=1}^{n} \psi(R_i, S_i) K\left(\psi(R_i, S_i)\frac{X_i - x}{h}\right) Y_i}{\sum_{i=1}^{n} \psi(R_i, S_i) K\left(\psi(R_i, S_i)\frac{X_i - x}{h}\right)}.$$

In either case, a methodology would need to be developed for finding optimal or near optimal functions depending on the data at hand. A simplified situation may be to use the function $\frac{1}{i^q}$ and to find an optimal value for q.

REFERENCES

Akaike, Hirotugu (1974). “A new look at the statistical model identification”. *IEEE Transactions on Automatic Control* 19, 716–723.

Azzalini, Adelchi (1985). “A class of distributions which includes the normal ones”. *Scandinavian journal of statistics*, 171–178.

Bartlett, Maurice S (1963). “Statistical estimation of density functions”. *Sankhyā: The Indian Journal of Statistics, Series A*, 245–254.

Bowman, Adrian (1984). “An alternative method of cross-validation for the smoothing of density estimates”. *Biometrika* 71, 353–360.

Cao, Ricardo, Cuevas, Antonio, and Manteiga, Wensceslao González (1994). “A comparative study of several smoothing methods in density estimation”. *Computational Statistics & Data Analysis* 17, 153–176.

Clark, RM (1977). “Non-parametric estimation of a smooth regression function”. *Journal of the Royal Statistical Society. Series B (Methodological)*, 107–113.

Collomb, Gérard (1976). “Estimation non paramétrique de la régression par la méthode du noyau”. PhD thesis.

Craven, Peter and Wahba, Grace (1978). “Smoothing noisy data with spline functions”. *Numerische Mathematik* 31, 377–403.

David, Herbert A and Nagaraja, Haikady N (2003). *Order Statistics*. Wiley, New York.

Fernández, Carmen and Steel, Mark FJ (1998). “On Bayesian modeling of fat tails and skewness”. *Journal of the American Statistical Association* 93, 359–371.

Greenwood, J Arthur et al. (1979). “Probability weighted moments: definition and relation to parameters of several distributions expressable in inverse form”. *Water Resources Research* 15, 1049–1054.

Hardle, Wolfgang, Marron, JS, and Wand, MP (1990). “Bandwidth choice for density derivatives”. *Journal of the Royal Statistical Society. Series B (Methodological)*, 223–232.

Hardle, Wolfgang et al. (2012). *Nonparametric and Semiparametric Models*. Springer Science & Business Media.

Hart, Jeffrey D and Yi, Seongbaek (1998). “One-sided cross-validation”. *Journal of the American Statistical Association* 93, 620–631.

Heidenreich, Nils-Bastian, Schindler, Anja, and Sperlich, Stefan (2013). “Bandwidth selection for kernel density estimation: a review of fully automatic selectors”. *AStA Advances in Statistical Analysis* 97, 403–433.

Hurvich, Clifford M, Simonoff, Jeffrey S, and Tsai, Chih-Ling (1998). “Smoothing parameter selection in nonparametric regression using an improved Akaike information criterion”. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 60, 271–293.

Jones, M Chris, Marron, James S, and Sheather, Simon J (1996). “A brief survey of bandwidth selection for density estimation”. *Journal of the American Statistical Association* 91, 401–407.

Jones, MC and Signorini, DF (1997). “A comparison of higher-order bias kernel density estimators”. *Journal of the American Statistical Association* 92, 1063–1073.

Köhler, Max, Schindler, Anja, and Sperlich, Stefan (2014). “A review and comparison of bandwidth selection methods for kernel regression”. *International Statistical Review* 82, 243–274.

Kruskal, William H and Wallis, W Allen (1952). “Use of ranks in one-criterion variance analysis”. *Journal of the American Statistical Association* 47, 583–621.

Li, Qi and Racine, Jeffrey Scott (2007). *Nonparametric Econometrics: Theory and Practice*. Princeton University Press.

Loader, Clive R (1999). “Bandwidth selection: classical or plug-in?” *Annals of Statistics*, 415–438.

Mann, Henry B and Whitney, Donald R (1947). “On a test of whether one of two random variables is stochastically larger than the other”. *The Annals of Mathematical Statistics*, 50–60.

Marron, James Stephen and Ruppert, David (1994). “Transformations to reduce boundary bias in kernel density estimation”. *Journal of the Royal Statistical Society. Series B (Methodological)*, 653–671.

Nadaraya, Elizbar A (1964). “On estimating regression”. *Theory of Probability & Its Applications* 9, 141–142.

Park, Byeong, Turlach, Berwin, et al. (1992). *Practical performance of several data driven bandwidth selectors*. Tech. rep. Université catholique de Louvain, Center for Operations Research and Econometrics (CORE).

Park, Byeong U and Marron, James S (1990). “Comparison of data-driven bandwidth selectors”. *Journal of the American Statistical Association* 85, 66–72.

Park, So Ryoung et al. (2013). “On the sums of probability functions of order statistics”. *Journal of the Korean Statistical Society* 42, 257–265.

Parzen, Emanuel (1962). “On estimation of a probability density function and mode”. *The Annals of Mathematical Statistics* 33, 1065–1076.

Rao, BLS Prakasa (2014). *Nonparametric functional estimation*. Academic press.

Rosenblatt, Murray et al. (1956). “Remarks on some nonparametric estimates of a density function”. *The Annals of Mathematical Statistics* 27, 832–837.

Rudemo, Mats (1982). “Empirical choice of histograms and kernel density estimators”. *Scandinavian Journal of Statistics*, 65–78.

Savchuk, Olga Y, Hart, Jeffrey D, and Sheather, Simon J (2010). "Indirect cross-validation for density estimation". *Journal of the American Statistical Association* 105, 415–423.

Scott, David W (1992). *Multivariate density estimation: theory, practice, and visualization*. John Wiley & Sons.

Scott, David W and Terrell, George R (1987). "Biased and unbiased cross-validation in density estimation". *Journal of the American Statistical Association* 82, 1131–1146.

Sheather, Simon J and Jones, Michael C (1991). "A reliable data-based bandwidth selection method for kernel density estimation". *Journal of the Royal Statistical Society. Series B (Methodological)*, 683–690.

Shibata, Ritei (1981). "An optimal selection of regression variables". *Biometrika* 68, 45–54.

Silverman, Bernard W (1986). *Density estimation for statistics and data analysis*. Vol. 26. CRC press.

Stoker, Thomas M (1993). "Smoothing bias in density derivative estimation". *Journal of the American Statistical Association* 88, 855–863.

Stuart, Alan and Ord, Keith J. (1987). *Kendall's Advanced Theory of Statistics*. Oxford University Press.

Timsina, Thagendra (2015). "Transformed nonparametric functions estimation". PhD thesis. Oklahoma State University.

Wand, Matt P and Jones, M Chris (1994). *Kernel Smoothing*. CRC Press.

Watson, Geoffrey S (1964). "Smooth regression analysis". *Sankhyā: The Indian Journal of Statistics, Series A*, 359–372.

Appendix A

Density Bandwidth Plots

Plots begin on the next page.

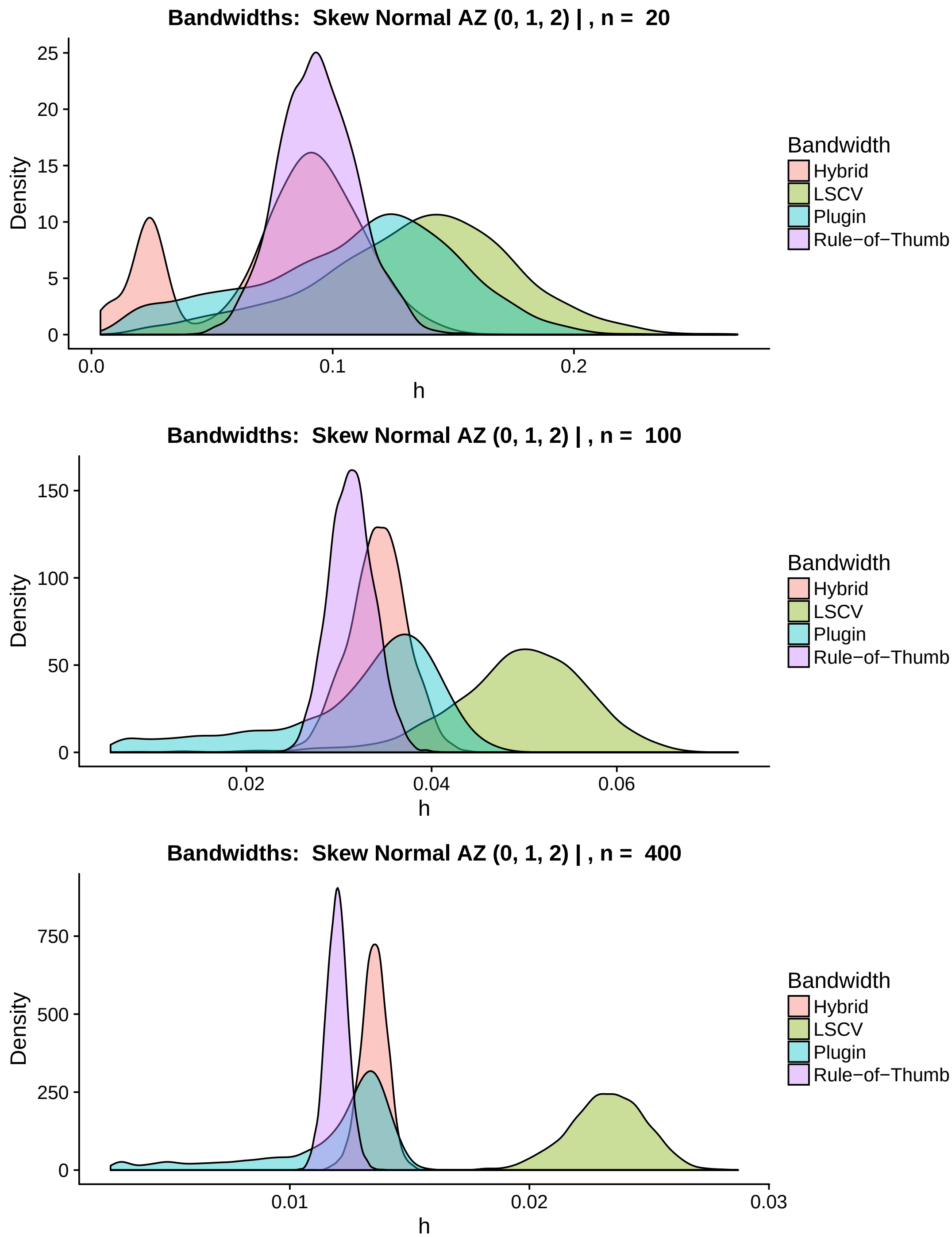


Figure A.1: Plot of bandwidths for each of the four bandwidth selection methods discussed for the rank weighted estimator at sample sizes 20, 100, 200, and 400 of the $SkewNormAZ(0,1,2)$.

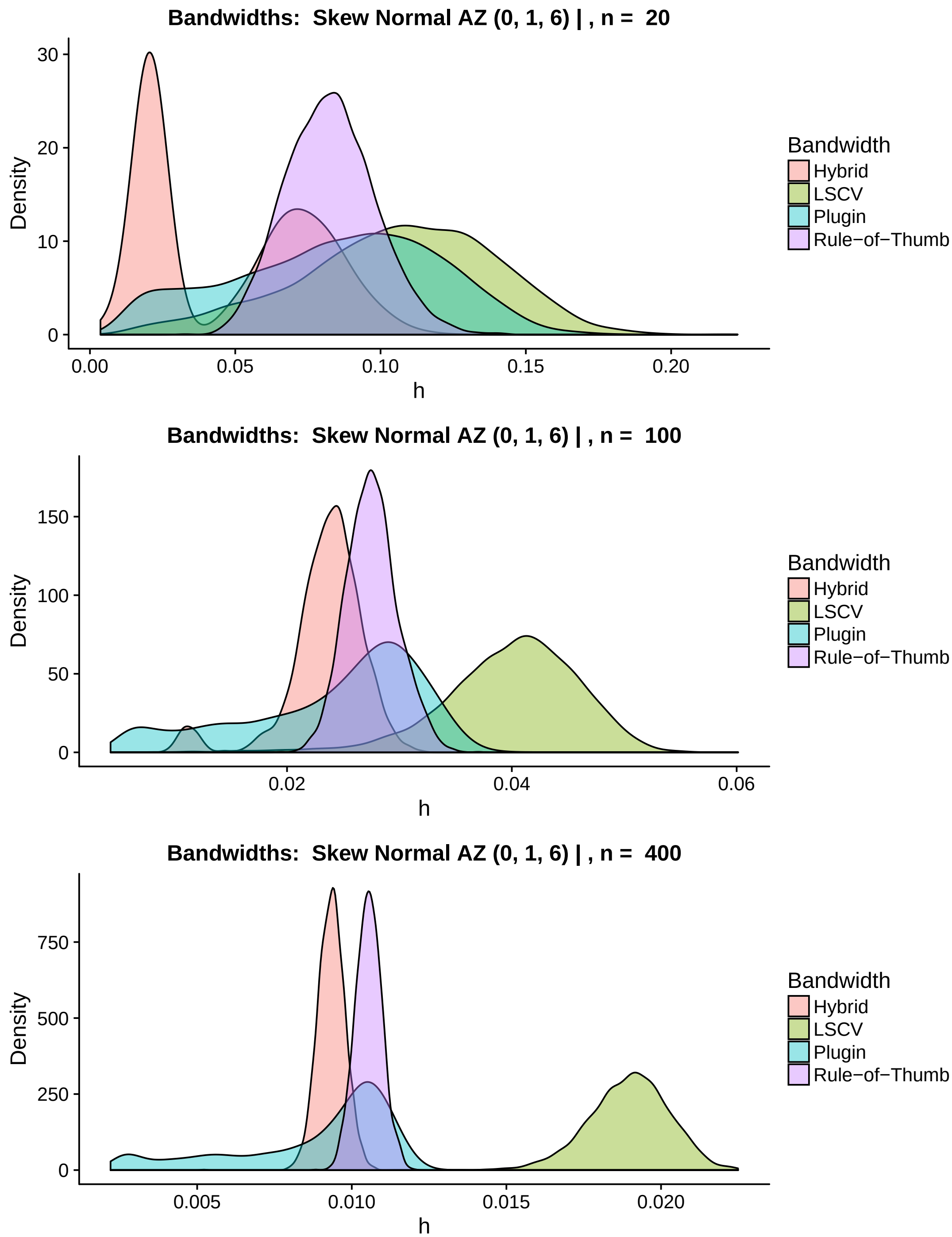


Figure A.2: Plot of bandwidths for each of the four bandwidth selection methods discussed for the rank weighted estimator at sample sizes 20, 100, 200, and 400 of the $SkewNormAZ(0,1,6)$.

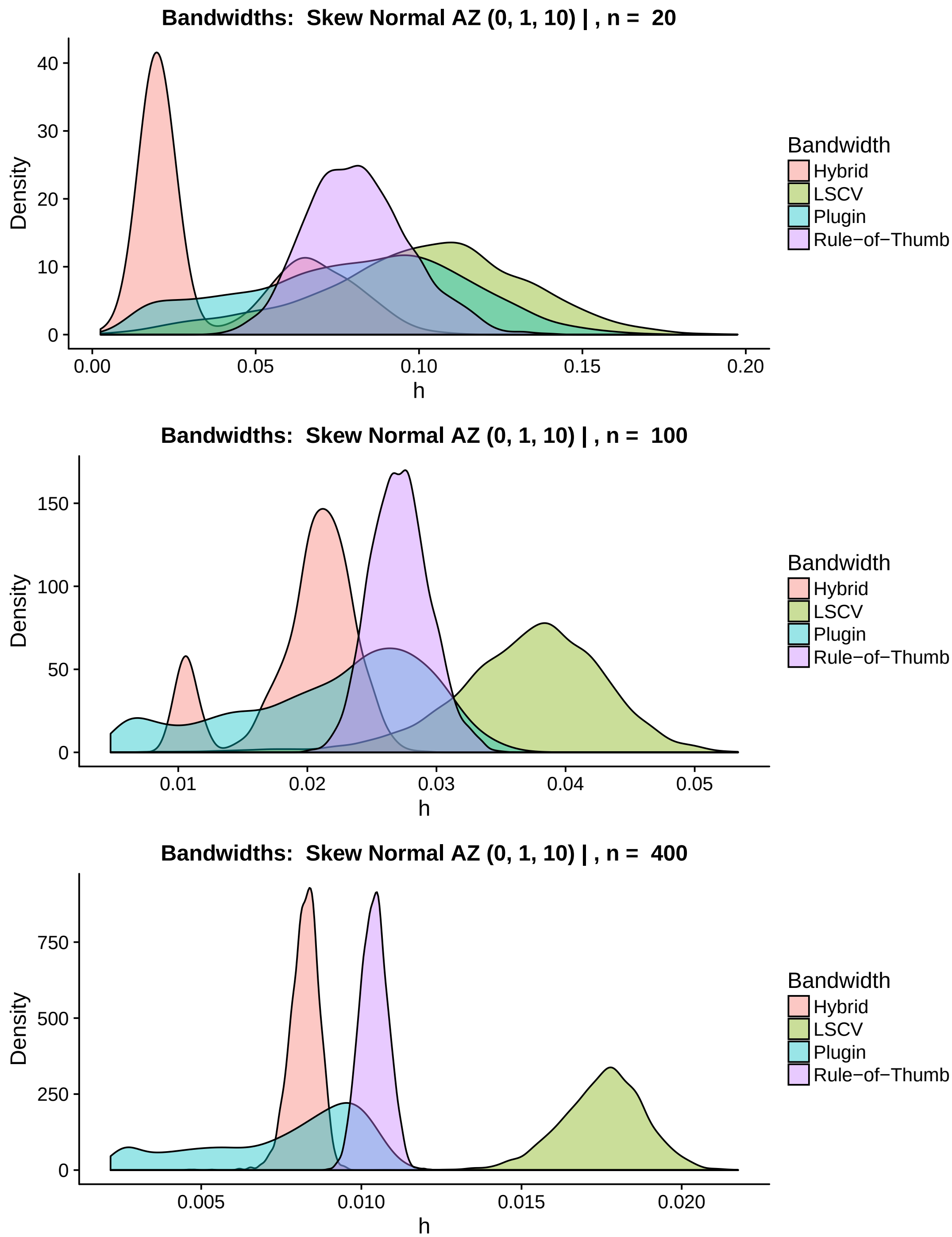


Figure A.3: Plot of bandwidths for each of the four bandwidth selection methods discussed for the rank weighted estimator at sample sizes 20, 100, 200, and 400 of the $SkewNormAZ(0,1,10)$.

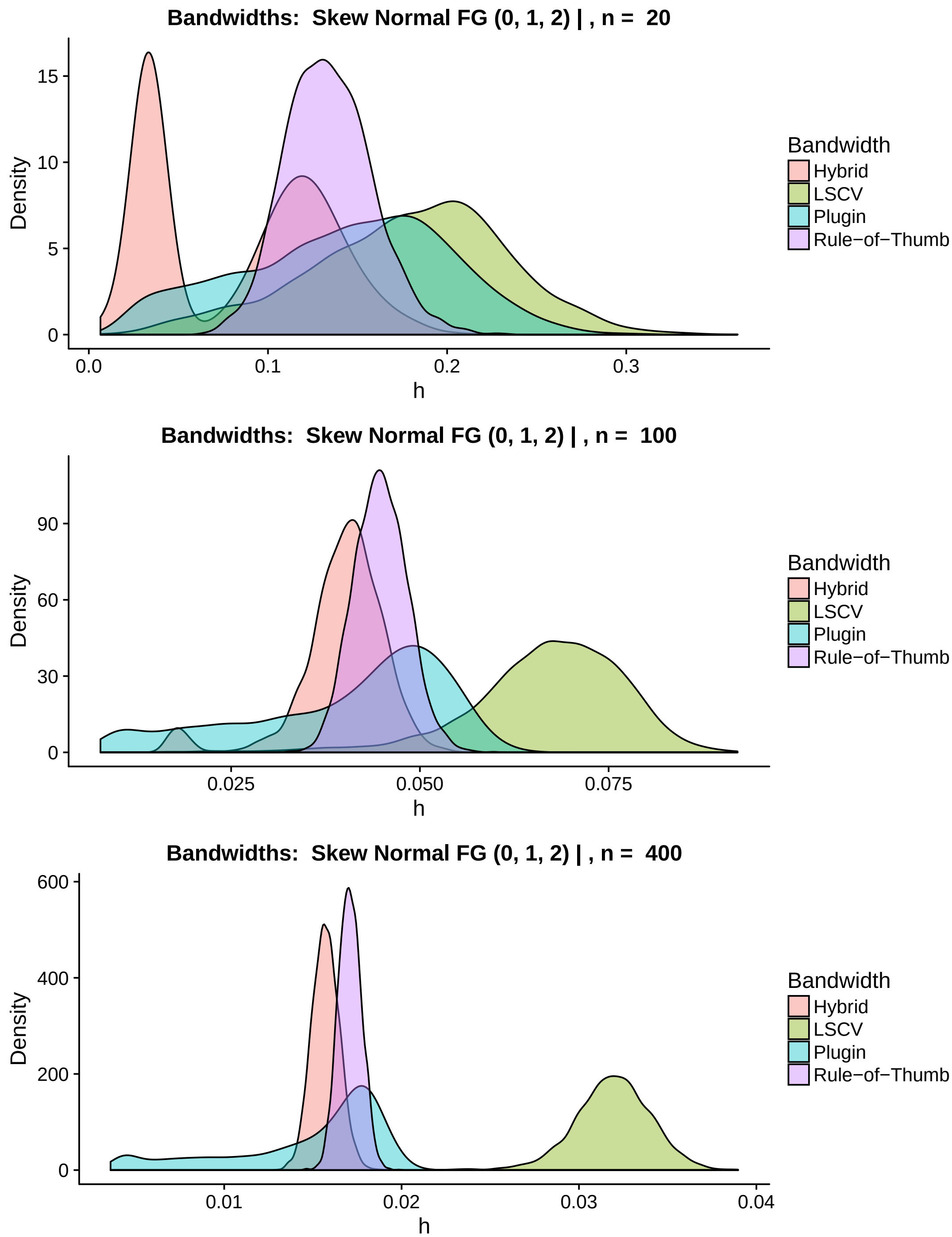


Figure A.4: Plot of bandwidths for each of the four bandwidth selection methods discussed for the rank weighted estimator at sample sizes 20, 100, 200, and 400 of the $SkewNormFG(0,1,2)$.

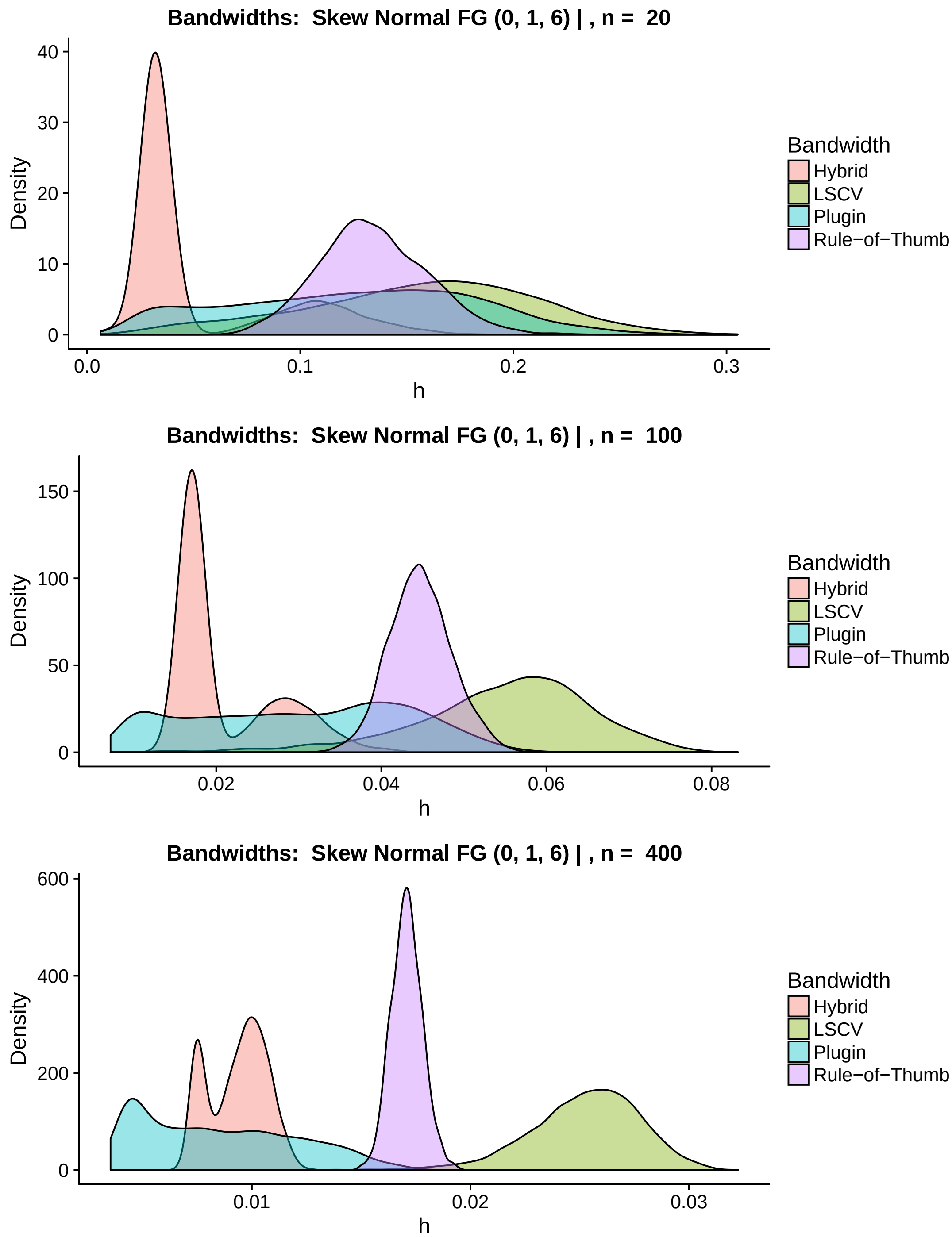


Figure A.5: Plot of bandwidths for each of the four bandwidth selection methods discussed for the rank weighted estimator at sample sizes 20, 100, 200, and 400 of the $SkewNormFG(0,1,6)$.

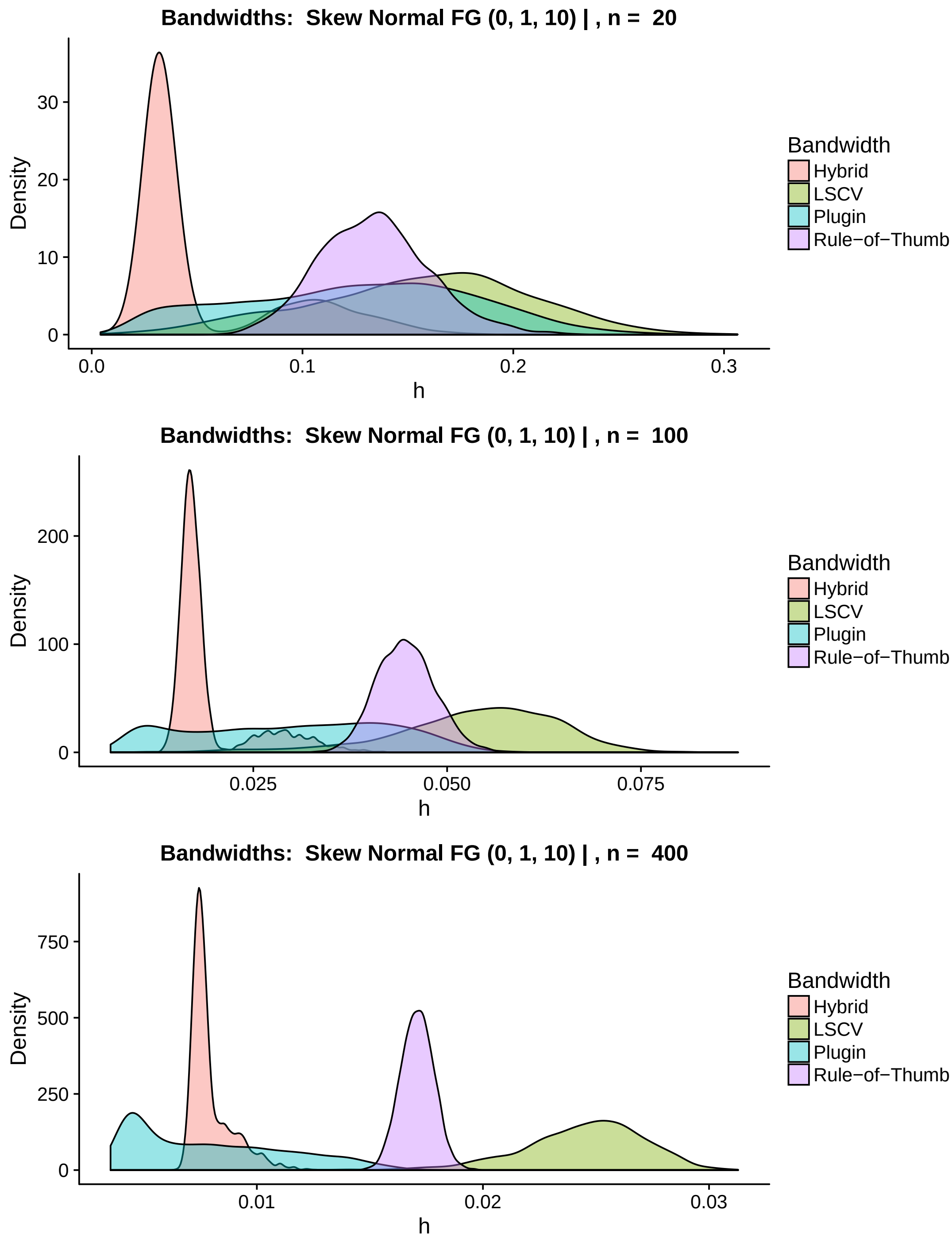


Figure A.6: Plot of bandwidths for each of the four bandwidth selection methods discussed for the rank weighted estimator at sample sizes 20, 100, 200, and 400 of the $SkewNormFG(0,1,10)$.

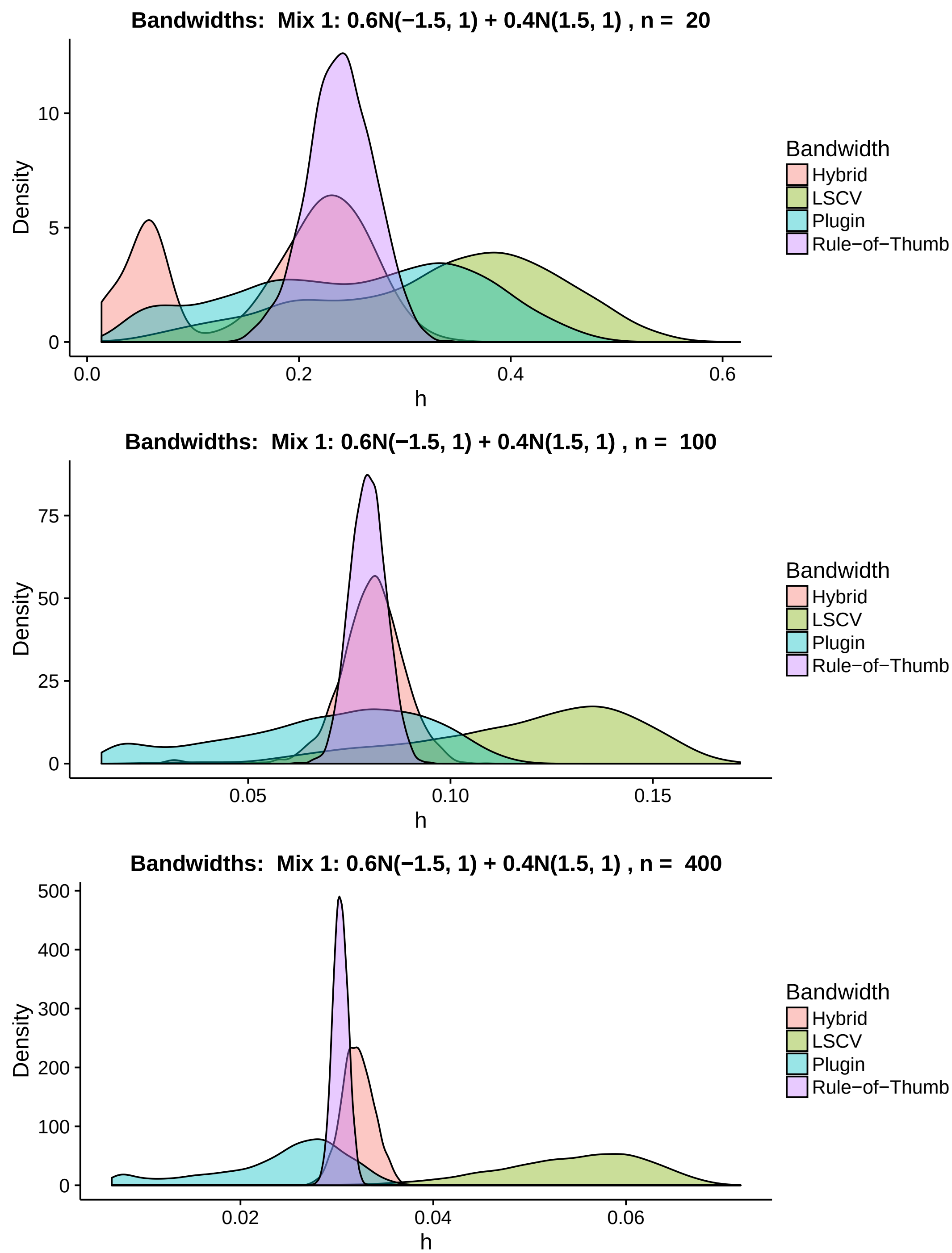


Figure A.7: Plot of bandwidths for each of the four bandwidth selection methods discussed for the rank weighted estimator at sample sizes 20, 100, 200, and 400 of the normal mixture 1.

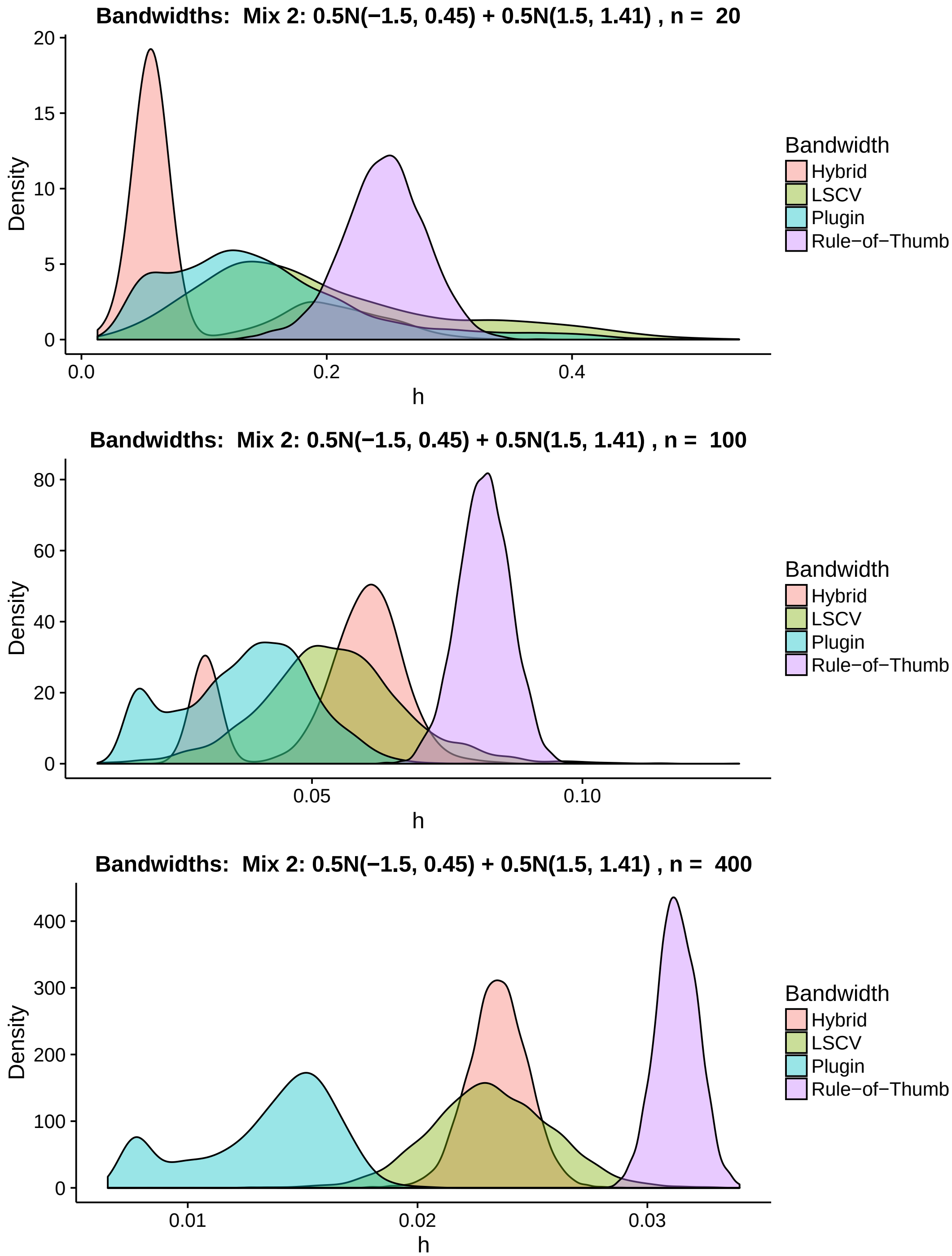


Figure A.8: Plot of bandwidths for each of the four bandwidth selection methods discussed for the rank weighted estimator at sample sizes 20, 100, 200, and 400 of the normal mixture 2.

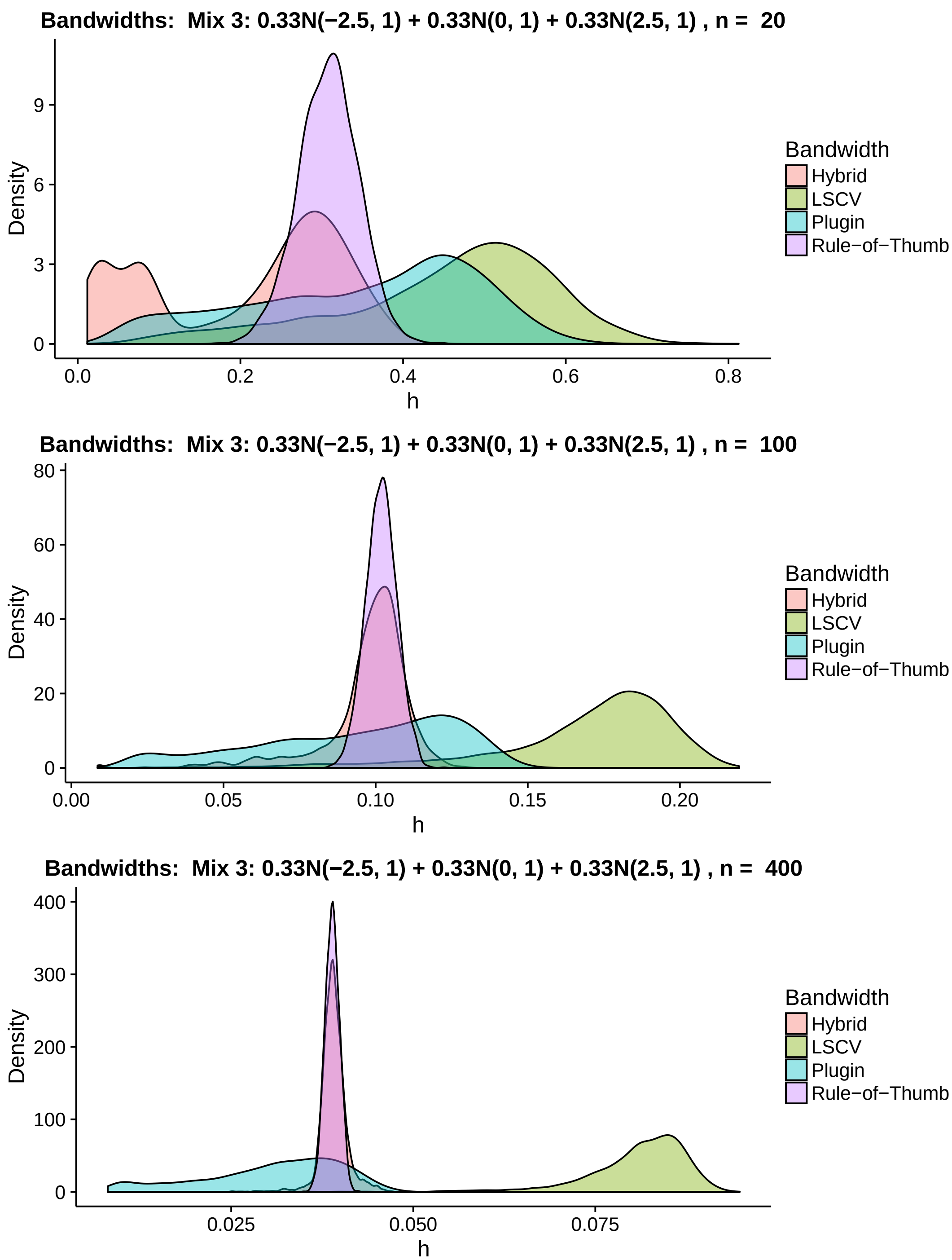


Figure A.9: Plot of bandwidths for each of the four bandwidth selection methods discussed for the rank weighted estimator at sample sizes 20, 100, 200, and 400 of the normal mixture 3.

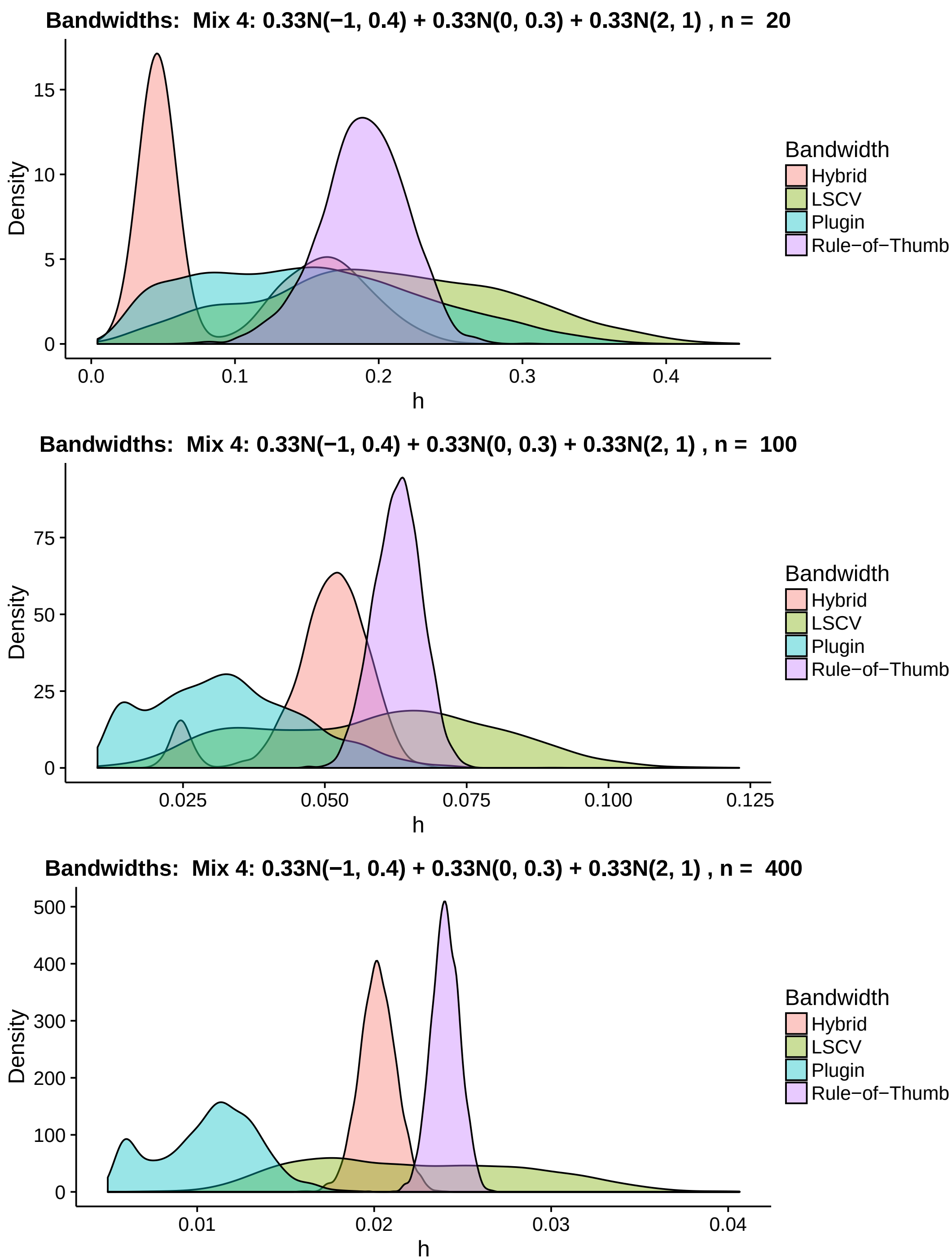


Figure A.10: Plot of bandwidths for each of the four bandwidth selection methods discussed for the rank weighted estimator at sample sizes 20, 100, 200, and 400 of the normal mixture 4.

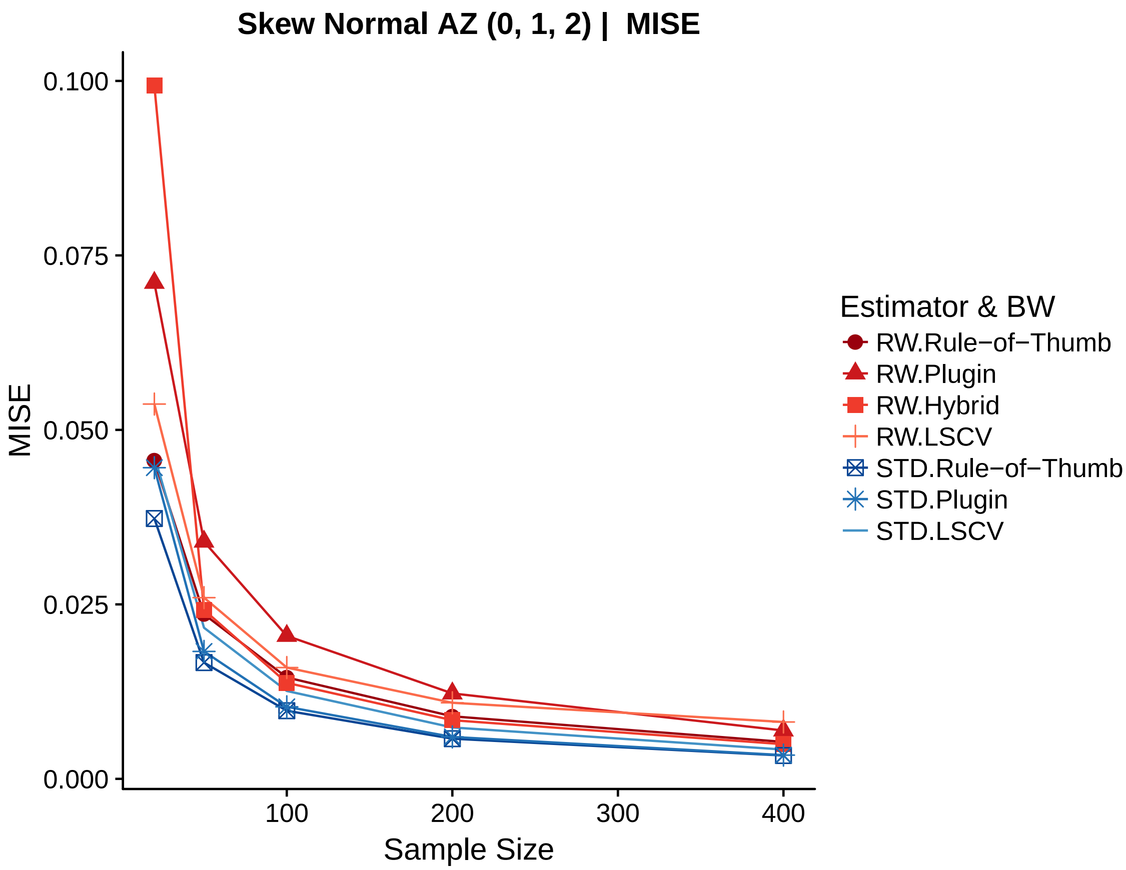


Figure A.11: MISEs for the $SkewNormAZ(0,1,2)$.

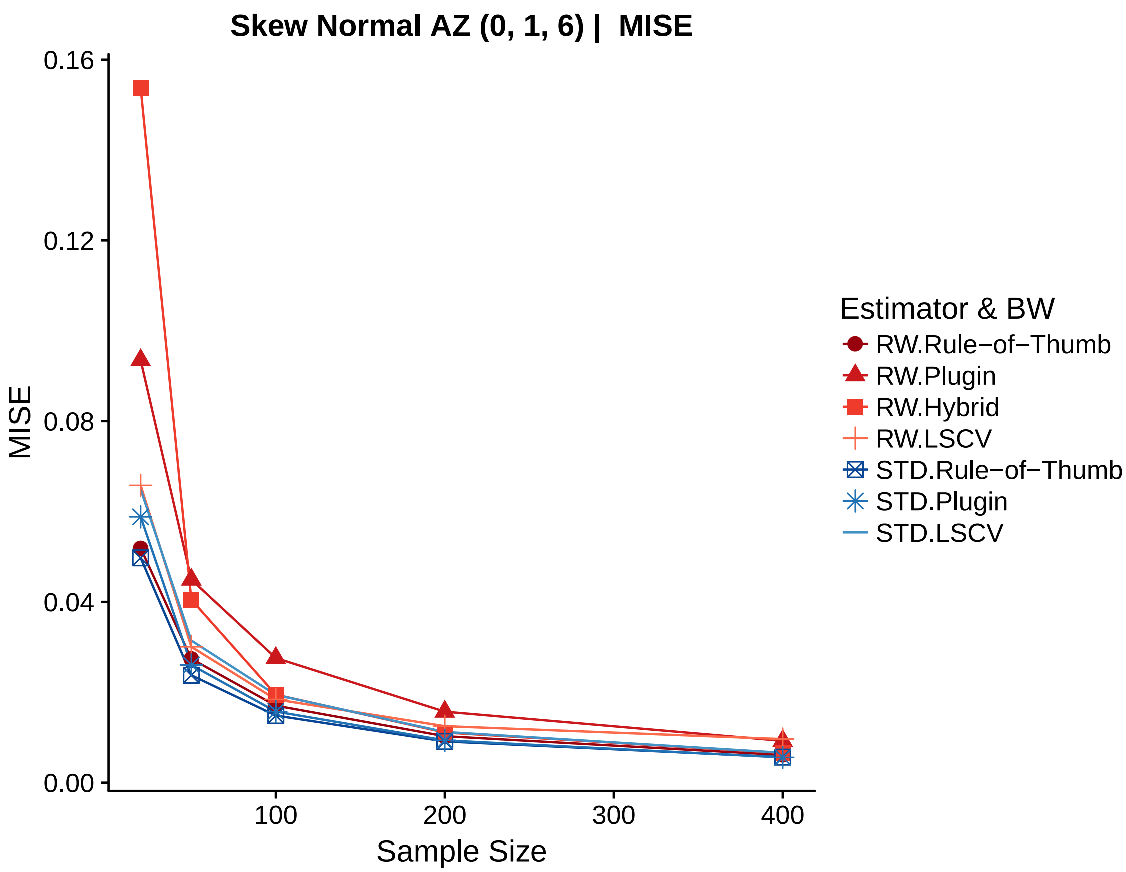


Figure A.12: MISEs for the $SkewNormAZ(0,1,6)$.

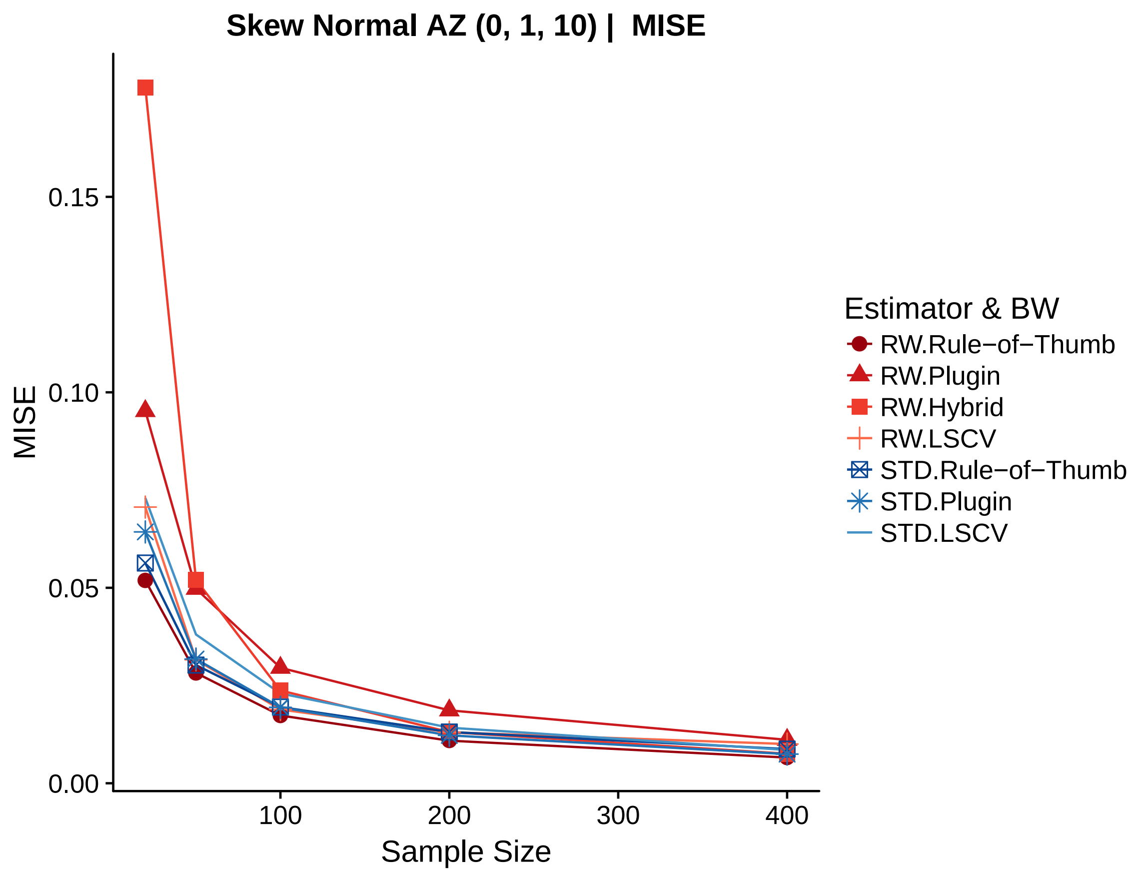


Figure A.13: MISEs for the $SkewNormAZ(0,1,10)$.

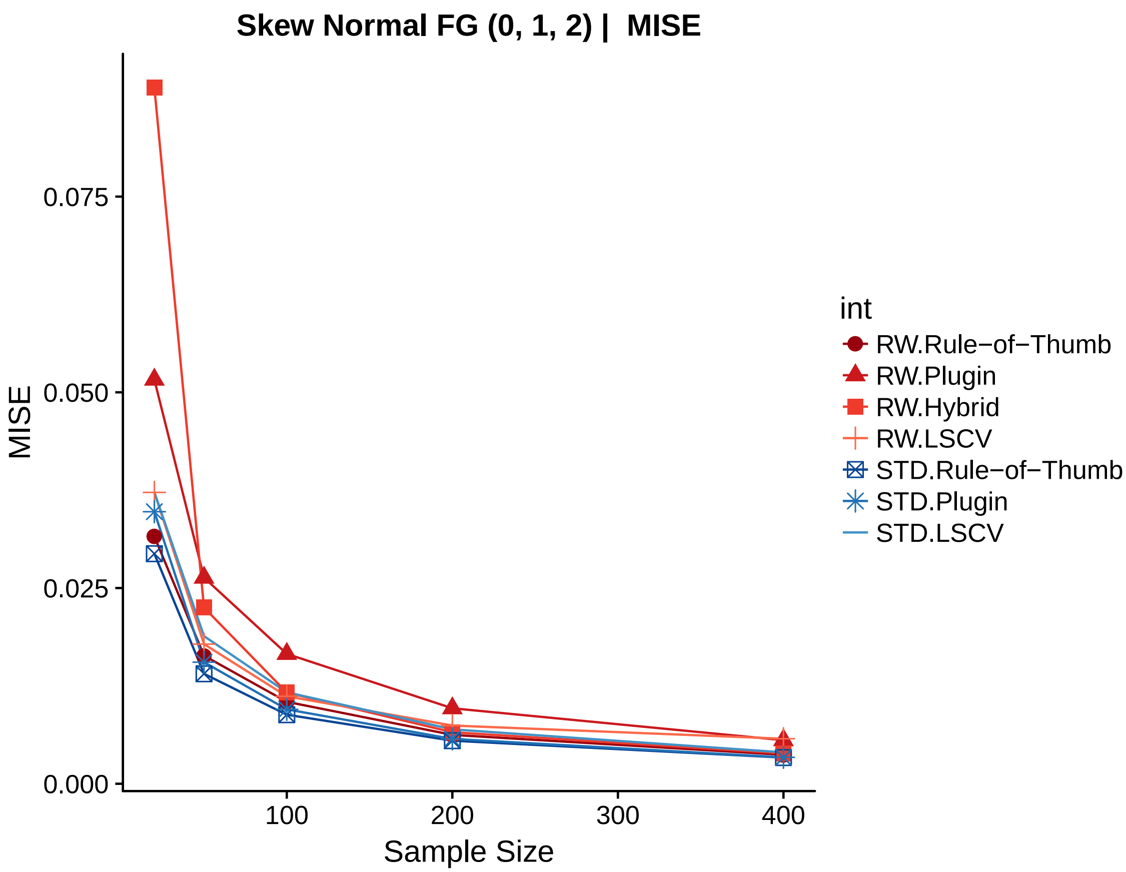


Figure A.14: MISEs for the $SkewNormFG(0,1,2)$.

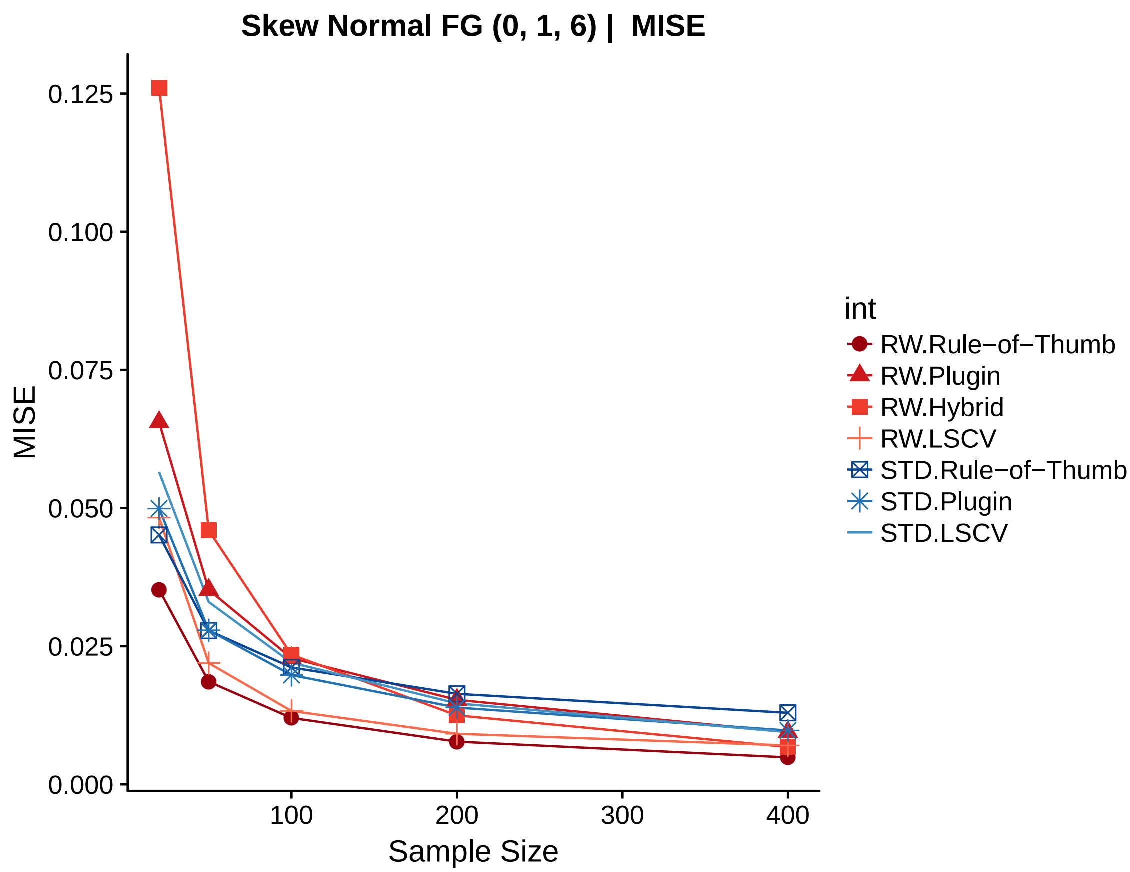


Figure A.15: MISEs for the $SkewNormFG(0, 1, 6)$.

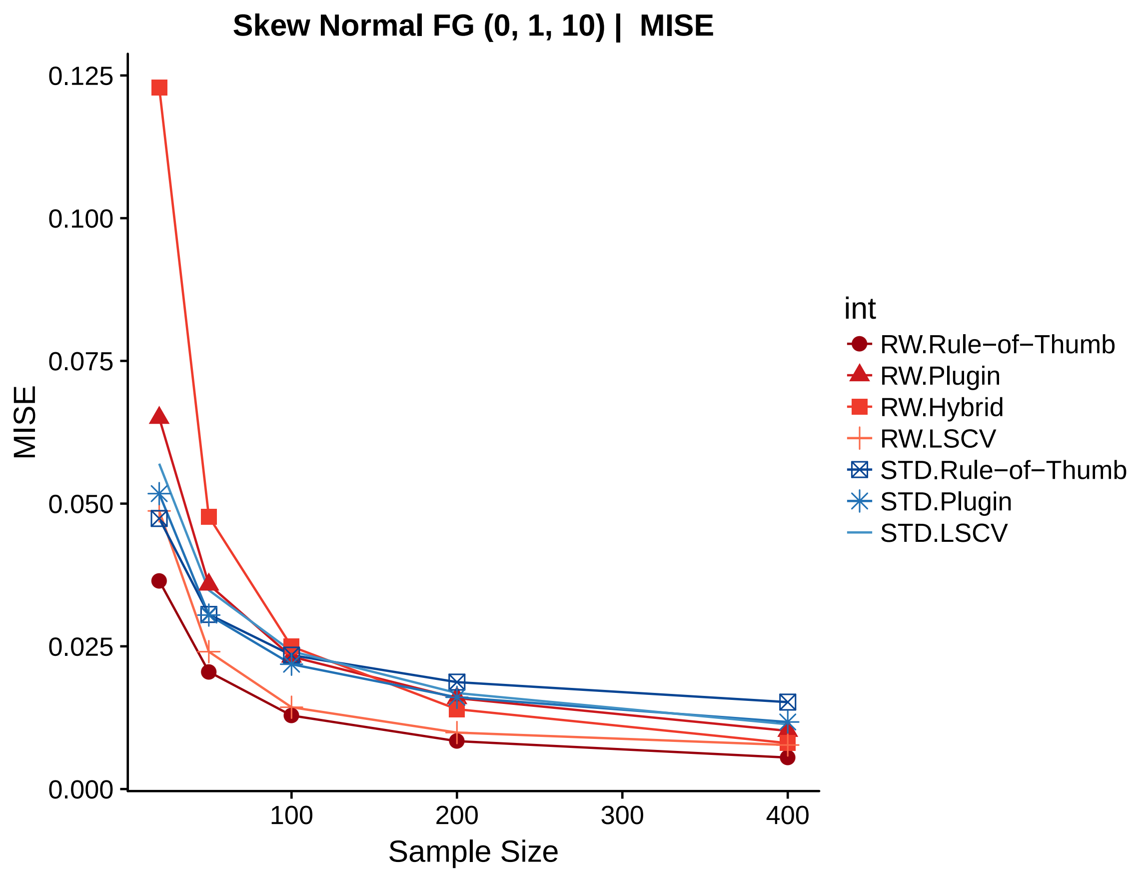


Figure A.16: MISEs for the $SkewNormFG(0,1,10)$.

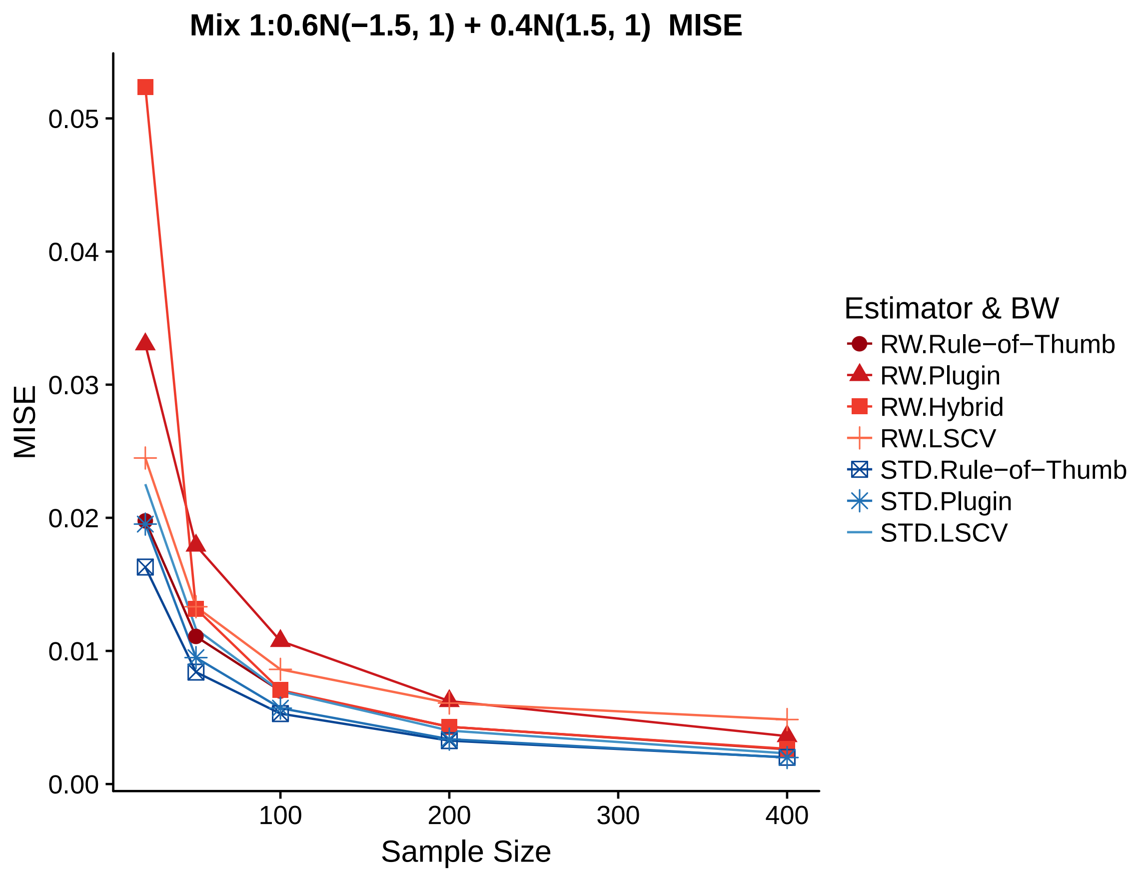


Figure A.17: MISEs for the normal mixture 1.

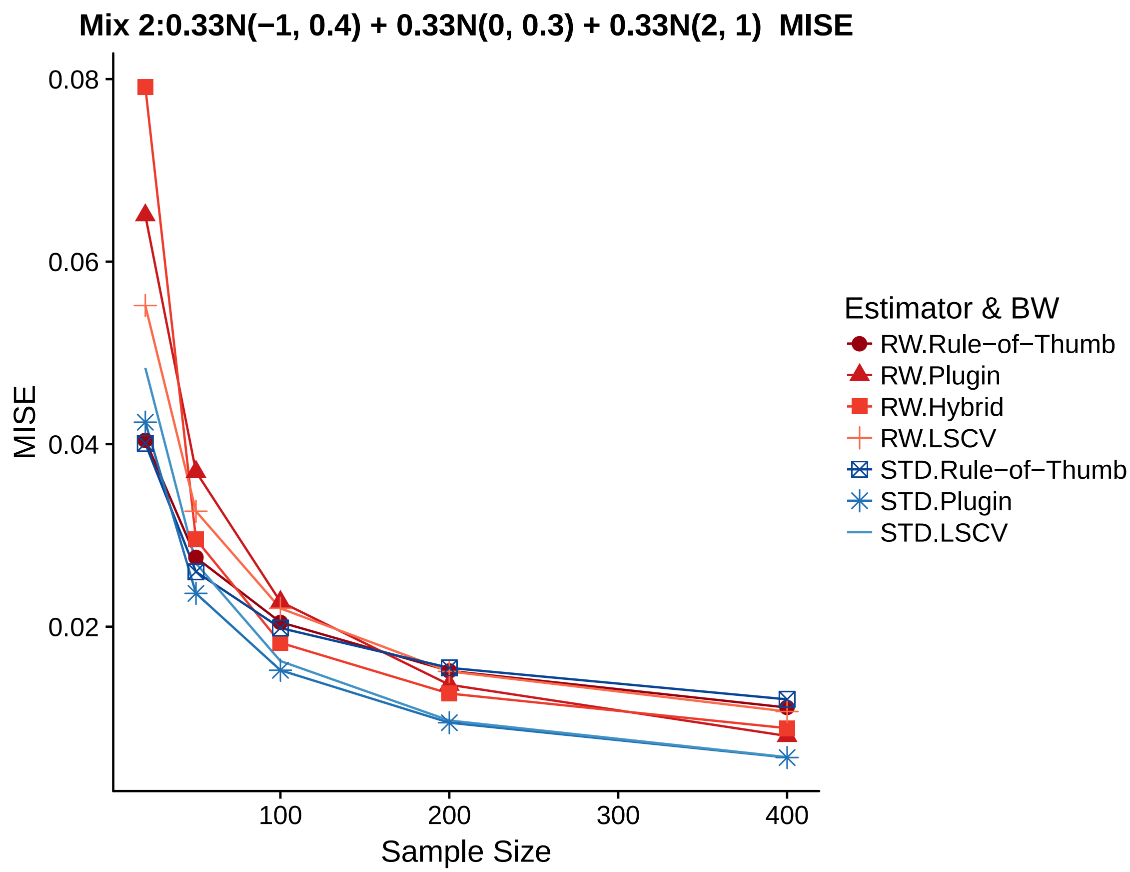


Figure A.18: MISEs for the normal mixture 2.

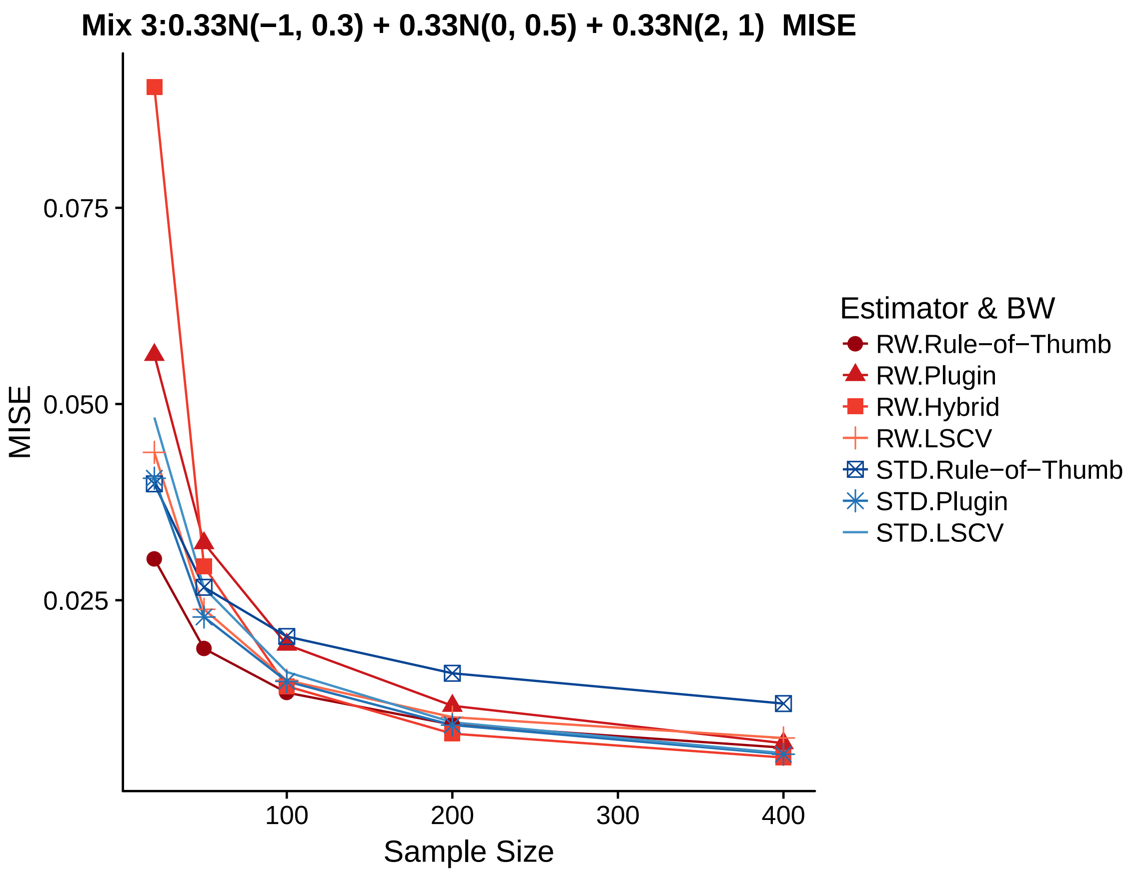


Figure A.19: MISEs for the normal mixture 3.

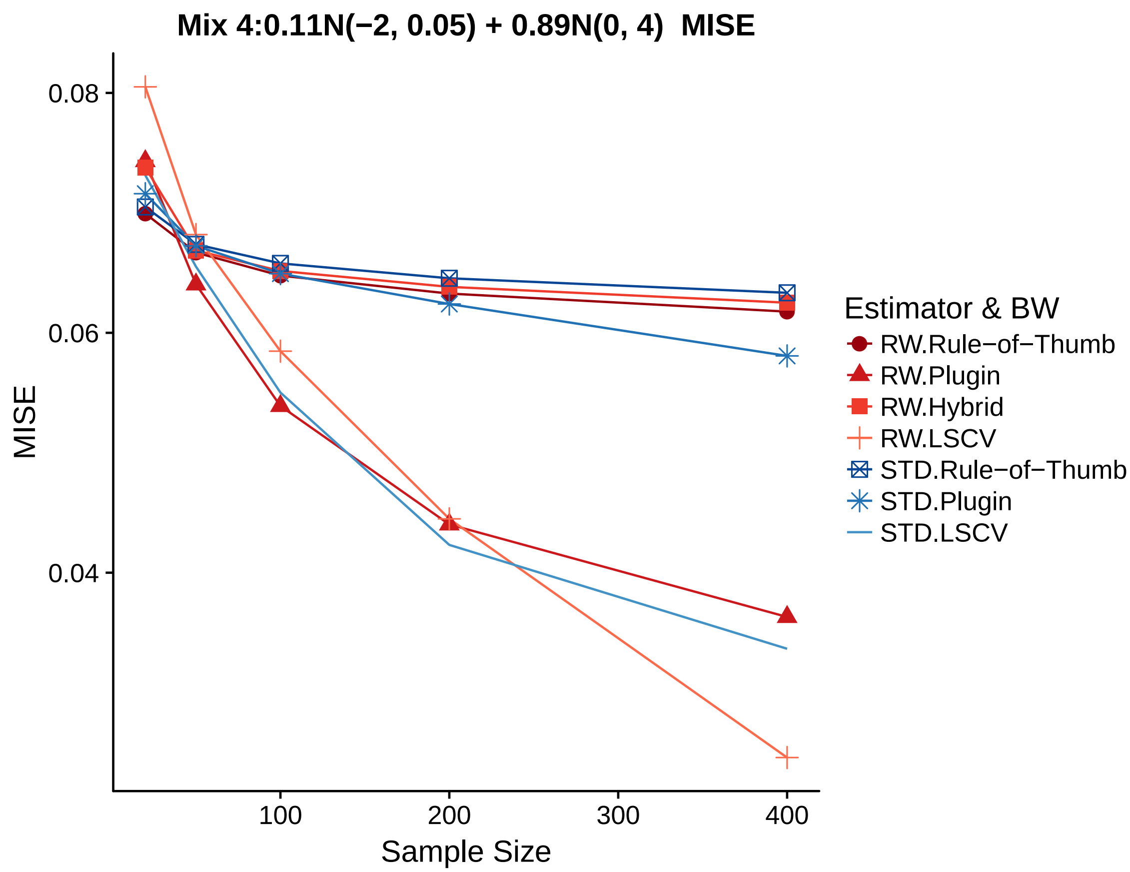
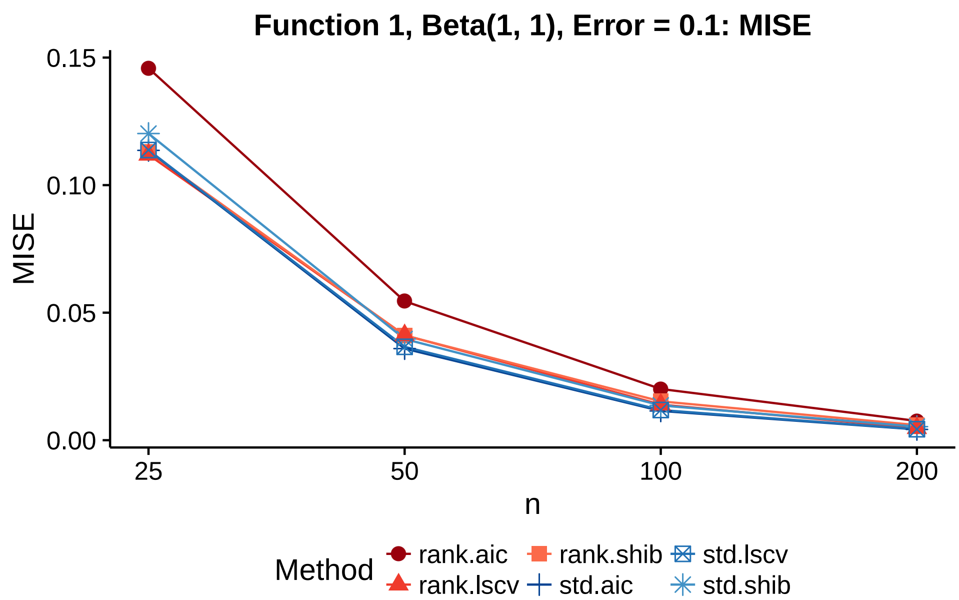


Figure A.20: MISEs for the normal mixture 4.



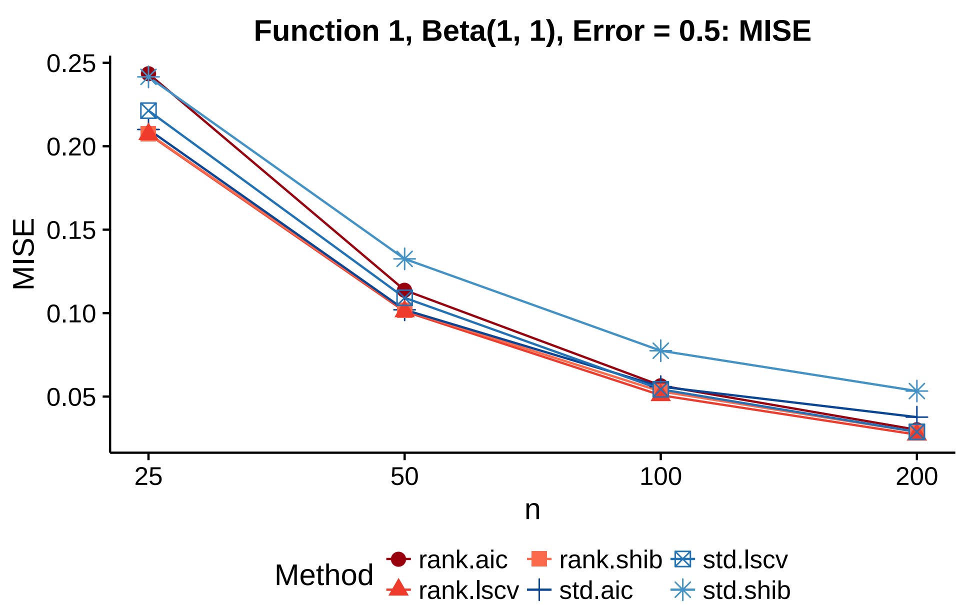


Figure A.21: MISEs for the NW estimator and RWNW estimator under the indicated bandwidth methods in the legend. Results are for function 1 under the Beta(1, 1) distribution.

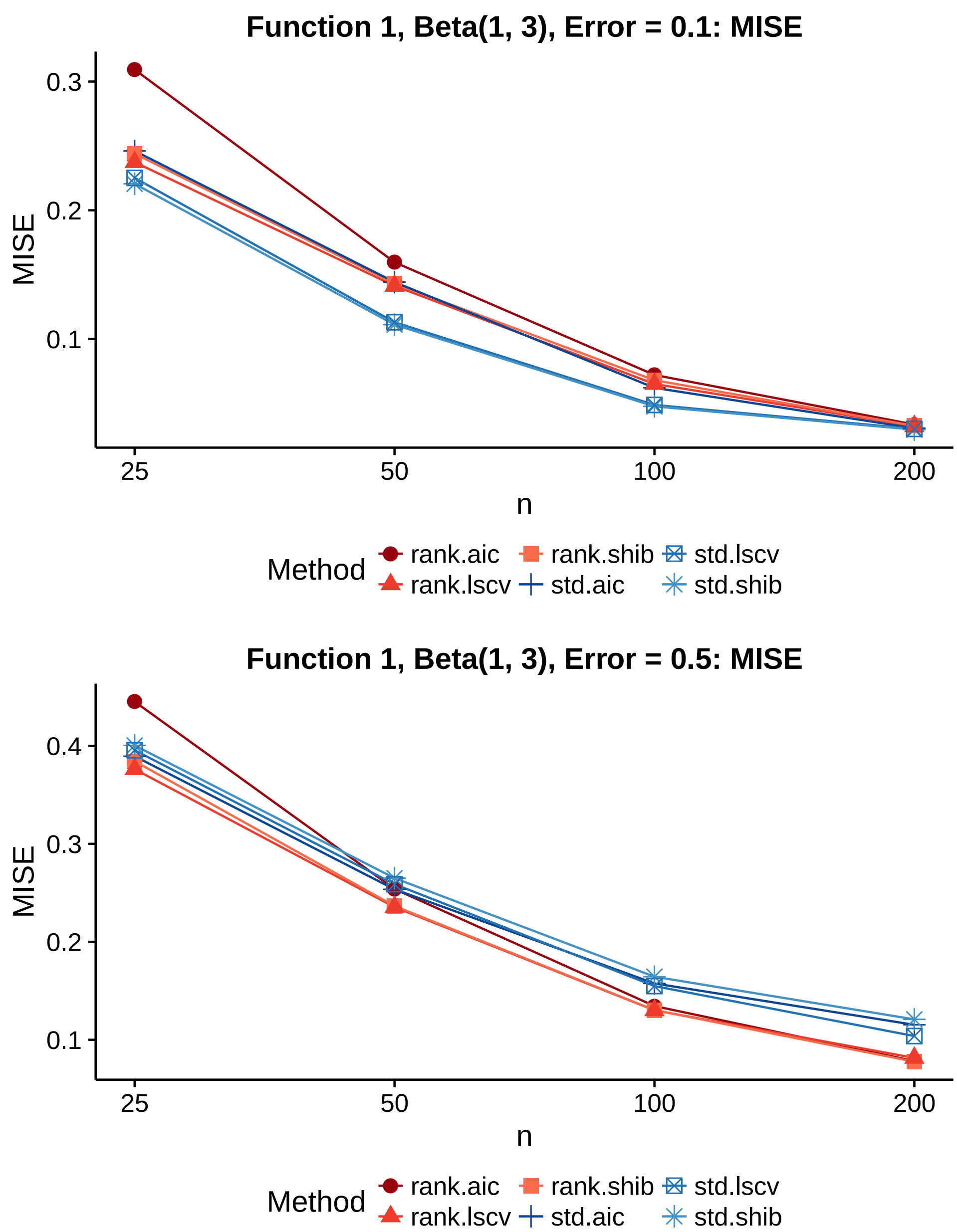


Figure A.22: MISEs for the NW estimator and RWNW estimator under the indicated bandwidth methods in the legend. Results are for function 1 under the Beta(1, 3) distribution.

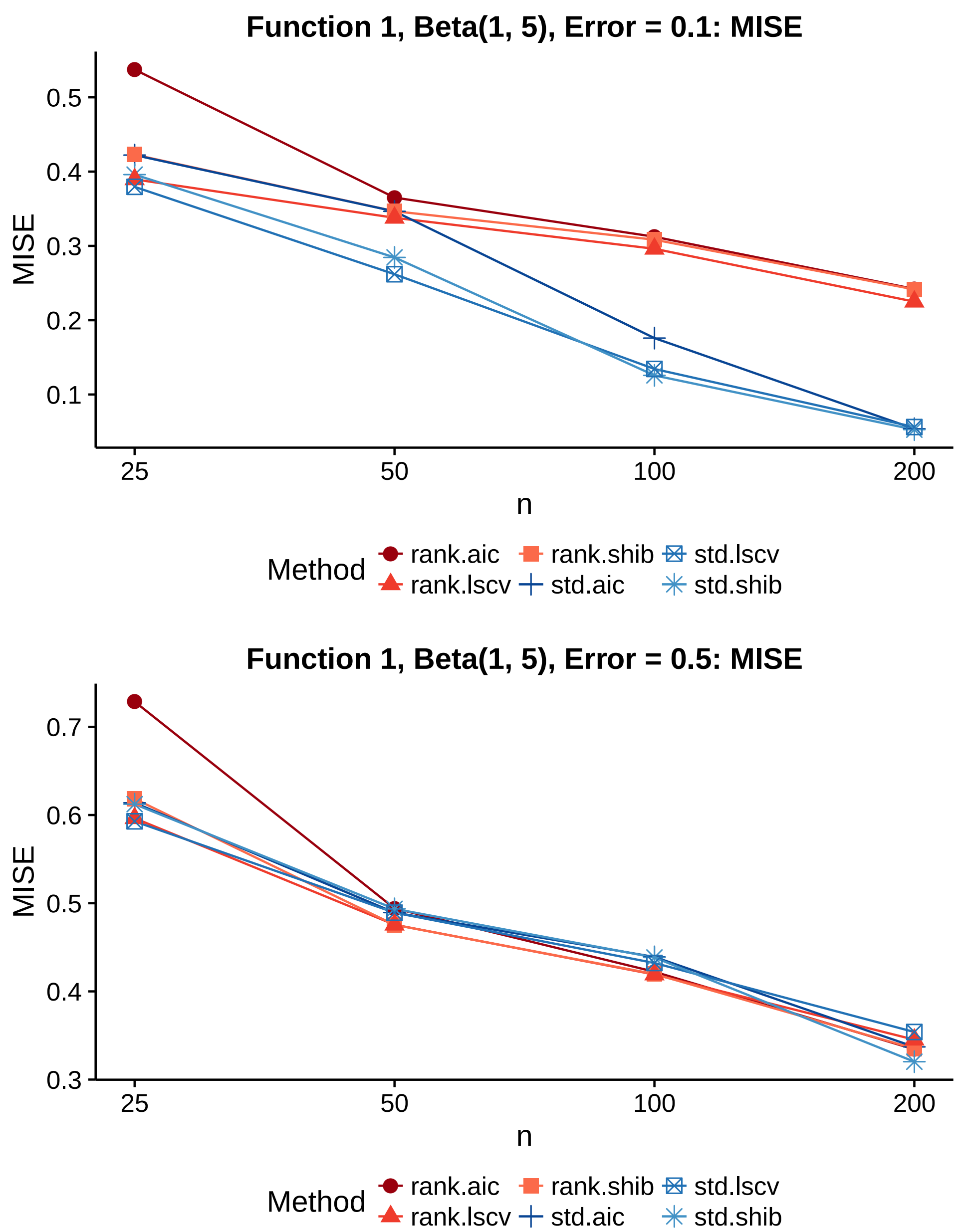
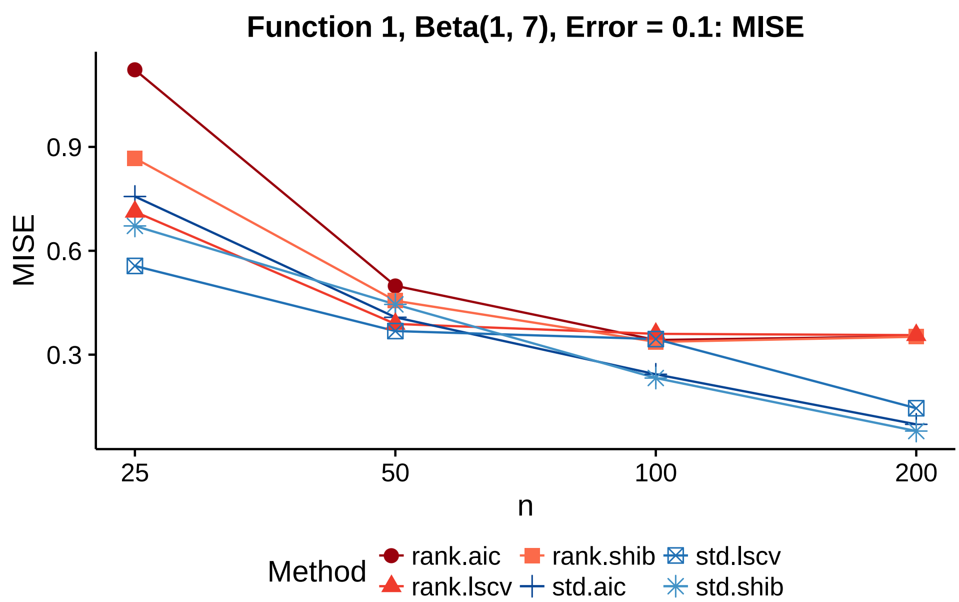


Figure A.23: MISEs for the NW estimator and RWNW estimator under the indicated bandwidth methods in the legend. Results are for function 1 under the Beta(1, 5) distribution.



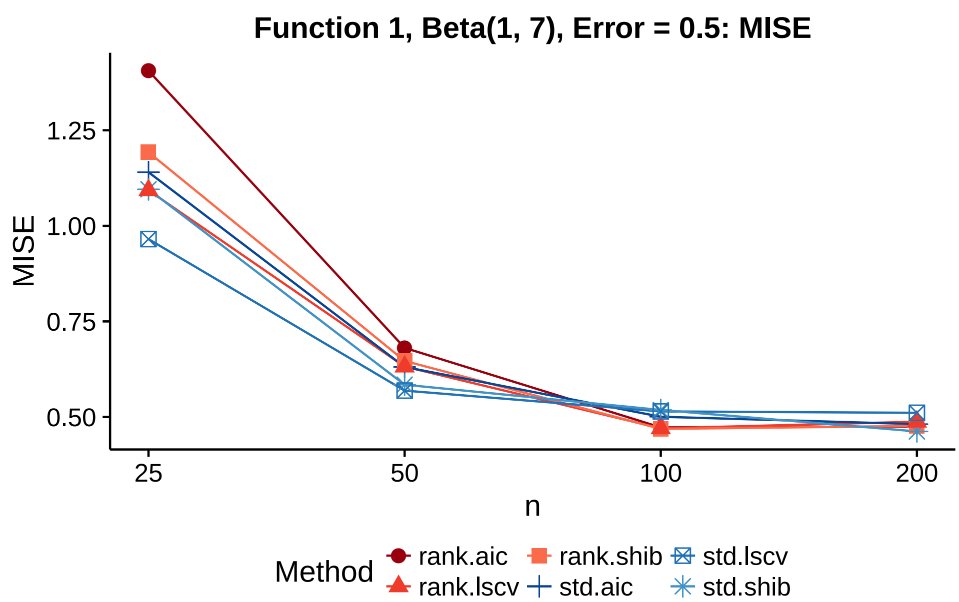
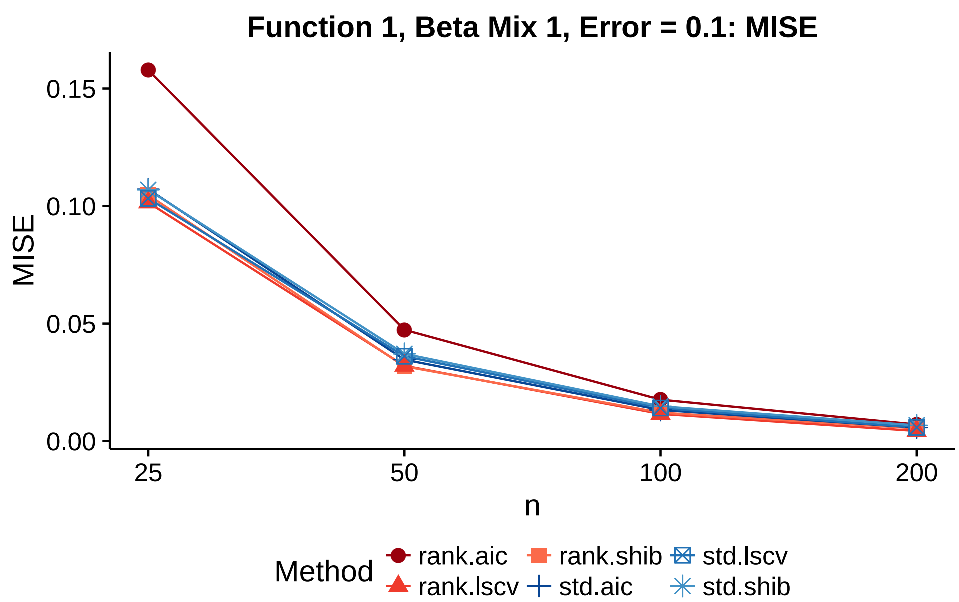


Figure A.24: MISEs for the NW estimator and RWNW estimator under the indicated bandwidth methods in the legend. Results are for function 1 under the Beta(1, 7) distribution.



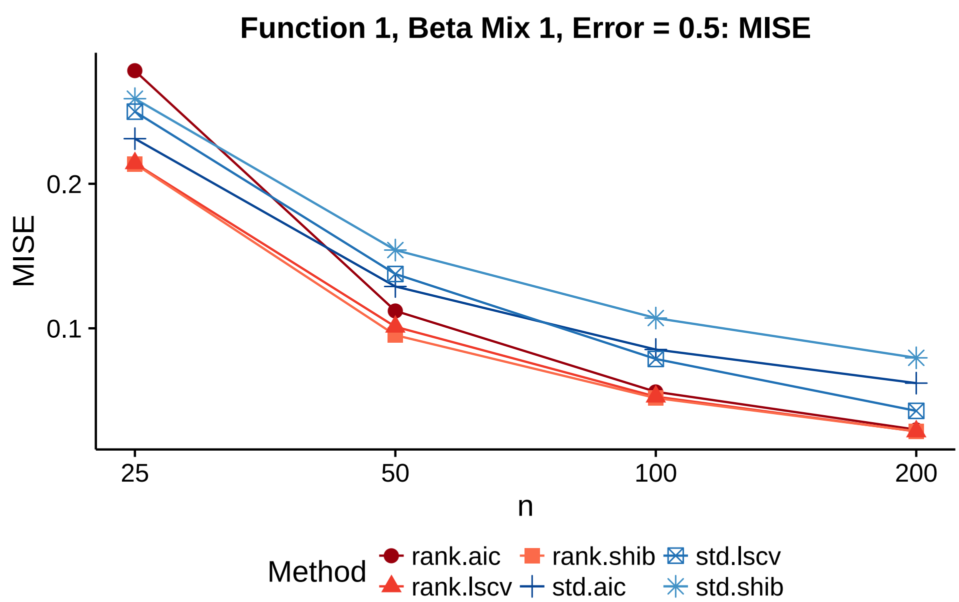
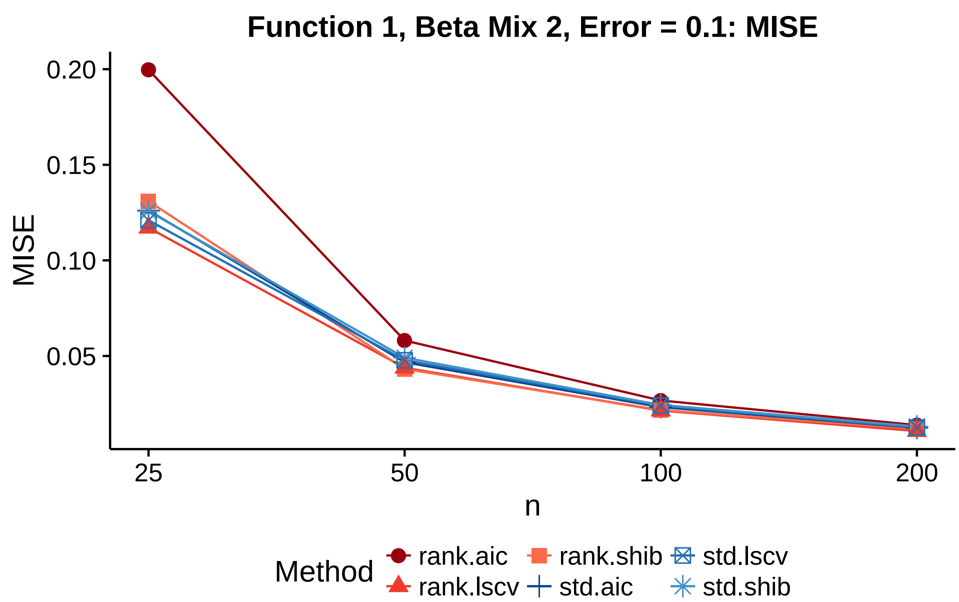


Figure A.25: MISEs for the NW estimator and RWNW estimator under the indicated bandwidth methods in the legend. Results are for function 1 under the beta mixture 1.



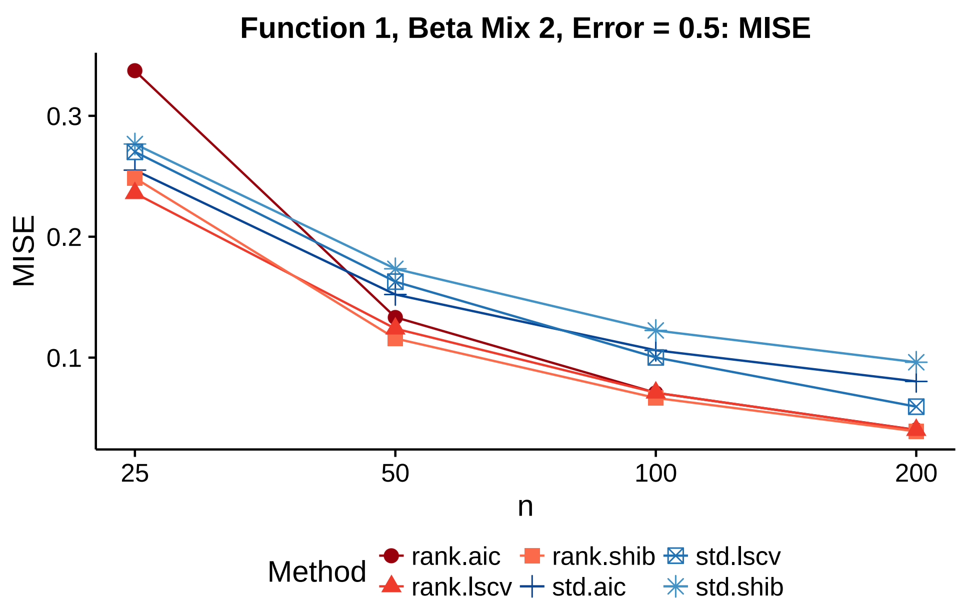
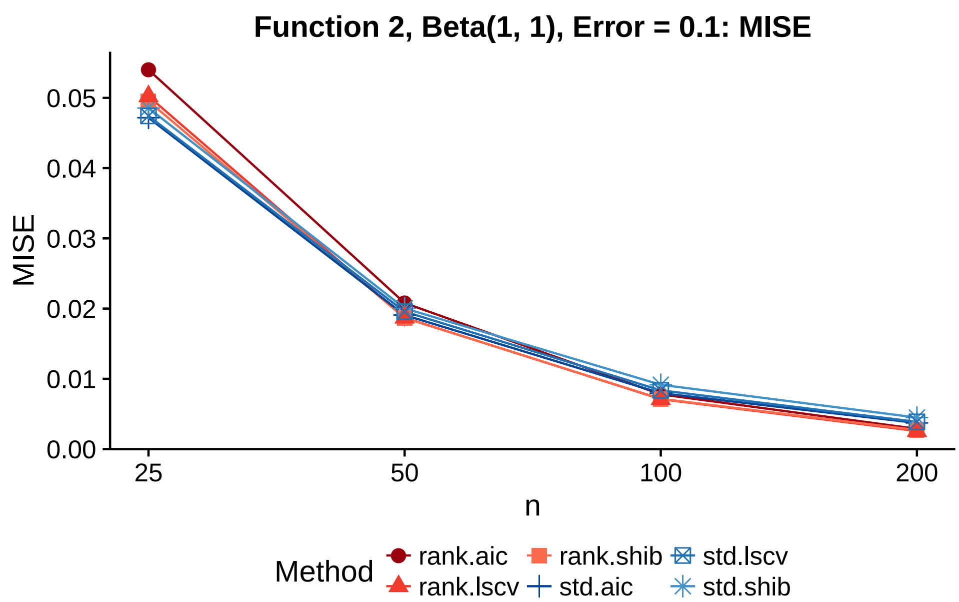


Figure A.26: MISEs for the NW estimator and RWNW estimator under the indicated bandwidth methods in the legend. Results are for function 1 under the beta mixture 2.



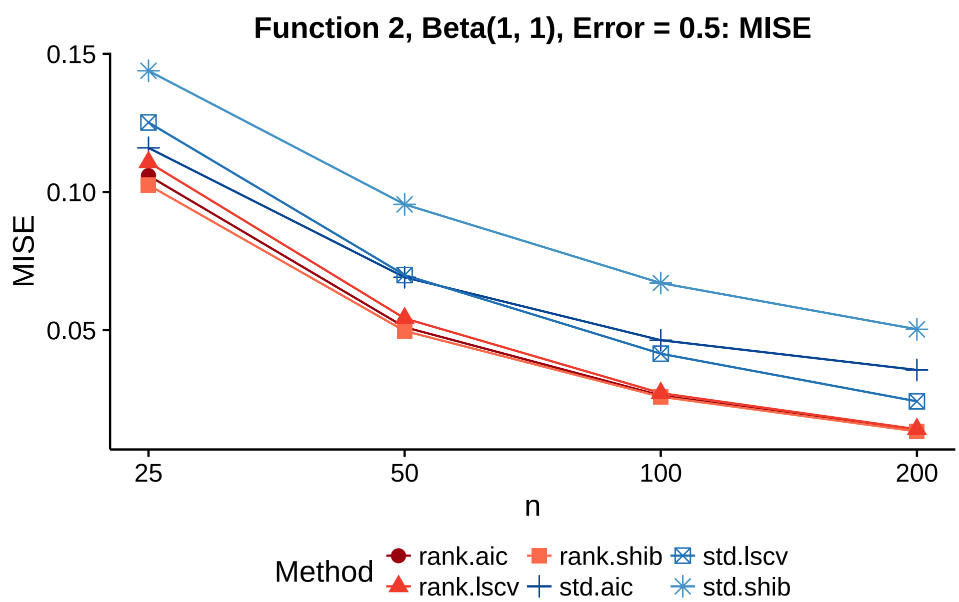


Figure A.27: MISEs for the NW estimator and RWNW estimator under the indicated bandwidth methods in the legend. Results are for function 2 under the Beta(1, 1) distribution.

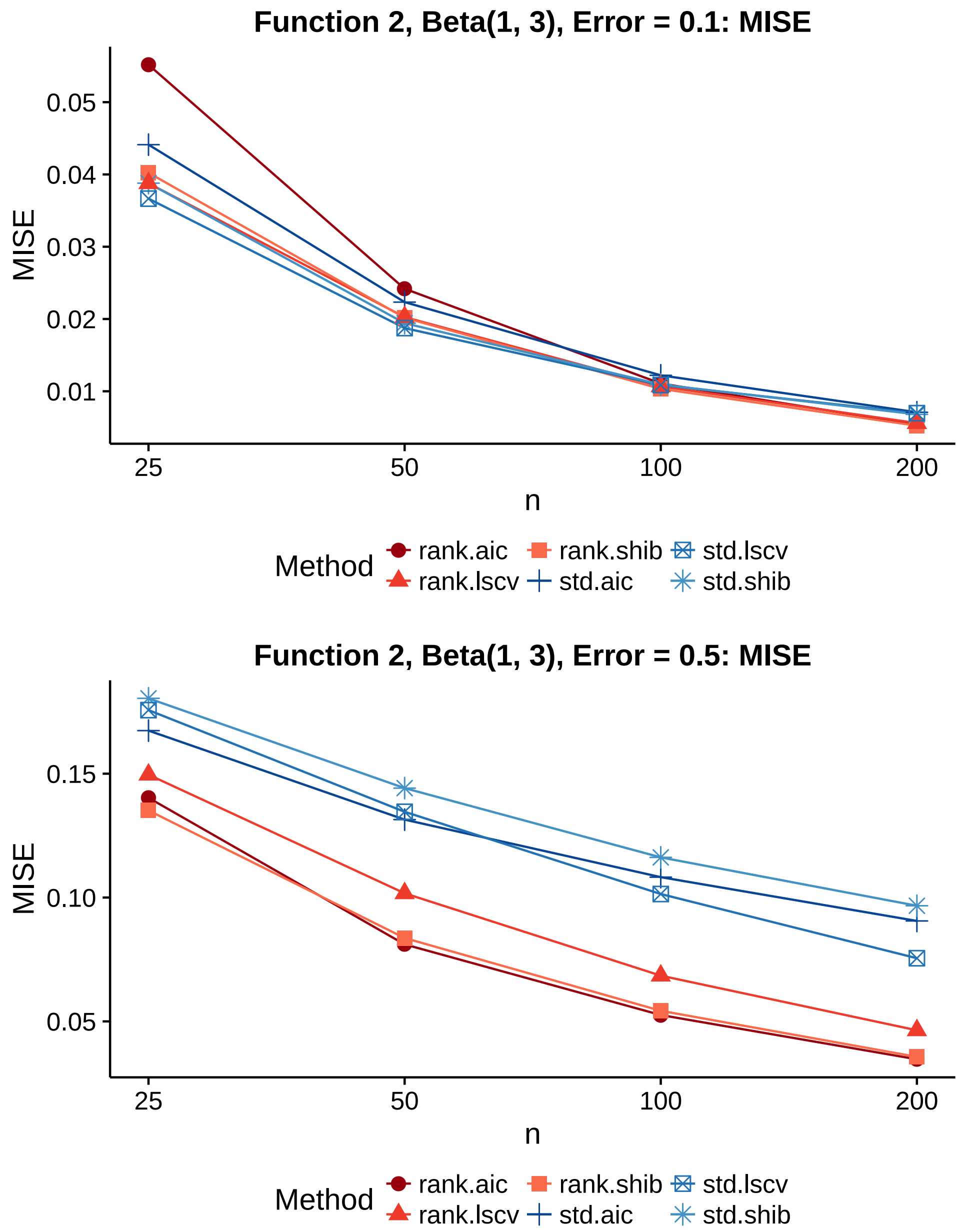
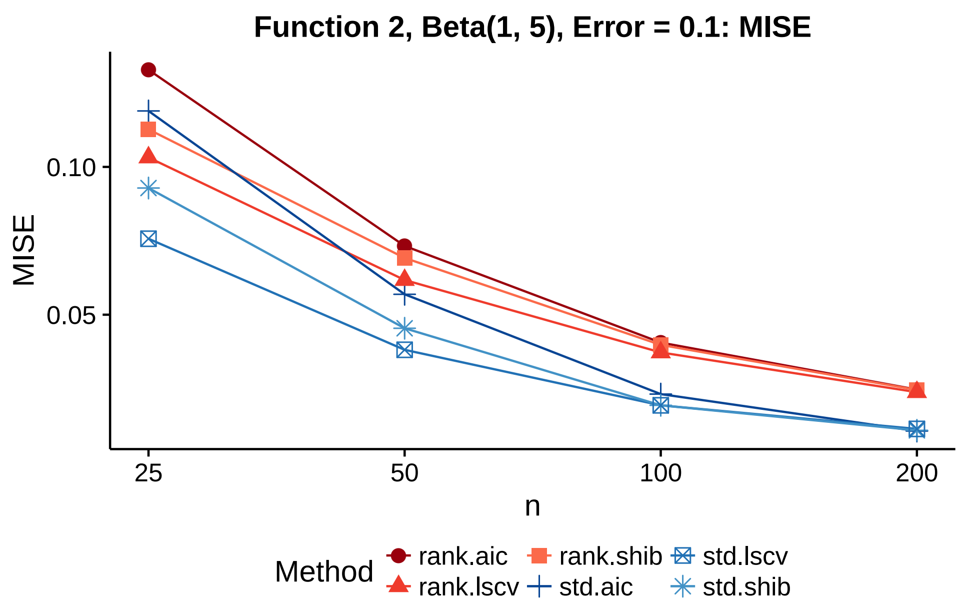


Figure A.28: MISEs for the NW estimator and RWNW estimator under the indicated bandwidth methods in the legend. Results are for function 2 under the Beta(1, 3) distribution.



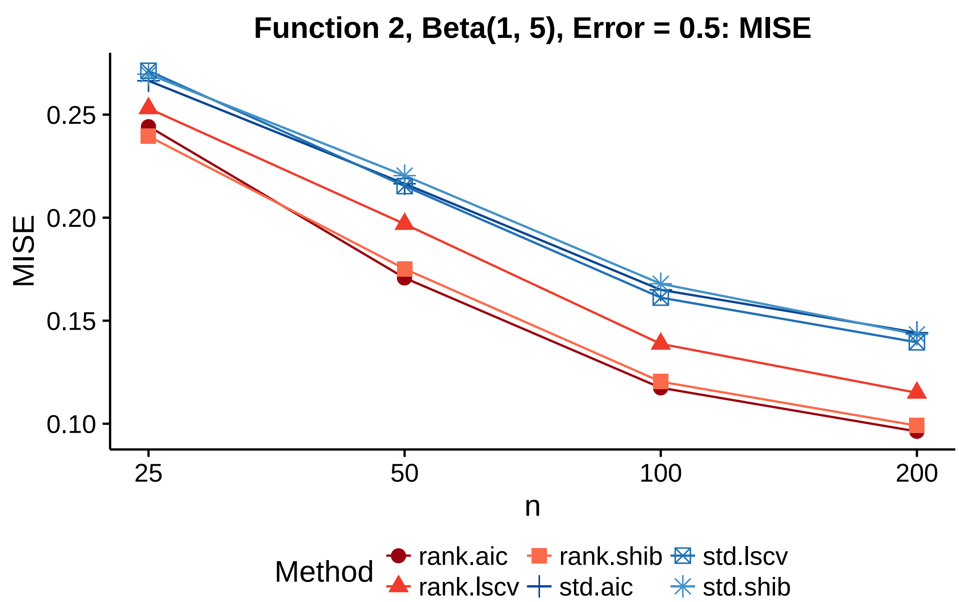
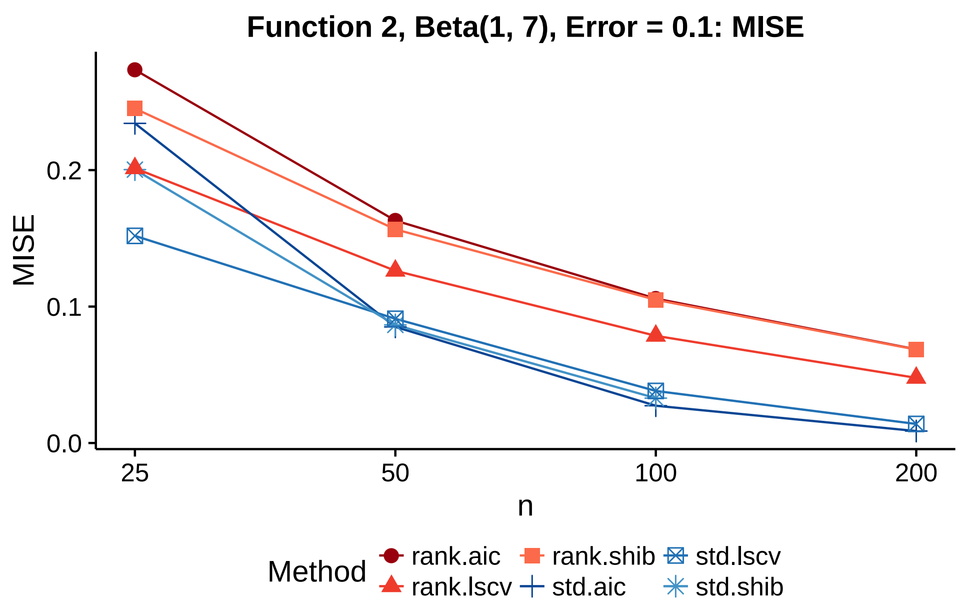


Figure A.29: MISEs for the NW estimator and RWNW estimator under the indicated bandwidth methods in the legend. Results are for function 2 under the Beta(1, 5) distribution.



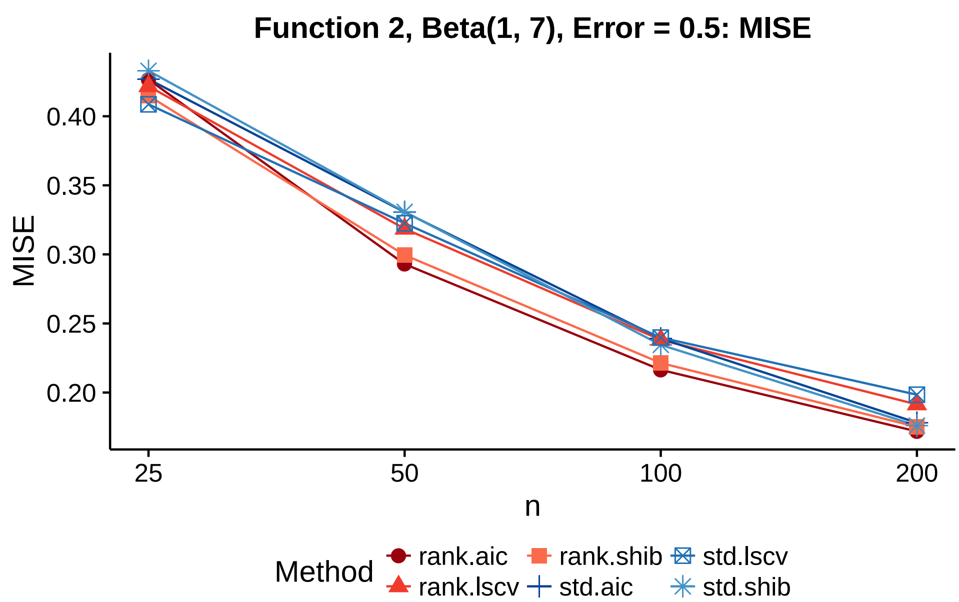


Figure A.30: MISEs for the NW estimator and RWNW estimator under the indicated bandwidth methods in the legend. Results are for function 2 under the Beta(1, 7) distribution.

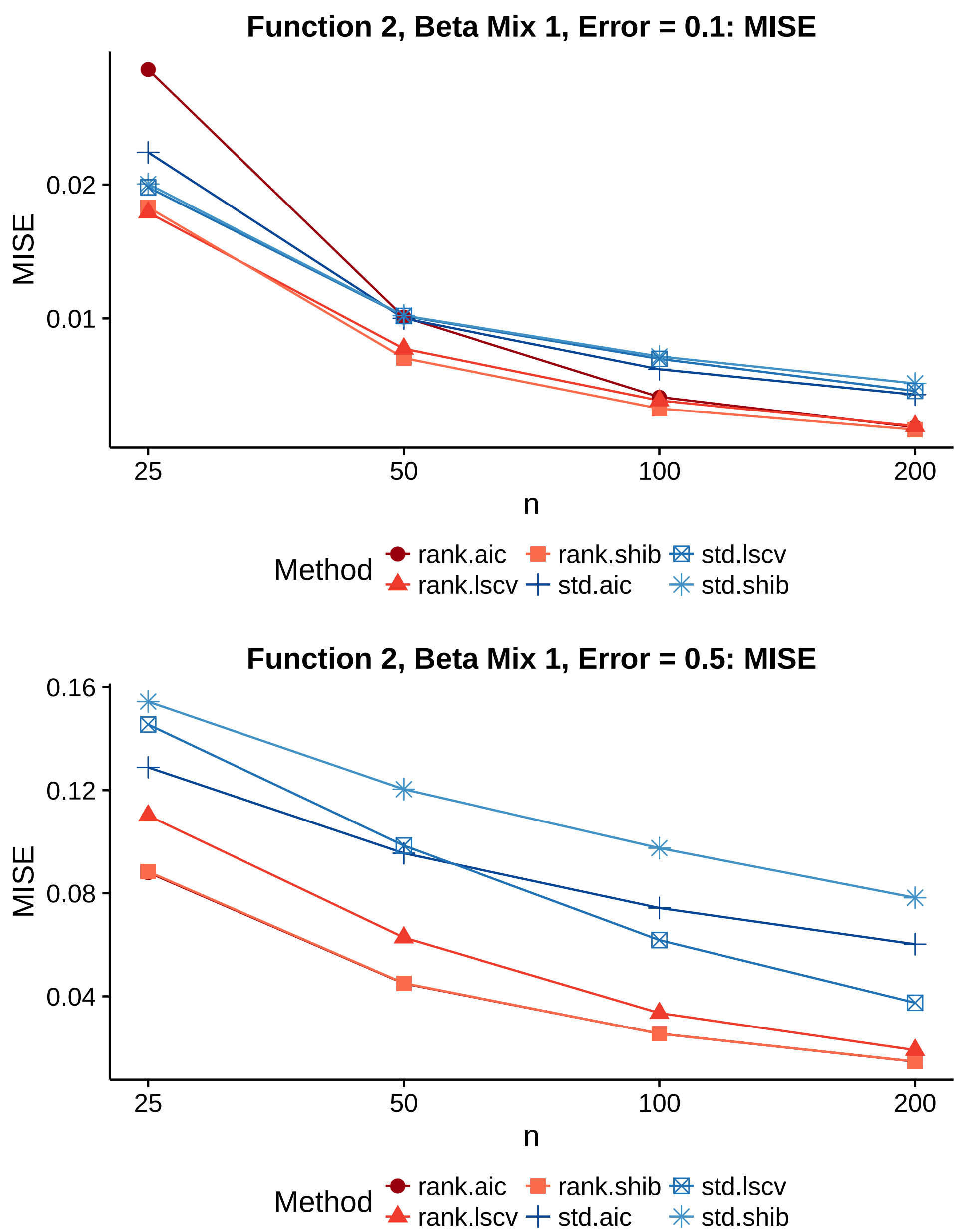
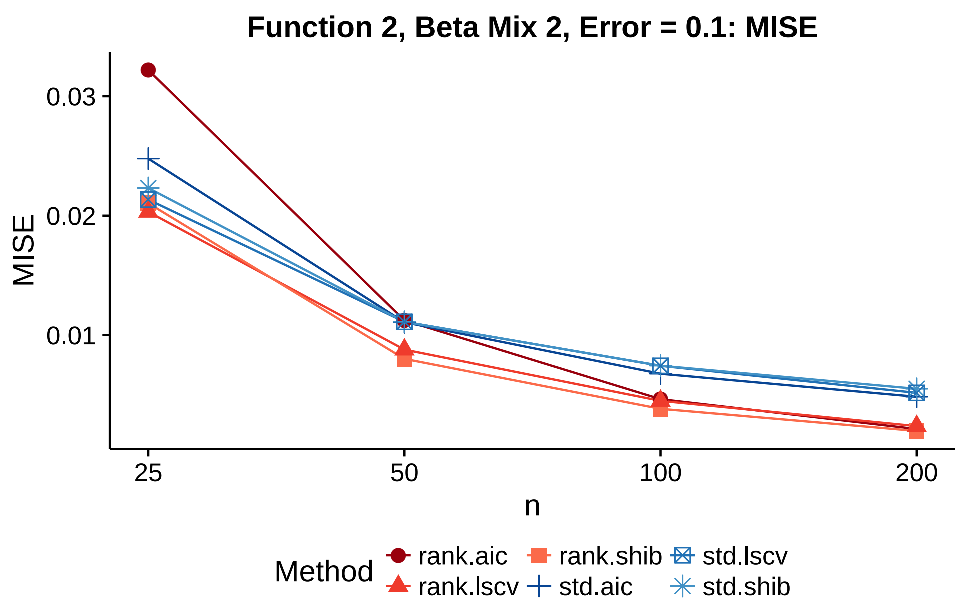


Figure A.31: MISEs for the NW estimator and RWNW estimator under the indicated bandwidth methods in the legend. Results are for function 2 under the beta mixture 1.



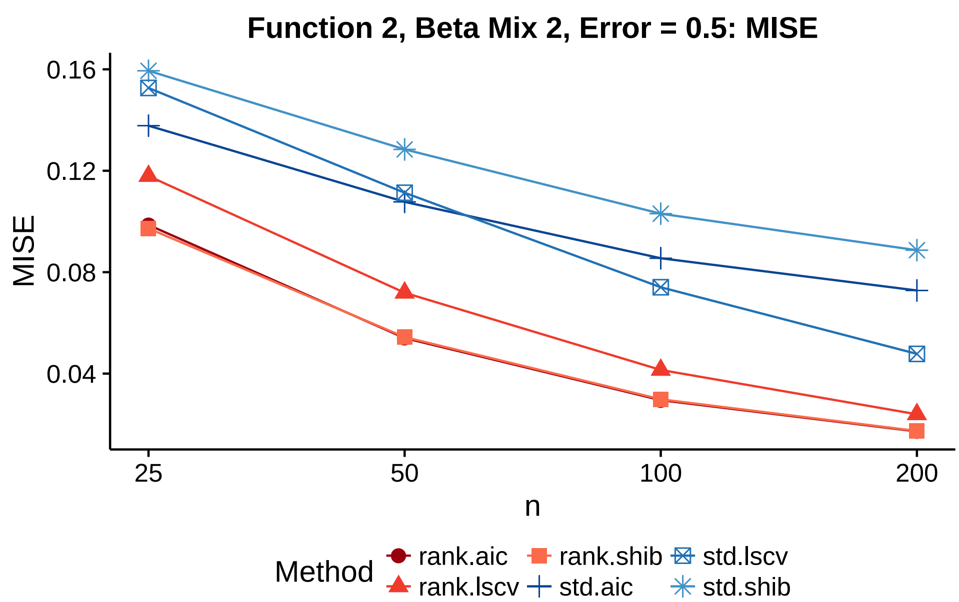
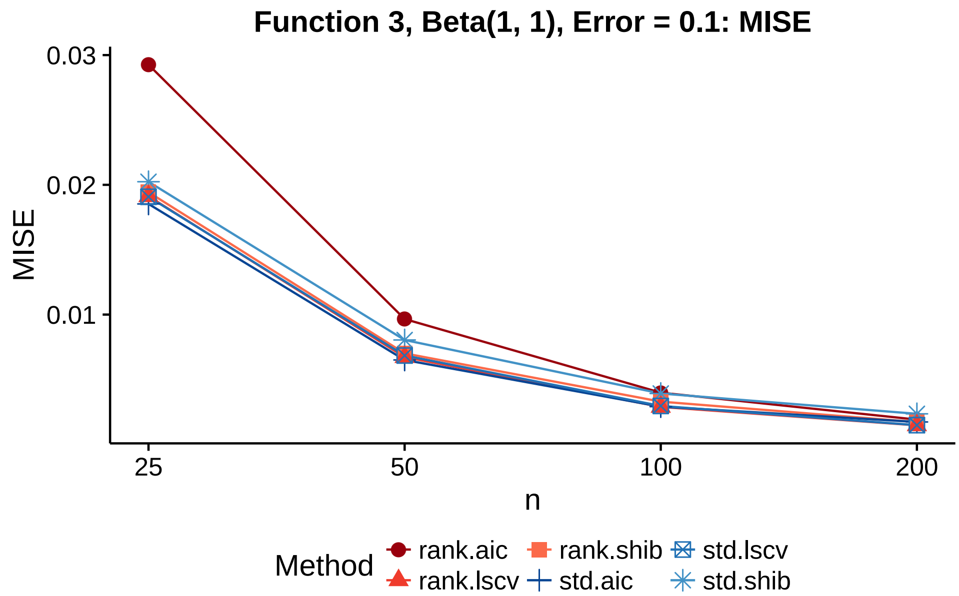


Figure A.32: MISEs for the NW estimator and RWNW estimator under the indicated bandwidth methods in the legend. Results are for function 2 under the beta mixture 2.



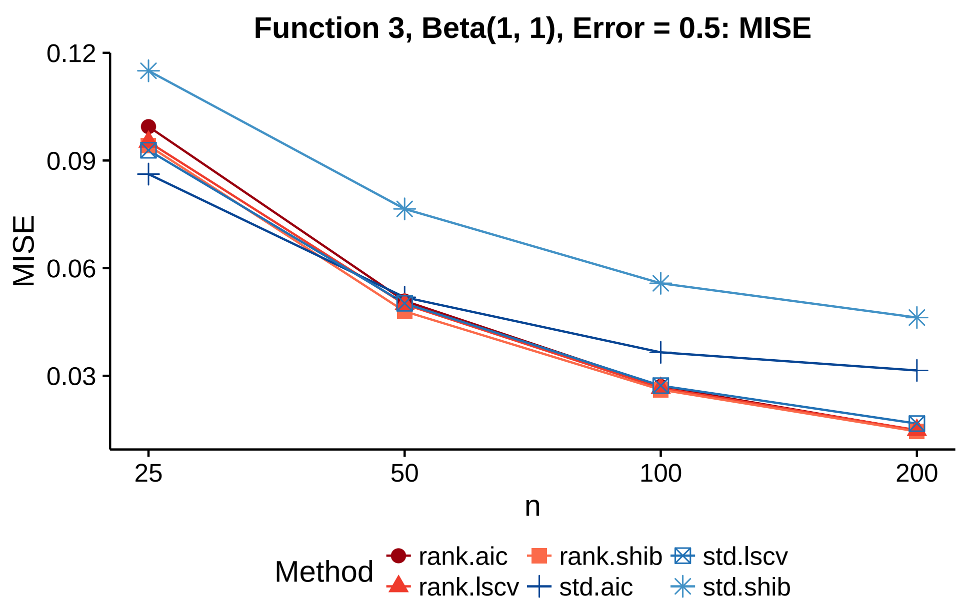


Figure A.33: MISEs for the NW estimator and RWNW estimator under the indicated bandwidth methods in the legend. Results are for function 3 under the Beta(1, 1) distribution.

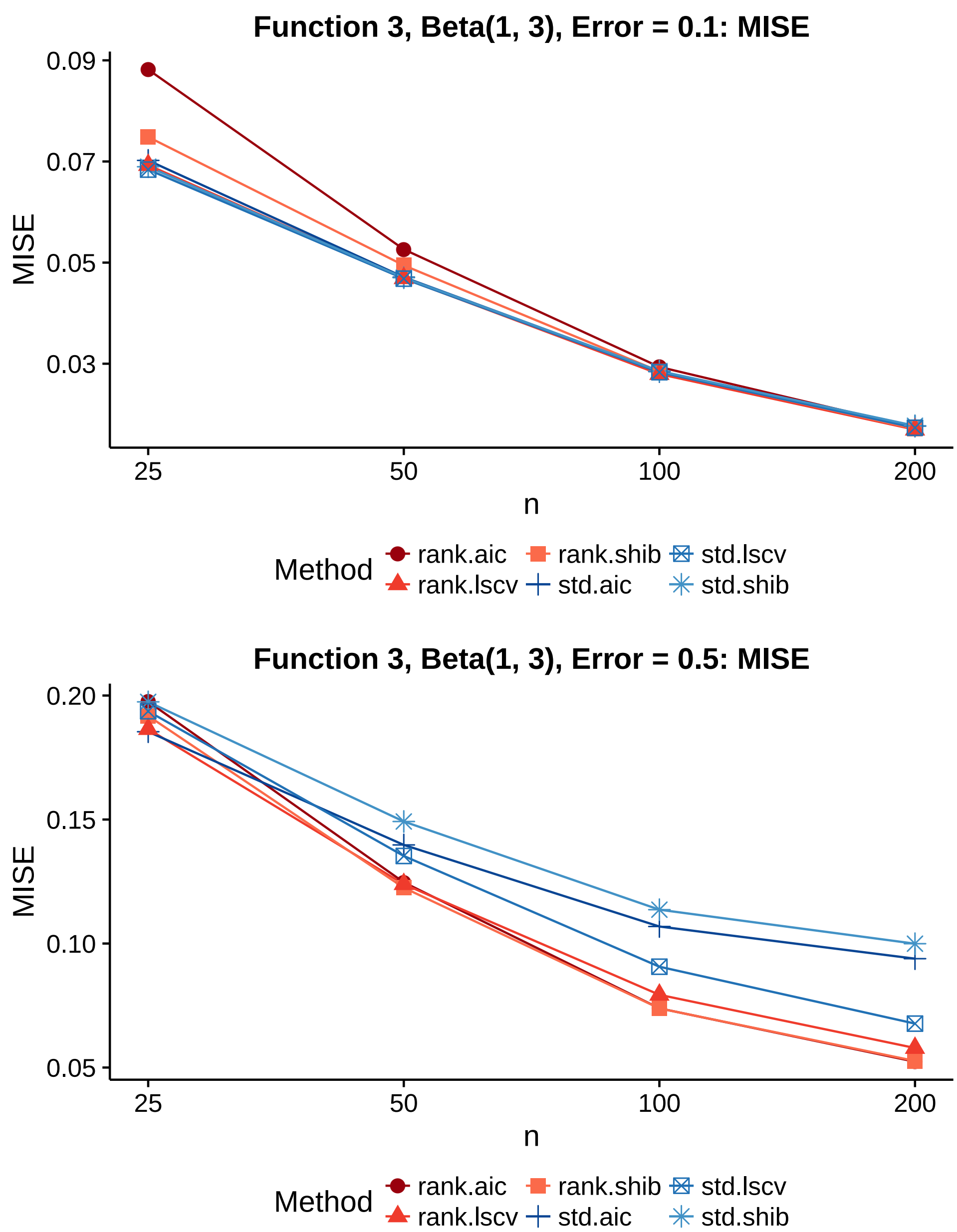
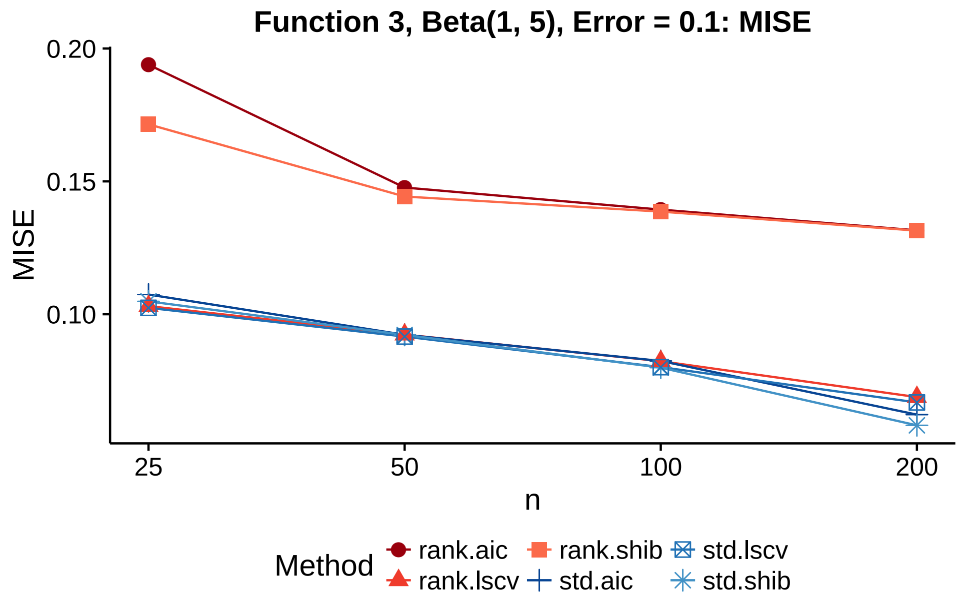


Figure A.34: MISEs for the NW estimator and RWNW estimator under the indicated bandwidth methods in the legend. Results are for function 3 under the Beta(1, 3) distribution.



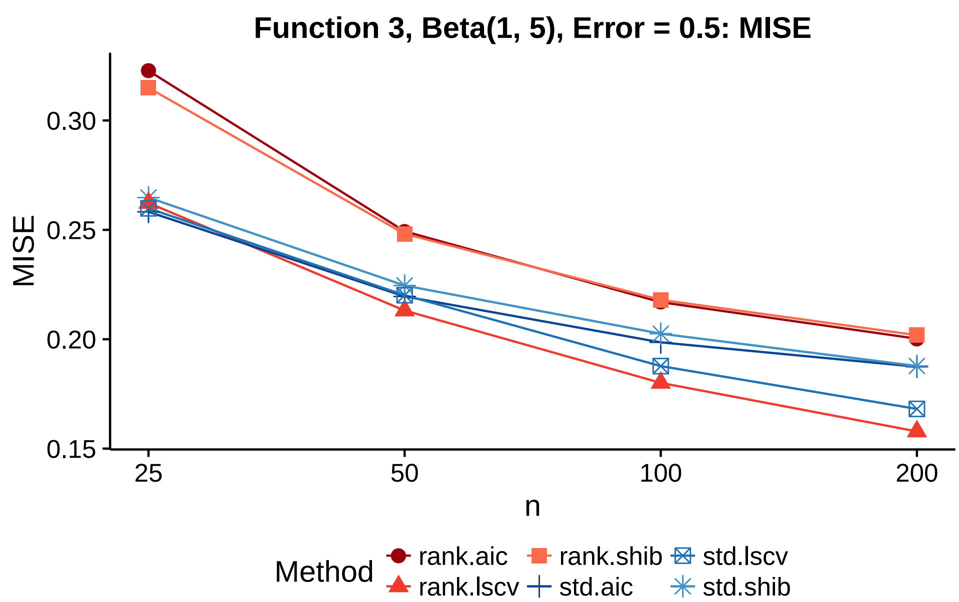


Figure A.35: MISEs for the NW estimator and RWNW estimator under the indicated bandwidth methods in the legend. Results are for function 3 under the Beta(1, 5) distribution.

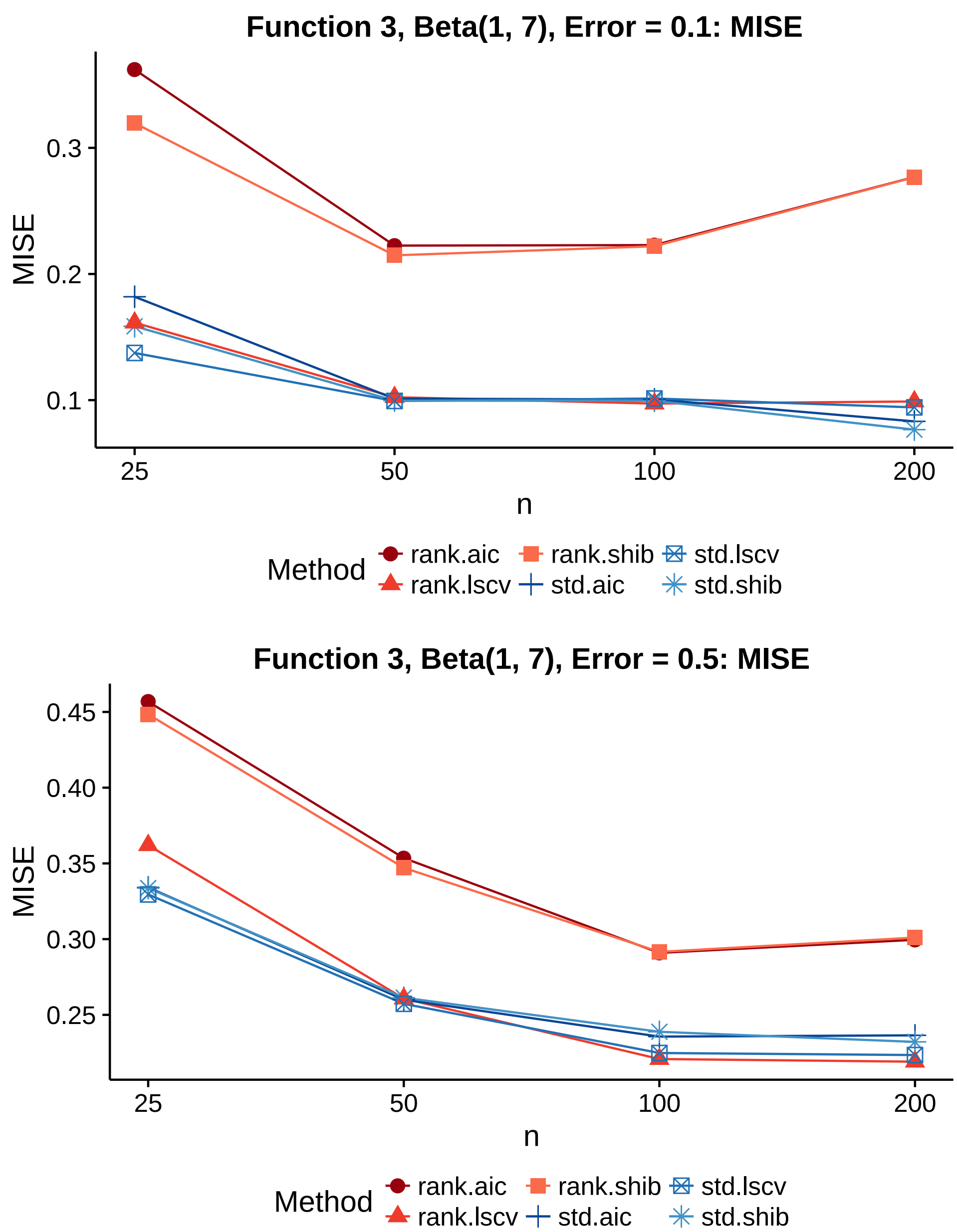
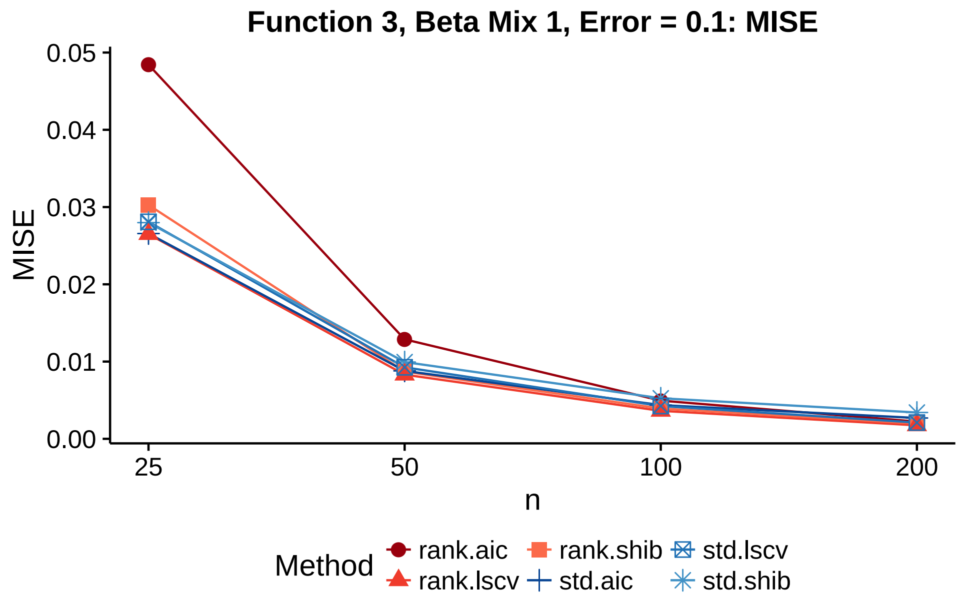


Figure A.36: MISEs for the NW estimator and RWNW estimator under the indicated bandwidth methods in the legend. Results are for function 3 under the Beta(1, 7) distribution.



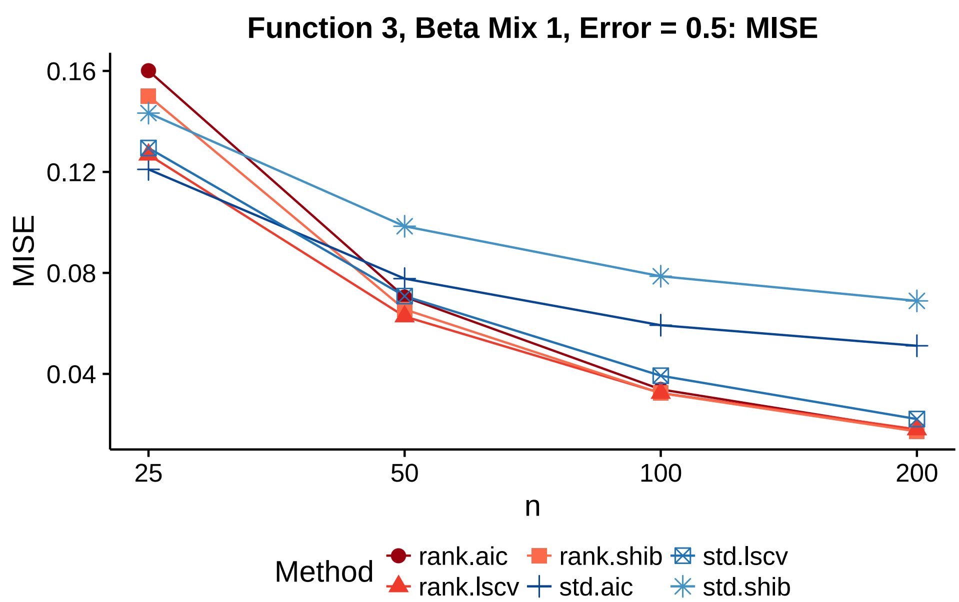
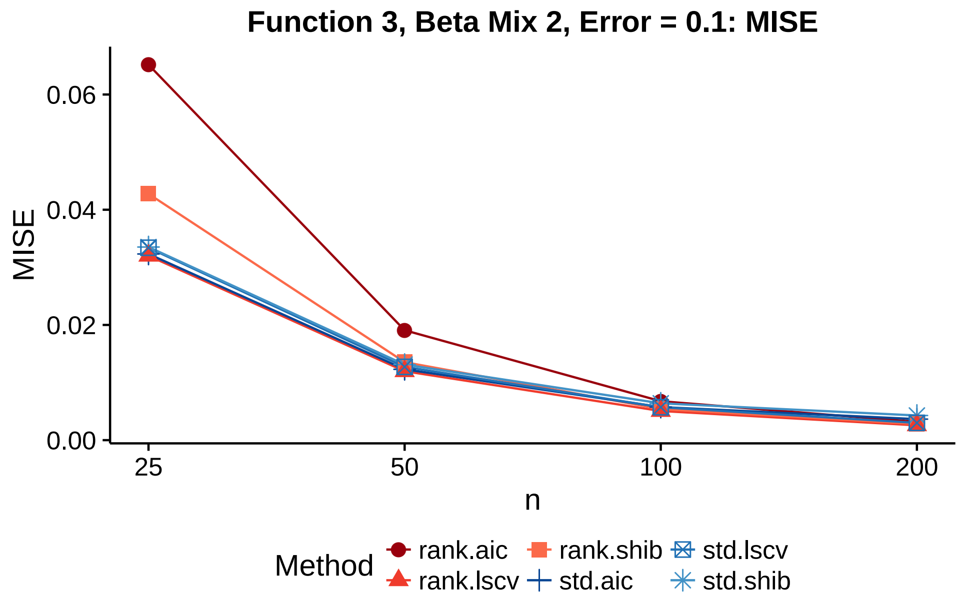


Figure A.37: MISEs for the NW estimator and RWNW estimator under the indicated bandwidth methods in the legend. Results are for function 3 under the beta mixture 1.



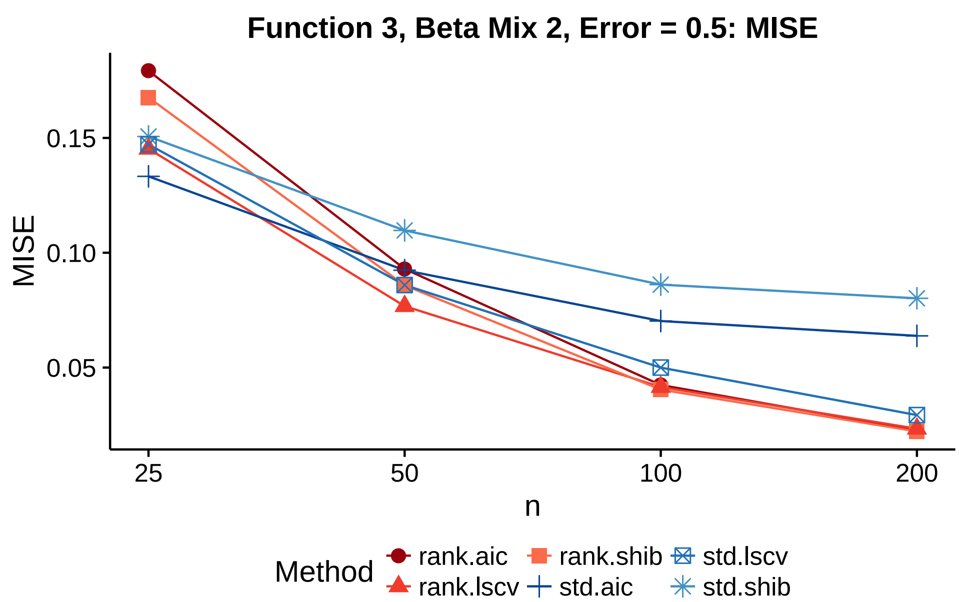


Figure A.38: MISEs for the NW estimator and RWNW estimator under the indicated bandwidth methods in the legend. Results are for function 3 under the beta mixture 2.

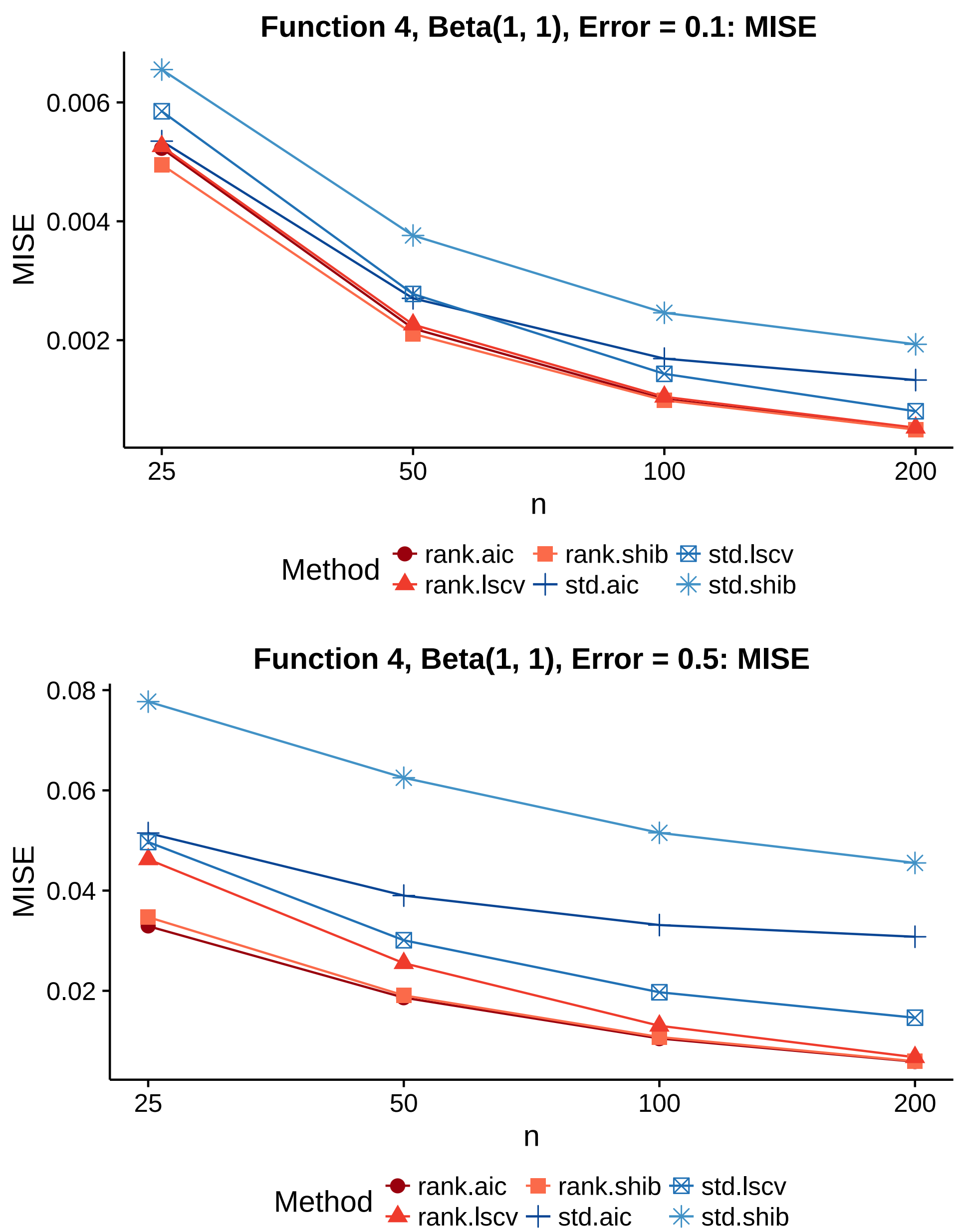
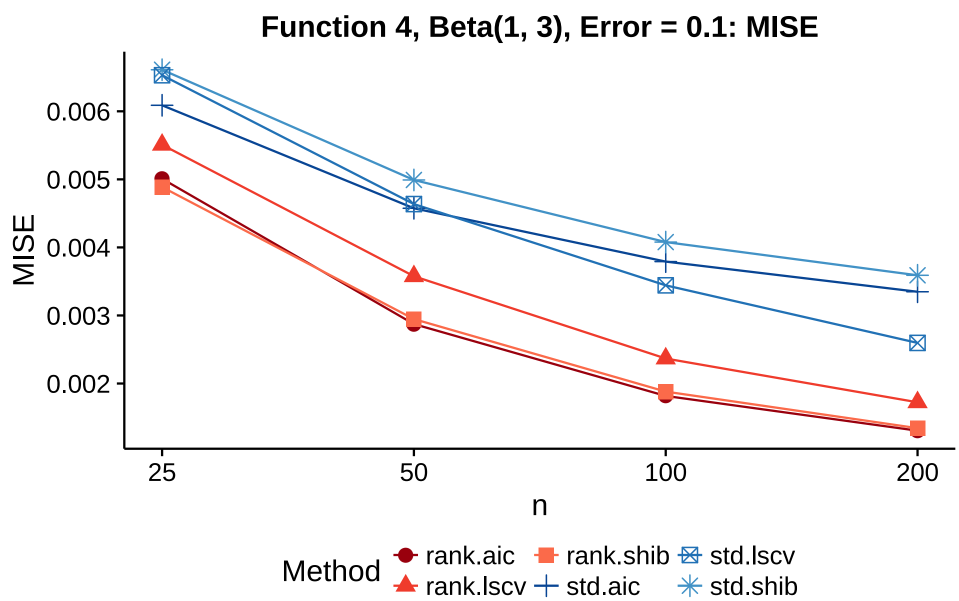


Figure A.39: MISEs for the NW estimator and RWNW estimator under the indicated bandwidth methods in the legend. Results are for function 4 under the Beta(1, 1) distribution.



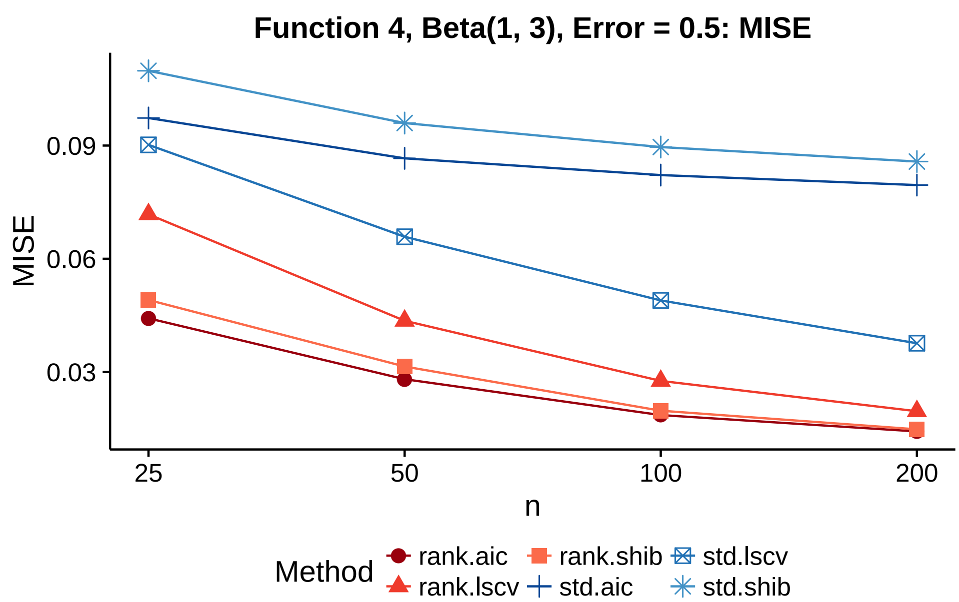
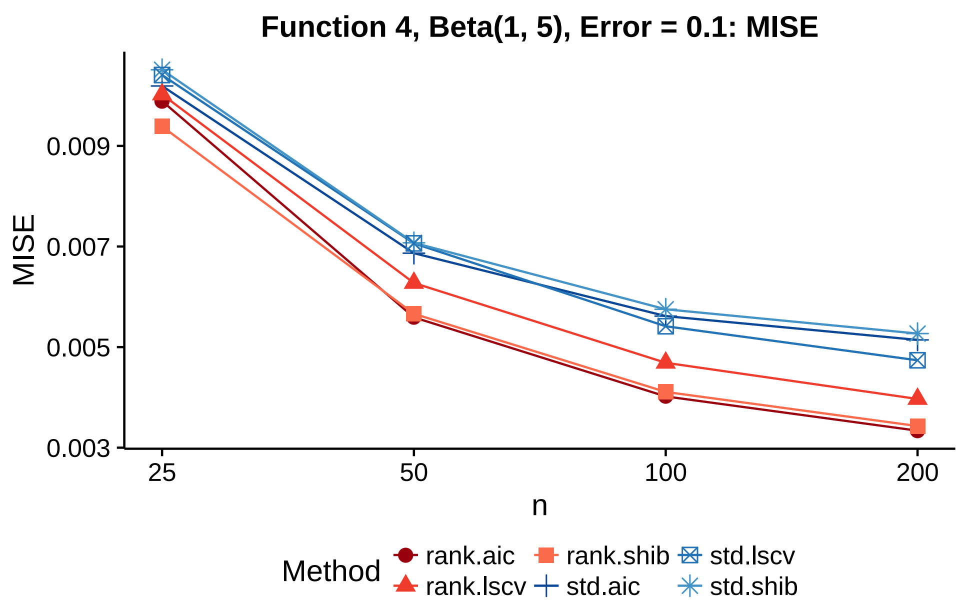


Figure A.40: MISEs for the NW estimator and RWNW estimator under the indicated bandwidth methods in the legend. Results are for function 4 under the Beta(1, 3) distribution.



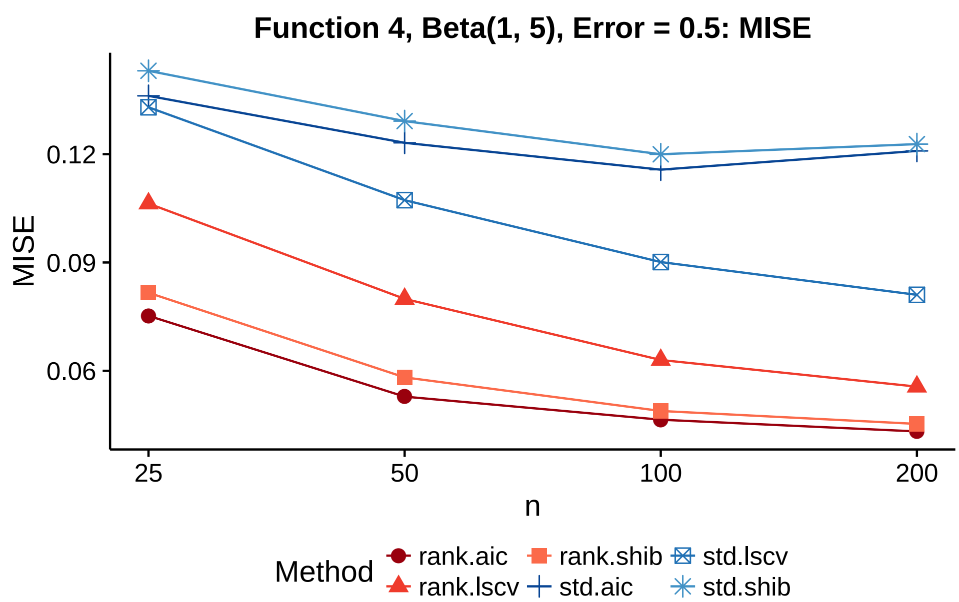
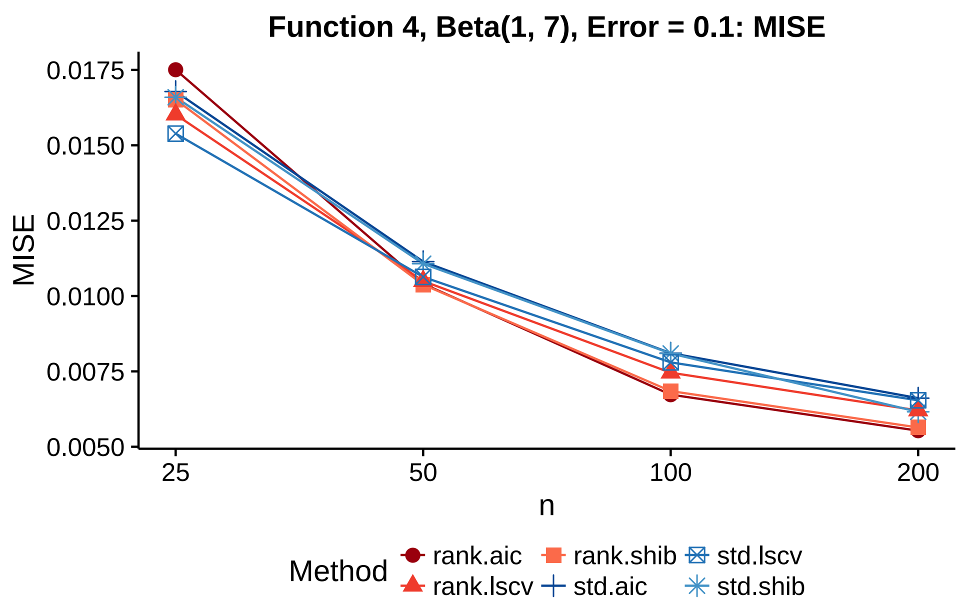


Figure A.41: MISEs for the NW estimator and RWNW estimator under the indicated bandwidth methods in the legend. Results are for function 4 under the Beta(1, 5) distribution.



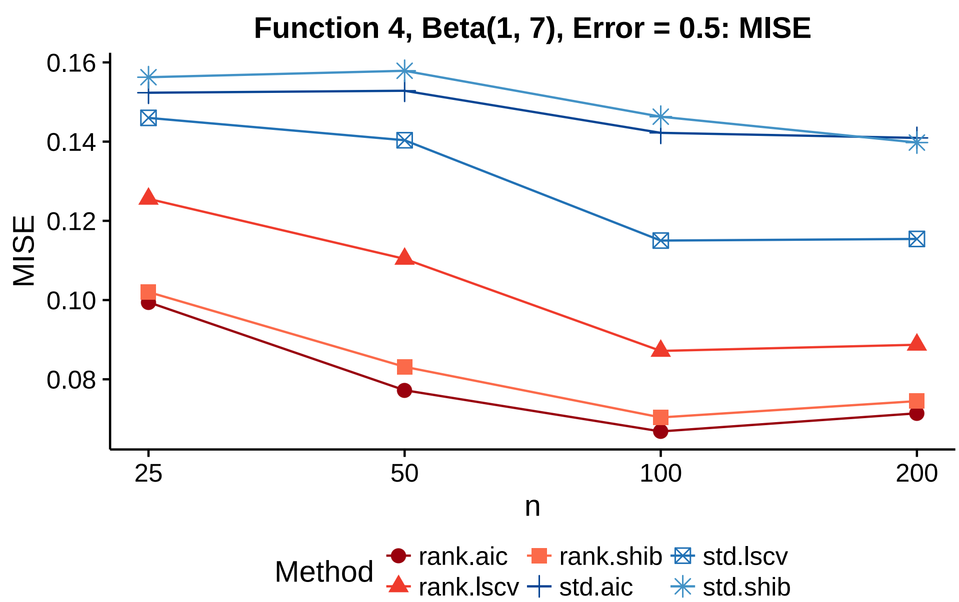
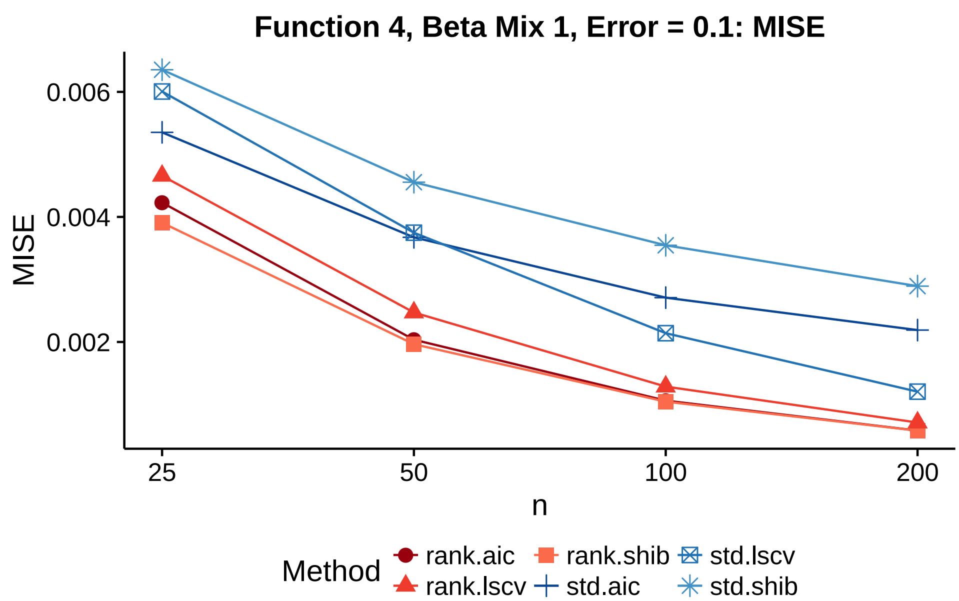


Figure A.42: MISEs for the NW estimator and RWNW estimator under the indicated bandwidth methods in the legend. Results are for function 4 under the Beta(1, 7) distribution.



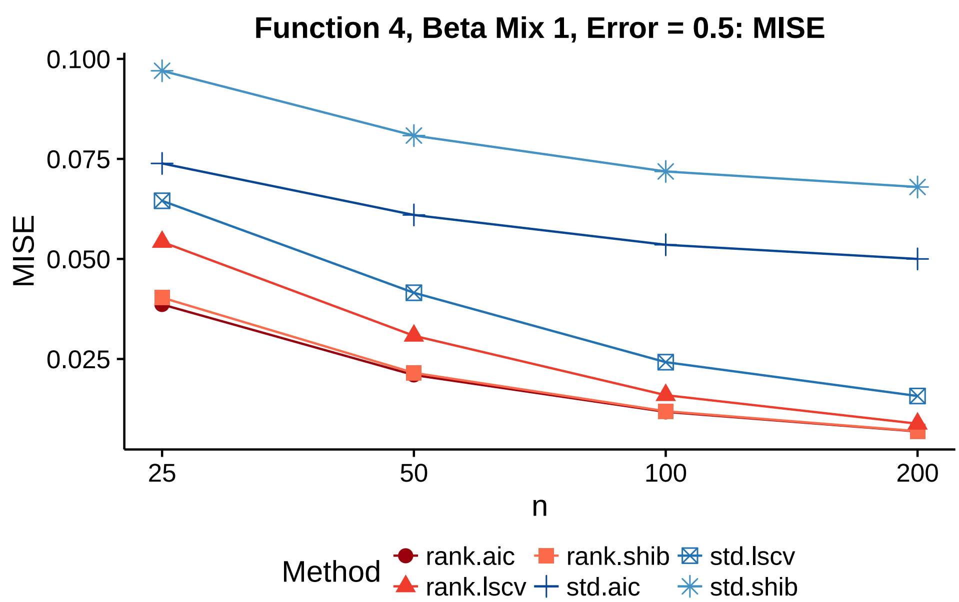
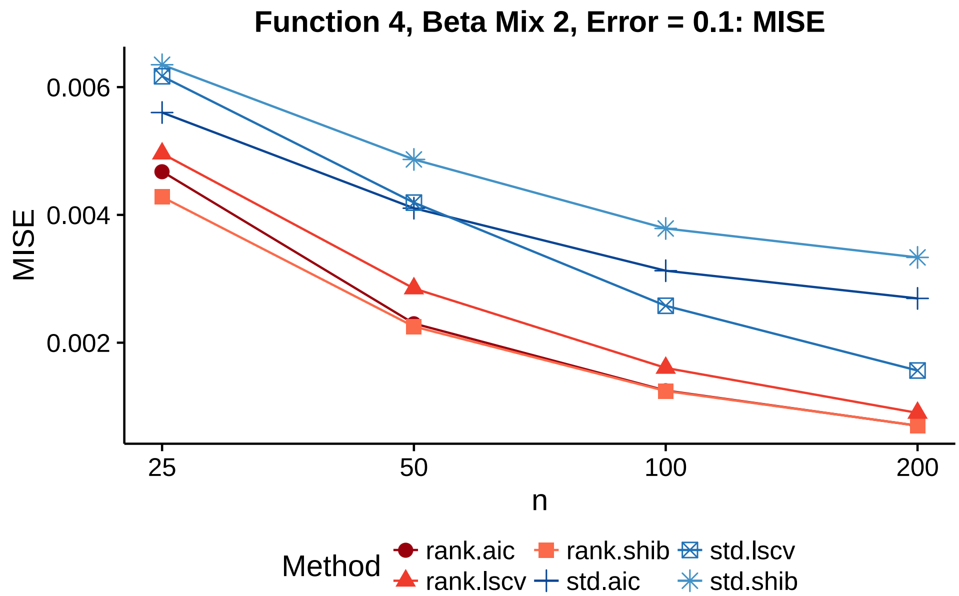


Figure A.43: MISEs for the NW estimator and RWNW estimator under the indicated bandwidth methods in the legend. Results are for function 4 under the beta mixture 1.



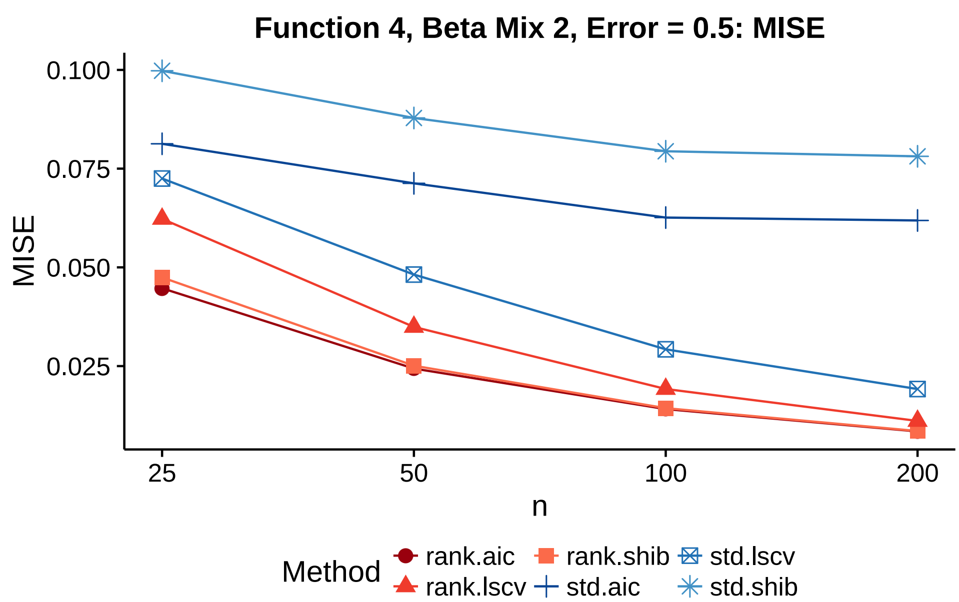


Figure A.44: MISEs for the NW estimator and RWNW estimator under the indicated bandwidth methods in the legend. Results are for function 4 under the beta mixture 2.

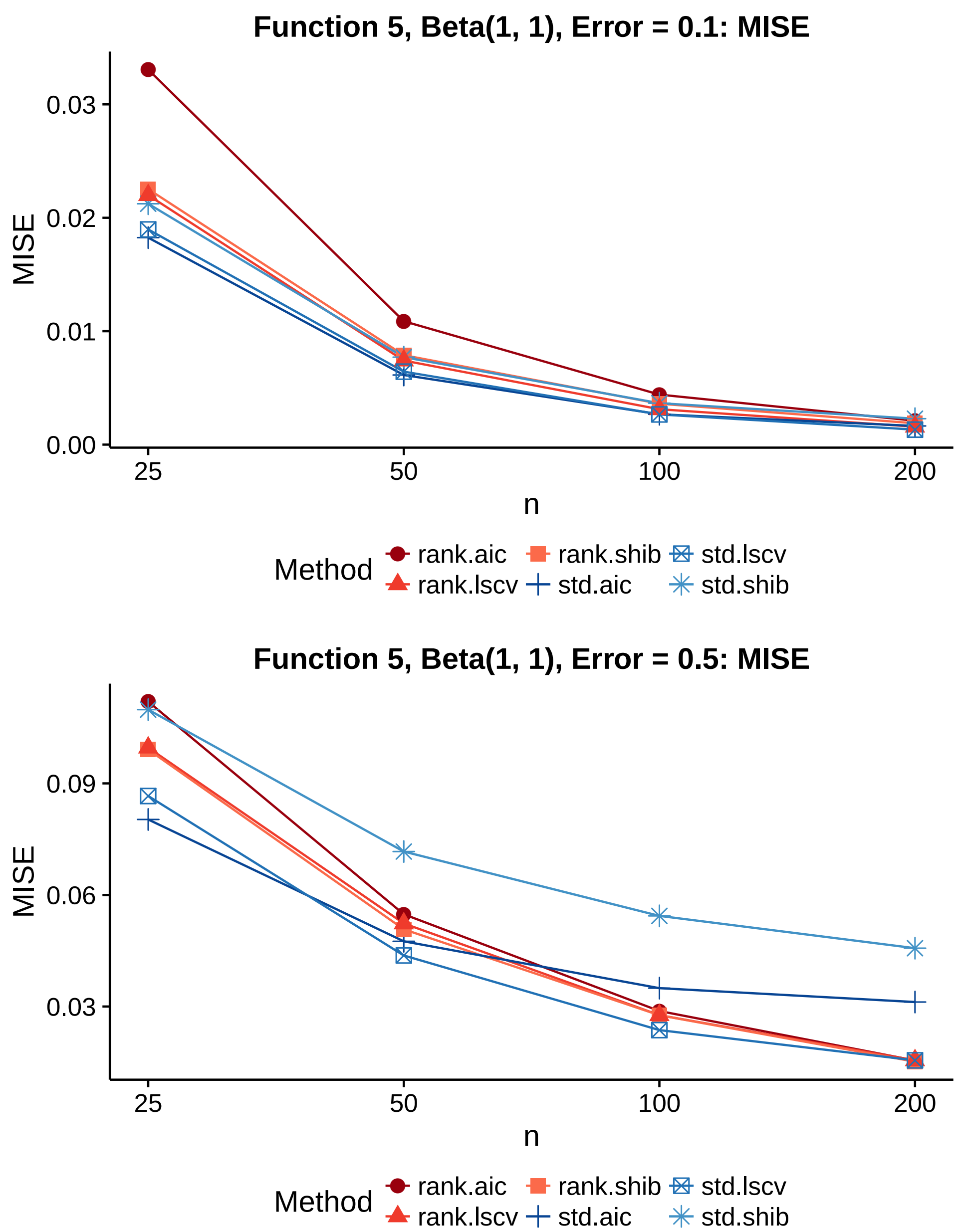
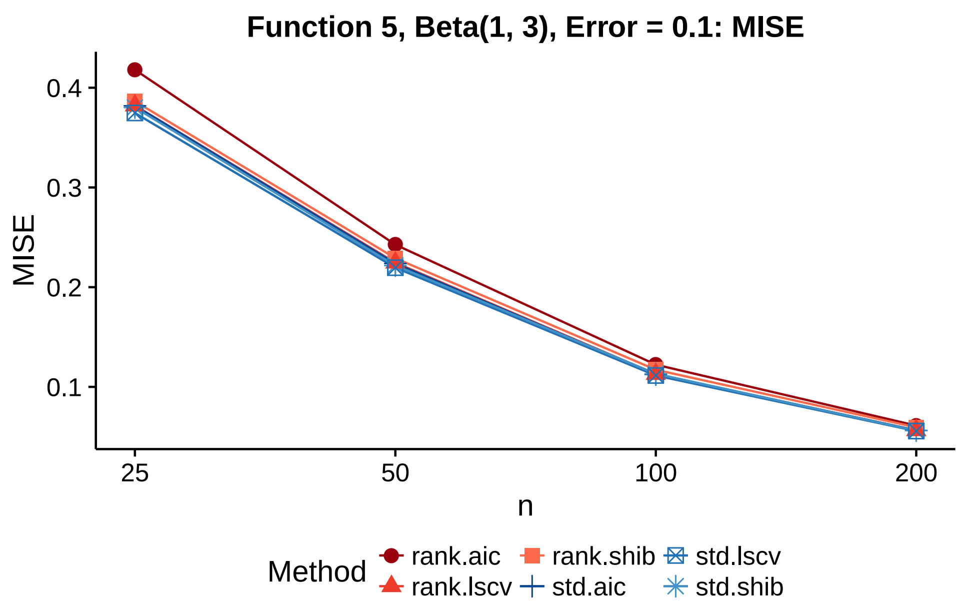


Figure A.45: MISEs for the NW estimator and RWNW estimator under the indicated bandwidth methods in the legend. Results are for function 5 under the Beta(1, 1) distribution.



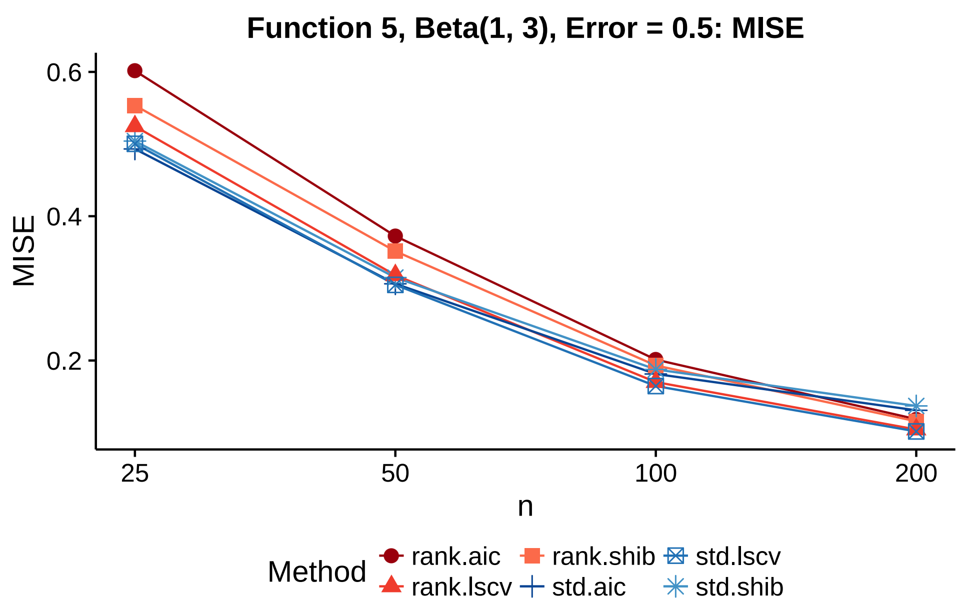
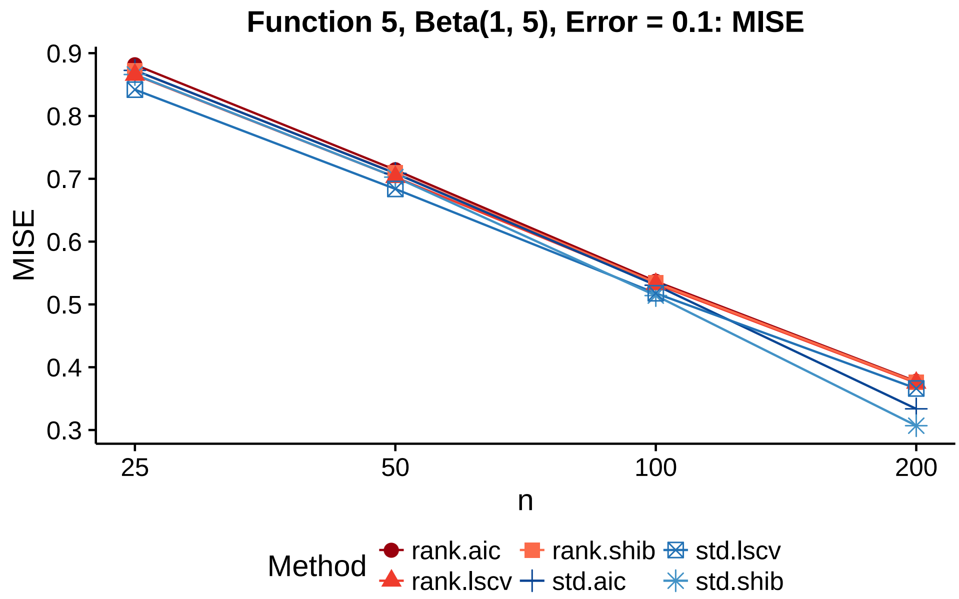


Figure A.46: MISEs for the NW estimator and RWNW estimator under the indicated bandwidth methods in the legend. Results are for function 5 under the Beta(1, 3) distribution.



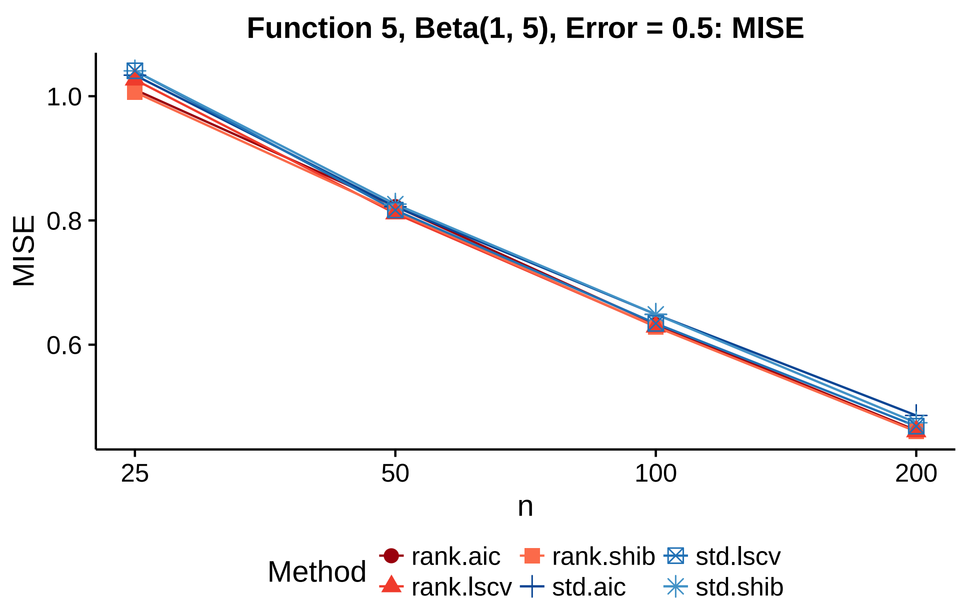
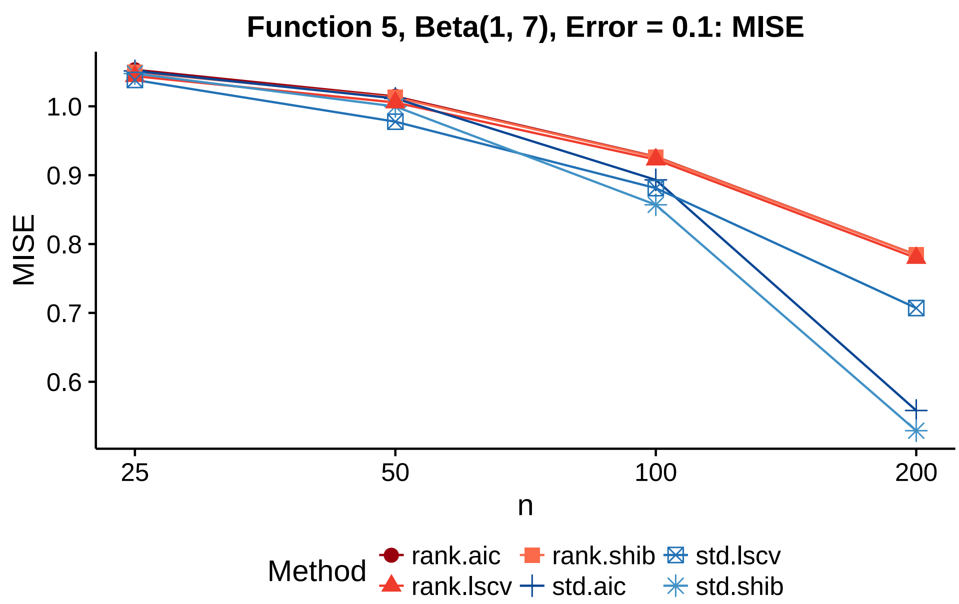


Figure A.47: MISEs for the NW estimator and RWNW estimator under the indicated bandwidth methods in the legend. Results are for function 5 under the Beta(1, 5) distribution.



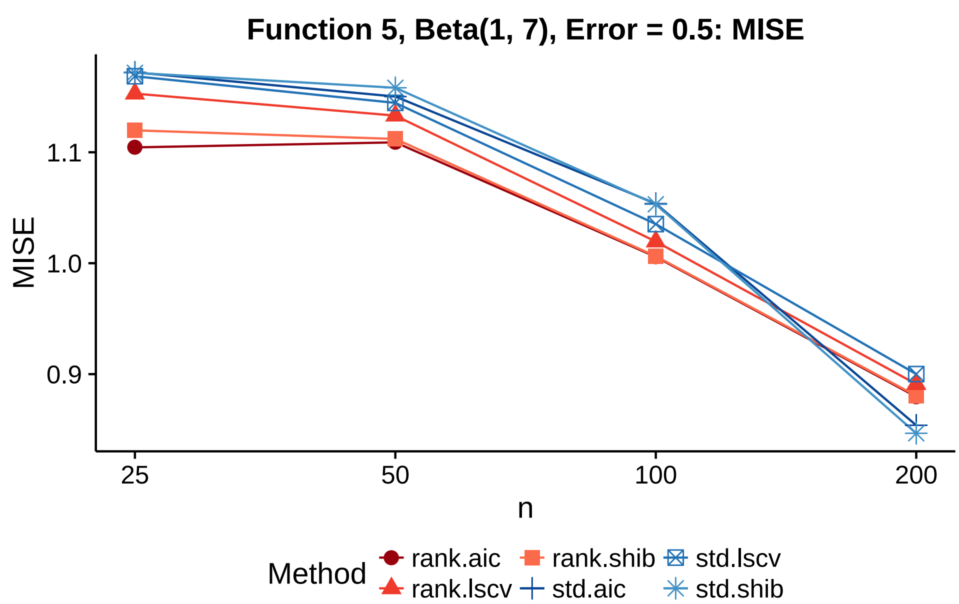
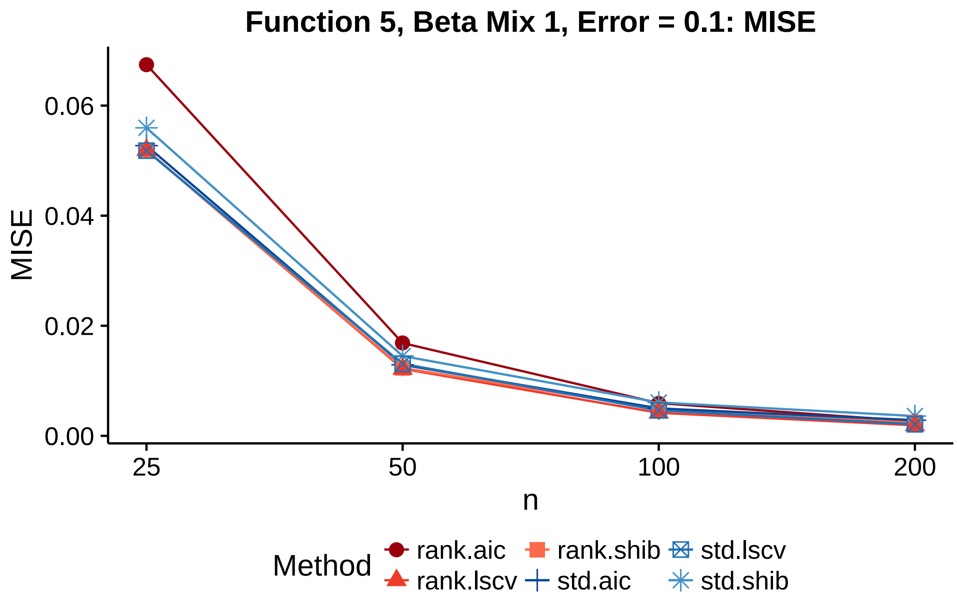


Figure A.48: MISEs for the NW estimator and RWNW estimator under the indicated bandwidth methods in the legend. Results are for function 5 under the Beta(1, 7) distribution.



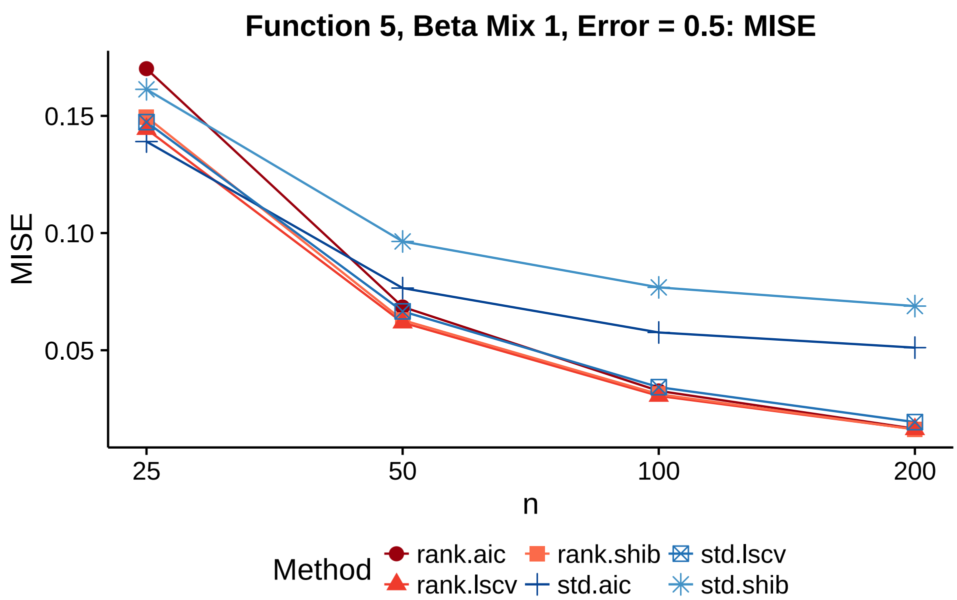
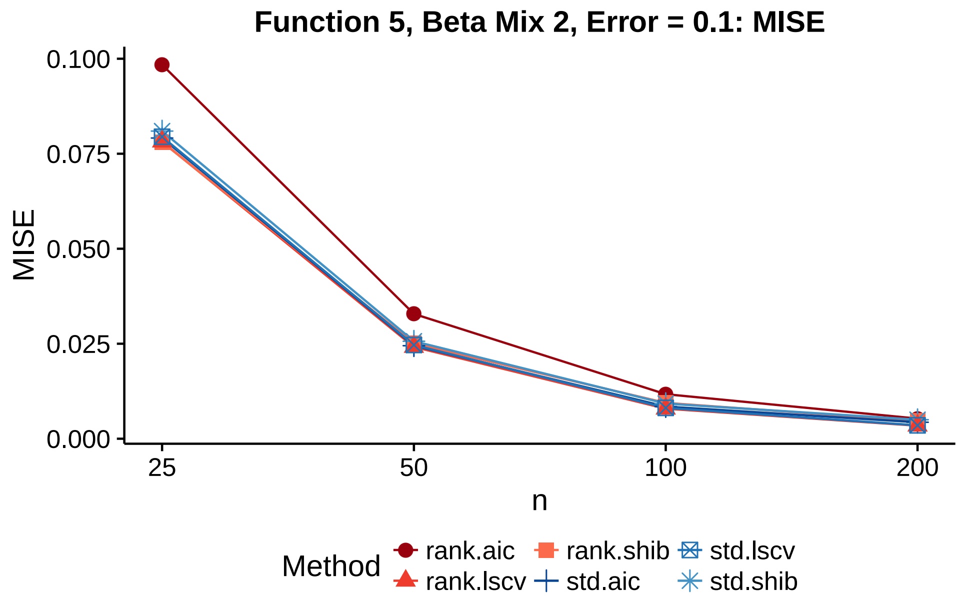


Figure A.49: MISEs for the NW estimator and RWNW estimator under the indicated bandwidth methods in the legend. Results are for function 5 under the beta mixture 1.



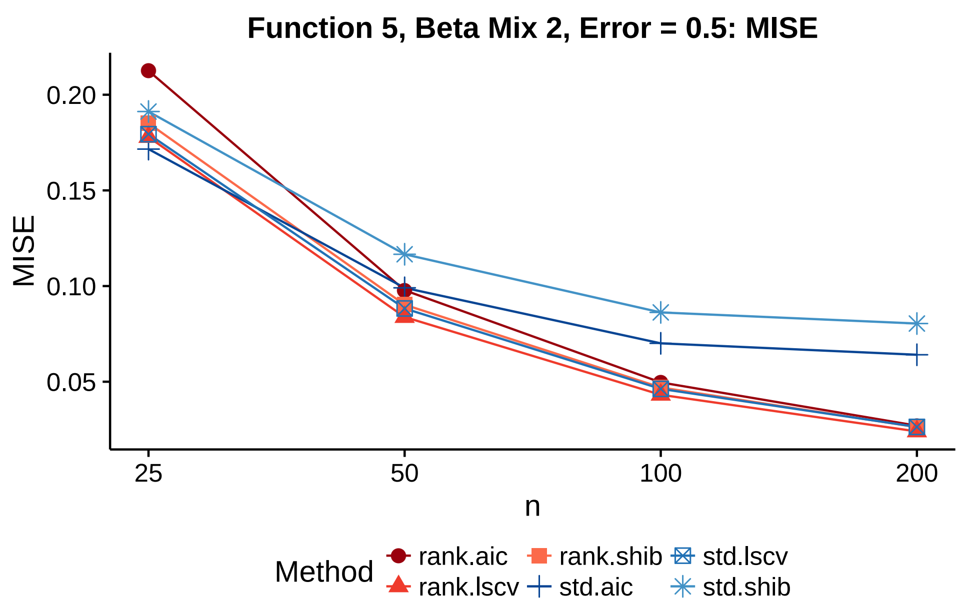


Figure A.50: MISEs for the NW estimator and RWNW estimator under the indicated bandwidth methods in the legend. Results are for function 5 under the beta mixture 2.

VITA

Nicholas Armstrong Kaukis

Candidate for the Degree of

Doctor of Philosophy

Thesis: UTILIZING OBSERVATIONS AND RANKS IN KERNEL ESTIMATION

Major Field: Statistics

Biographical:

Education:

Completed the requirements for the Doctor of Philosophy in statistics at Oklahoma State University, Stillwater, Oklahoma in December, 2016.

Completed the requirements for the Master of Science in statistics at Oklahoma State University, Stillwater, Oklahoma in 2014.

Completed the requirements for the Bachelor of Science/Arts in statistics at Oklahoma State University, Stillwater, Oklahoma in 2009.

Experience:

- Instructor of record at Oklahoma State University, Department of Statistics for the following courses (2009-2016):
    - Elementary Statistics (STAT-2013)
    - Elementary Statistics for Business and Economics (STAT-2023)
    - Elementary Statistics for Social Sciences (STAT-2053)
    - Intermediate Statistical Analysis (STAT-3013)
    - Statistical Methods I (STAT-4013)
    - Statistical Methods for Social Sciences I (STAT-4053)
    - Engineering Statistics (STAT-4033)
    - Engineering Statistics with Experimental Design (STAT-4073)

- Statistics Club Vice President (2014-15)

- Statistics Club Grad. Student Government Rep. (2009-2010)

Professional Memberships:

American Statistical Association (since 2013)