Investigating Confidence Interval Coverage for Inference Using Penalized Splines in Mixed Linear Models

1 Introduction

In a mixed linear model with predictors, $\mathbf{X}$, and response, $\mathbf{y}$, we assume the relationship $\mathbf{y} = f(\mathbf{X}) + \epsilon$ for an unknown smooth function, $f(\mathbf{X}) = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u}$, and independent error term, $\epsilon \overset{i.i.d}{\sim} N(0, \sigma_\epsilon^2)$. Often we use penalized splines to fit this relationship with the goal of making inference on observed $\mathbf{x}_i$, for $i = 1, \ldots, n$, or perhaps predictions for a new $\mathbf{x}^\star$. In a conventional analysis, we must first specify our choice of basis, knots, and penalty. Ruppert et al. (2003) emphasize the truncated power bases and Hodges (2014) supports this with further remarks that higher-order polynomials allow a smoother fit between knots. For this paper, I use a truncated cubic basis with equally-spaced knots and a simple quadratic penalty of the form $\lambda^2 (\boldsymbol{\beta}, \mathbf{u})^T \mathbf{D} \binom{\boldsymbol{\beta}}{\mathbf{u}}$, where $\lambda^2 = \sigma_\epsilon^2/\sigma_s^2$, $\mathbf{D}$ is a diagonal matrix with either a 0 on the diagonal entry, corresponding to the fixed effects $\boldsymbol{\beta}$, or a 1 on the diagonal entry, corresponding to the random effects $\mathbf{u}$, and σ_s^2 is a smoothing variance for $\mathbf{u}$. For smaller σ_s^2 the random effects are shrunk more toward zero.

Now, we can maximize the restricted log likelihood to obtain estimates for $\boldsymbol{\beta}$ and $\mathbf{u}$. Using the estimates of $\boldsymbol{\beta}$ and $\mathbf{u}$, we can calculate our fitted values, $\hat{f}(\mathbf{x}_i)$ for each i. However in this framework, $f(\mathbf{x}_i)$ is random due to the randomness of $\mathbf{u}$. Variance estimates for $f(\mathbf{x}_i)$ then differ depending on whether the randomness of $\mathbf{u}$ is taken into account. So, how should we construct confidence intervals (CIs) for $\hat{f}(\mathbf{x}_i)$? Ruppert et al. (2003) argue that $\mathbf{u}$ is a modeling device used to capture curvature, thus variance calculations should be carried out treating $\mathbf{u}$ as fixed but unknown. If we treat $\mathbf{u}$ as fixed but unknown, we introduce bias. We can denote the conditional bias as $\widehat{bias}(x \mid \mathbf{u})$, which is given by

$$E[\hat{f}(\mathbf{X}) - f(\mathbf{X})|\mathbf{u}] = -r\mathbf{C}(\mathbf{C}'\mathbf{C} + r\mathbf{I})^{-1} \begin{pmatrix} \mathbf{0_P} \\ \mathbf{u} \end{pmatrix}, \tag{1}$$

where $r = \hat{\sigma}_\epsilon^2/\hat{\sigma}_s^2$ and $\mathbf{C} = [\mathbf{X}|\mathbf{Z}]$. If we average over all $\mathbf{u}$ in equation (1), the bias turns out to be zero. This result has lead researchers to propose bias-correction methods that incorporate the conditional bias in either the variance estimate or the center (fitted value) of the confidence interval. Through a simulation study, this paper compares the coverage probabilities (CPs) of three different bias-correction methods.

2 Methods

I consider three methods proposed by Ruppert et al. (2003), Hodges (2014), and Sun and Loader (1994), as described below. The three methods under consideration differ in the adjustment of the CI to correct for the conditional bias.

2.1 Ruppert, Wand and Carroll (RWC)

To account for bias, Ruppert et al. (2003) suggest replacing the conditional variance with the conditional mean squared error, averaged over $\mathbf{u}$ (or the unconditional mean squared error). Assuming we have one predictor, the RWC CIs can be calculated as

$$\hat{f}(x) \pm z_{1-\frac{\alpha}{2}} \sqrt{\hat{E}\left[\{\hat{f}(x) - f(x)\}^2\right]},$$

where

$$\hat{E}\left[\{\hat{f}(x) - f(x)\}^2\right] = \mathbf{C}_i \widehat{Cov}\begin{pmatrix} \hat{\boldsymbol{\beta}} \\ \hat{\mathbf{u}} - \mathbf{u} \end{pmatrix} \mathbf{C}_i^T,$$

and

$$\widehat{Cov}\begin{pmatrix} \hat{\boldsymbol{\beta}} \\ \hat{\mathbf{u}} - \mathbf{u} \end{pmatrix} = \left(\frac{1}{\hat{\sigma}_\epsilon^2}\mathbf{C}^T\mathbf{C} + \begin{pmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \frac{1}{\hat{\sigma}_s^2} \end{pmatrix}\right)^{-1}.$$

This correction widens the confidence interval by the calculated conditional bias.

2.2 Hodges (JH)

Rather than increasing the CI variability, Hodges (2014) suggests using the conditional variance and alternatively shifting the center of the CIs up or down by the calculated conditional bias. The JH CIs are given by

$$[\hat{f}(x) - \widehat{bias}(x \mid \mathbf{u})] \pm z_{1-\frac{\alpha}{2}} \ \widehat{\text{st.dev.}}\{\hat{f}(x)|\mathbf{u}\},$$

where

$$\widehat{\text{var}}\{\hat{f}(x)|\mathbf{u}\} = \mathbf{C}_i \widehat{Cov}\left(\begin{array}{c|c} \hat{\boldsymbol{\beta}} & \\ & \mathbf{u} \\ \hat{\mathbf{u}} & \end{array}\right) \mathbf{C}_i^T,$$

and

$$\widehat{Cov}\left(\begin{array}{c|c} \hat{\boldsymbol{\beta}} & \\ & \mathbf{u} \\ \hat{\mathbf{u}} & \end{array}\right) = \frac{1}{\hat{\sigma}_\epsilon^2}\widehat{Cov}\begin{pmatrix} \hat{\boldsymbol{\beta}} \\ \hat{\mathbf{u}} - \mathbf{u} \end{pmatrix} \mathbf{C}^T\mathbf{C}\widehat{Cov}\begin{pmatrix} \hat{\boldsymbol{\beta}} \\ \hat{\mathbf{u}} - \mathbf{u} \end{pmatrix}.$$

The $\widehat{bias}(x \mid \mathbf{u})$ is defined as in Section 1.

2.3 Sun and Loader (S&L)

Others such as Sun and Loader (1994), have investigated confidence band coverage after adjusting for bias and have shown no improvements. However, the work of Sun and Loader (1994) predates the developing distinction between old- and new-style random effects and potentially used inappropriate data generation mechanisms. Similar to Ruppert et al. (2003), Sun and Loader (1994) account for the bias in the confidence band's variance estimate. They calculate the bias-corrected confidence band as $(\hat{f}(x) - (c + \hat{m}(x))\hat{\sigma}, \hat{f}(x) + (c + \hat{m}(x))\hat{\sigma})$, where $\hat{m}(x)$ is the estimated bias and c is a suitable constant. For inference based on CIs rather than confidence bands, we can consider using the conditional mean squared error to adjust for the bias in the CI's variance estimate. This will also allow us to compare the conditional and unconditional mean squared errors. I will denote these as S&L CIs, which can be calculated as

$$\hat{f}\left(x\right) \pm z_{1-\frac{\alpha}{2}} \sqrt{\hat{E}\left[\{\hat{f}\left(x\right) - f\left(x\right)|\mathbf{u}\}^2\right]},$$

where I calculate the conditional mean squared error as the sum of the conditional variance and the squared conditional bias, as previously given.

3 Simulation Study

To investigate each method's effect on coverage, I conduct a simulation study to generate several datasets using various functions. Each dataset is simulated using a single predictor. The data generation functions imitate artifacts such as fast turns or long valleys seen in real data. To generate such data, I use various "broken-stick" and "bathtub" models, denoted as $f(\cdot)$ and $g(\cdot)$, respectively. In this simulation study, I assume we have $n = 100$ observations and consider three curves from the "broken-stick" model and six curves from the "bathtub" model.

First, I generate M = 1000 sets of error terms, $\epsilon_{n\times 1} \overset{i.i.d}{\sim} N(0, \sigma_\epsilon^2)$ and consider n equally-spaced sample values, $x_i \in [0, 5]$. This setting is appropriate with a regularly observed independent variable, such as weekly medical visits. For more general settings, I simulate observations from $f(\mathbf{x})$ with higher frequency at lower values of $\mathbf{x}$, and lower frequency at higher values of $\mathbf{x}$. This sampling scheme is achieved by generating sample values, $\mathbf{x}$, from a *Gamma*(1, 5/4) (this density specification has a large mass near zero and a long tail across the domain).

Next, I apply the nine functions (three from $f(\cdot)$ and six from $g(\cdot)$) to my predictor $\mathbf{x}_{n\times 1}$ and add each set of error terms, $\epsilon_{n\times 1}$. This results in M datasets for each curve. Note these datasets have repeated measures in the sample values and error terms across the nine curves. The repeated measures allow us to compare coverage at different angles and curvature types across the nine curves. To investigate coverage performance, I chose x_i at and near locations where the data generation function displays high curvature. The curves and points for evaluation (depicted as hashmarks) are displayed for each model in Figure 1 below.

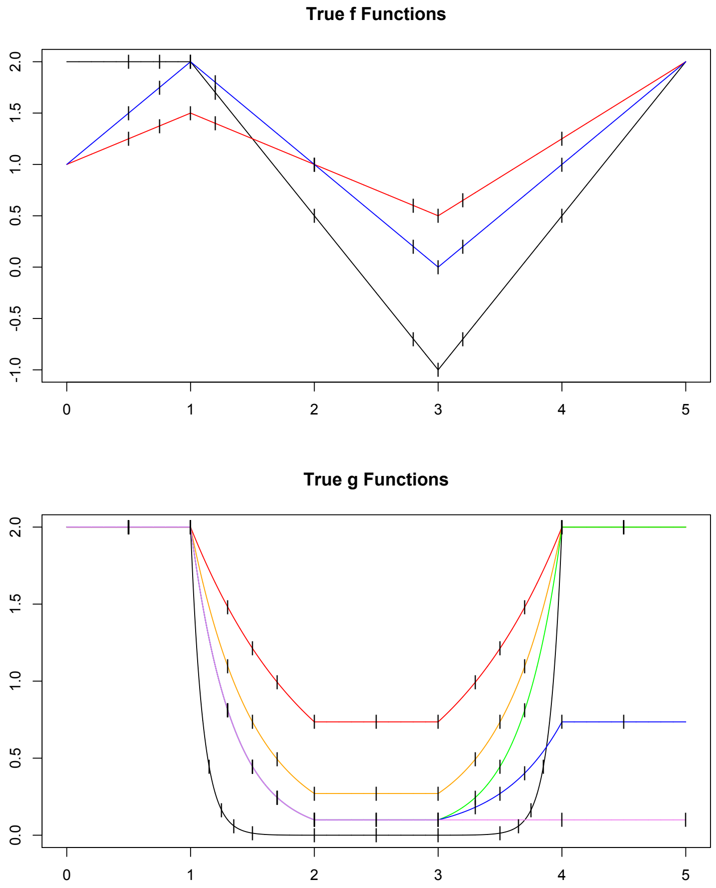


Figure 1: The "broken stick" model (top) and "bathtub" model (bottom) with CI coverage evaluated at the hashmarks.

Coverage probability (CP) is computed as the number of times the $100 \times (1-\alpha)\,\%$ CI contains the true data generation function over the $M = 1000$ simulations. For this paper, I assume $\alpha = 0.05$. In my analysis, I consider $K = \{10, 25, 100\}$ equally-spaced knots taken between 0.2 and 4.8, where 25 is the number of knots recommended by Ruppert et al. (2003), and 100 is a fully saturated model (one knot for each observation). Lastly, to investigate CP under various settings, I consider three levels of error deviation, $\sigma_\epsilon = \{0.05, 0.5, 1\}$. Figure 1 shows that most of the functions are in the range [0, 2]. The black "broken stick" curve ranges over [−1, 2]. Therefore, the values under consideration for σ_ϵ cover an assortment of settings: very little, mild, and heavy variation relative to $\mathbf{y}$.

4 Results

Given the varying factors, my simulation study produced over 54 graphs, not including plots from simulations assuming the sample values come from a Gamma distribution. Thus, I've chosen the interesting graphs to display for different σ_ϵ. The following graphs are the most extreme "broken stick" and "bathtub" models (f_3 in Figure 2, and g_4 in Figures 3 and 4). The remaining curves, assuming $\sigma_\epsilon = 0.5$, are displayed in the Appendix. If you are curious, I can send you the other 50 or so graphs.

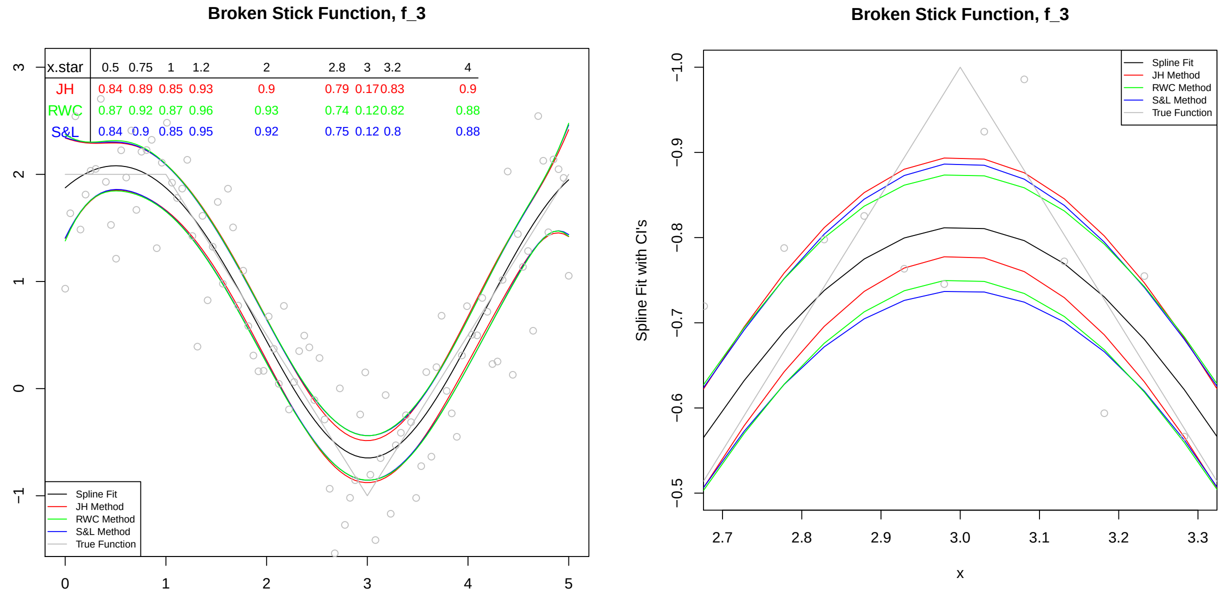
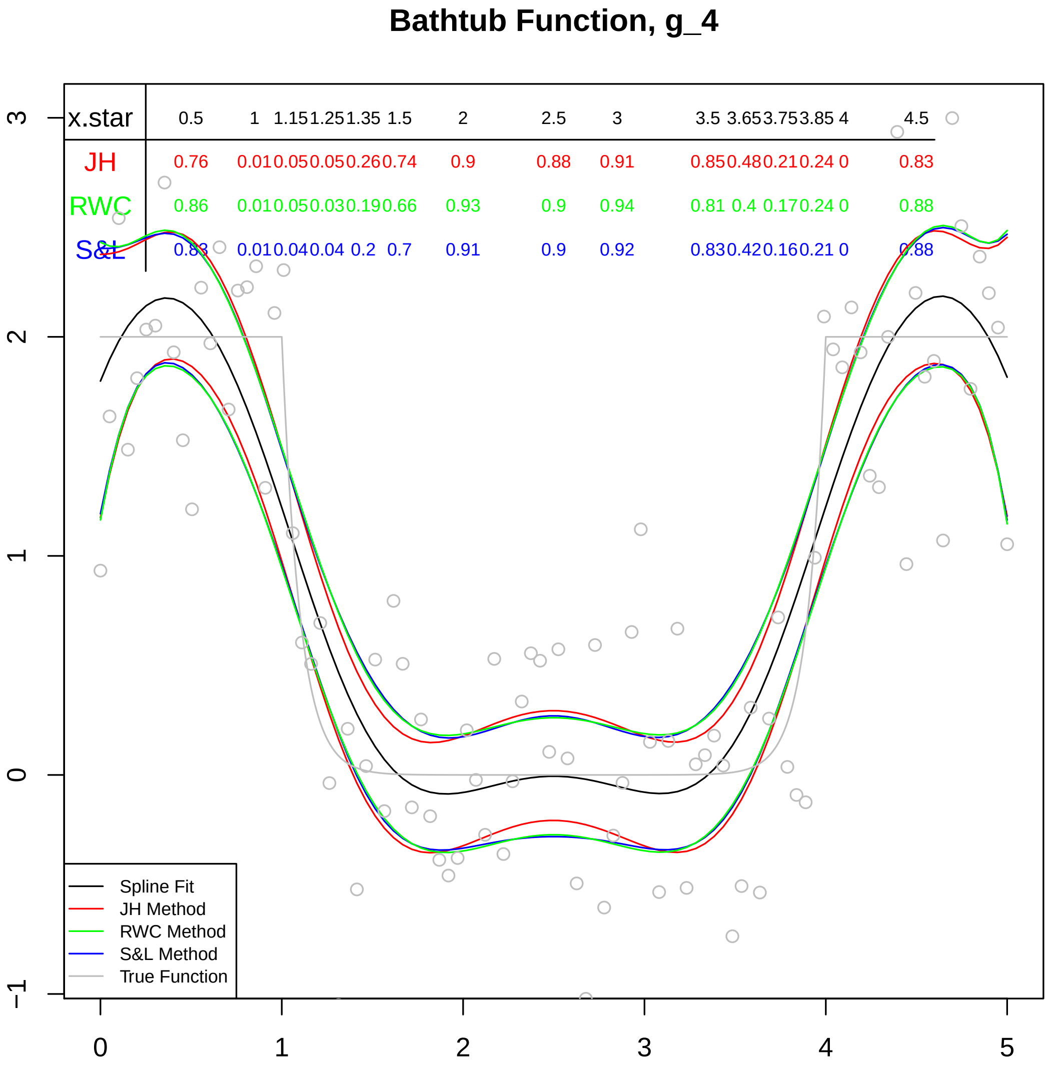


Figure 2: (Left) The last (most severe) "broken stick" model with $\sigma_\epsilon = 0.5$. (Right) The bottom of the same curve with $\sigma_\epsilon = 0.05$.

For the most part, CP comparisons between the three methods are similar for all three sets of knots (recall, $K = \{10, 25, 100\}$ equally-spaced knots), thus all plots presented use the recommended 25 knots in the analysis. Nevertheless, we see dramatic changes in CP between the JH and other two methods at difficult



x.star	0.5	1	1.15	1.25	1.35	1.5	2	2.5	3	3.5	3.65	3.75	3.85	4	4.5
JH	0.76	0.01	0.05	0.05	0.26	0.74	0.9	0.88	0.91	0.85	0.48	0.21	0.24	0	0.83
RWC	0.86	0.01	0.05	0.03	0.19	0.66	0.93	0.9	0.94	0.81	0.4	0.17	0.24	0	0.88
S&L	0.83	0.01	0.04	0.04	0.2	0.7	0.91	0.9	0.92	0.83	0.42	0.16	0.21	0	0.88

Figure 3: The forth (most severe) “bathtub” model with $\sigma_\epsilon = 0.5$.

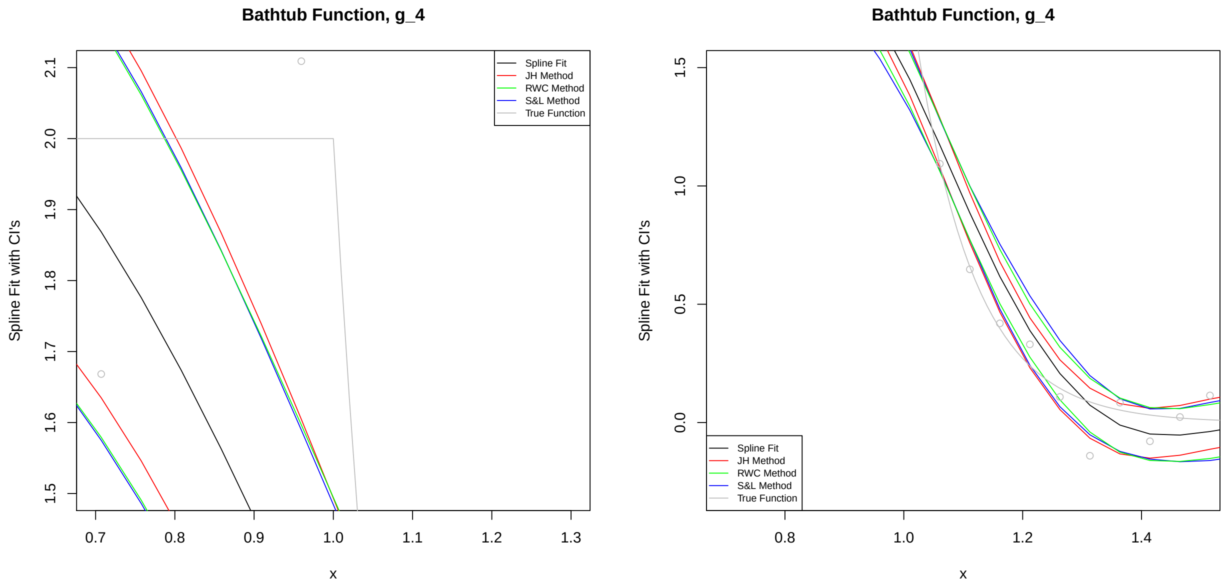


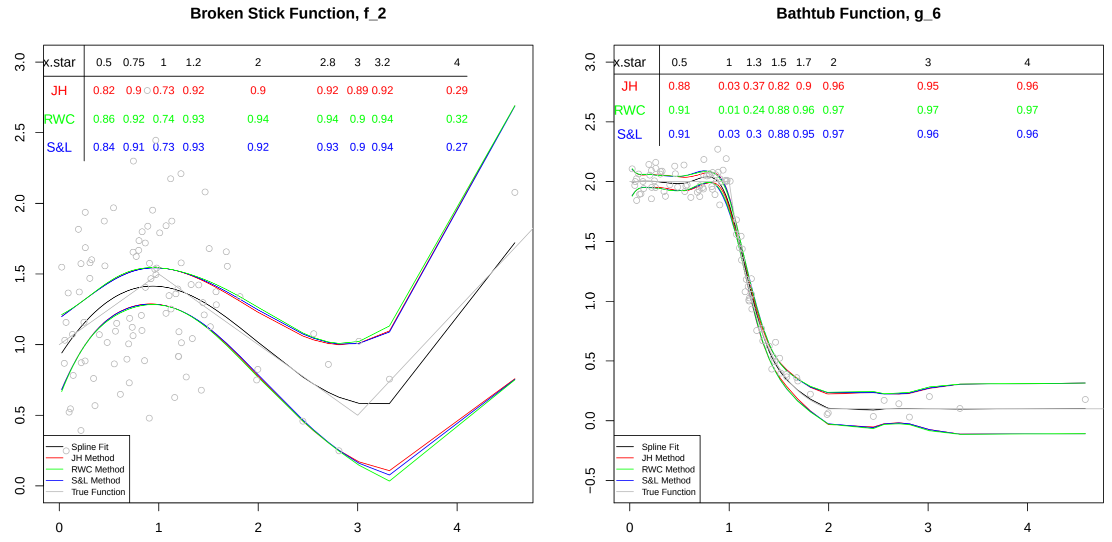
Figure 4: (Left) The top curvature of the forth (most severe) "bathtub" model with $\sigma_\epsilon = 0.5$. (Right) The bottom of the same curve with $\sigma_\epsilon = 0.05$.

points (points of high curvature) as we decrease K for small deviations, $\sigma_\epsilon = 0.05$. For g_4 with $\sigma_\epsilon = 0.05$, at $x = 1.25$ (located on the steep slope after the roughly 90° angle drop) with 10 knots the CPs for JH, RWC, and S&L are 0.38, 0.04, and 0.28, respectively. And as the number of knots increases the CPs drop dramatically. For JH, RWC, and S&L, with 25 knots: 0.06, 0, and 0.03, respectively, and with 100 knots all CPs equal 0; but this is the most extreme case considered of all the simulations. If we look at $x = 1.2$ for a less extreme curve, say f_3 ($x = 1.2$ is after an approximately 135° drop, and still assuming $\sigma_\epsilon = 0.05$), with 10 knots the CPs for JH, RWC, and S&L are 0.32, 0.62, and 0.62, respectively; and with 100 knots: 0.61, 0.78, 0.78, respectively. For a mild standard deviation ($\sigma_\epsilon = 0.5$), the CP is near the nominal value for $x = 1.2$, as displayed in Figure 2 (left).

In other curves, we see a slight increase in CP at difficult points, such as $x = 1, 1.2$, as we decrease the number of knots. I expected this on the linear points of a curve, but not the difficult points with high curvature. However, I think this might be due to the location of the knots. Larger K placed a knot at the difficult points, whereas $K = 10$ placed two knots near but not at the difficult points.

In general, as we increase σ_ϵ, CP increases at difficult points of a curve but decreases at linear points of a curve. However, the comparisons between the three methods' CPs does not change greatly as we vary σ_ϵ. For the most part, I didn't see a drastic difference between the three methods' CPs, except RWC seemed to perform slightly better in most settings. However, I might just be going crazy at this point from looking at so many plots and probabilities. The three bias-correction methods actually appeared so similar I had to try to find ways to separate the lines.

While generating the sample values from a Gamma distribution did not change CP comparisons, it was still interesting since the generated observations with a mild standard deviation resembled so many real datasets I've encountered. Two curves assuming $\mathbf{x} \sim Gamma$ are displayed in Figure 5. The left plot displays f_2 with $\sigma_\epsilon = 0.5$ and the right plot displays g_6 with $\sigma_\epsilon = 0.05$. Again assuming a small error deviation is one of the few times we can see a difference between JH and the other two methods at places were there is a steep slope or large bias (consider $x = 1.3$ in g_6).



x.star	0.5	0.75	1	1.2	2	2.8	3	3.2	4
JH	0.82	0.9	0.73	0.92	0.9	0.92	0.89	0.92	0.29
RWC	0.86	0.92	0.74	0.93	0.94	0.94	0.9	0.94	0.32
S&L	0.84	0.91	0.73	0.93	0.92	0.93	0.9	0.94	0.27

x.star	0.5	1	1.3	1.5	1.7	2	3	4
JH	0.88	0.03	0.37	0.82	0.9	0.96	0.95	0.96
RWC	0.91	0.01	0.24	0.88	0.96	0.97	0.97	0.97
S&L	0.91	0.03	0.3	0.88	0.95	0.97	0.96	0.96

Figure 5: (Left) The least severe "broken stick" function with $\sigma_\epsilon = 0.5$ and $\mathbf{x} \sim Gamma$. (Right) The last "bathtub" function with $\sigma_\epsilon = 0.05$.

5 Remarks

Based on these simulations, it is difficult to extract some meaningful take-home messages when there are no big distinctions between methods in the large number of results. However, it is important to note our ability to estimate the conditional bias is really poor, regardless of K. The largest estimated bias, which is less than 0.1, is at the boundary rather than points that we know have a large bias due to high curvature. This most likely explains why the three methods' CPs are so similar.

Generally, RWC displays consistently higher CP at linear points of a curve and comparable CP at difficult points of a curve. JH only performs slightly better when the bias is larger than the variance, this is apparent when the curve forms a 90° angle or less. S&L performance is similar to RWC, except when the estimated variance is less than the estimated bias, as displayed in Figure 2 (right). Lastly, the unconditional mean squared error is slightly larger than the conditional mean squared error (comparing CI widths for RWC and S&L).

In light of my simulation results, I carried out an additional simulation to examine the impact on CP by adjusting the CIs center and variance estimate by the conditional bias. The goal of such an adjustment is to further accommodate the under-estimated bias, but these results proved to be futile. I also considered user-specified knots, which resulted in an increase in CP. However, there was not much of difference across methods. The above simulations are dependent on the chosen basis and penalty. As an extension, we could consider alternative bases to investigate how the CPs of the three methods are affected.

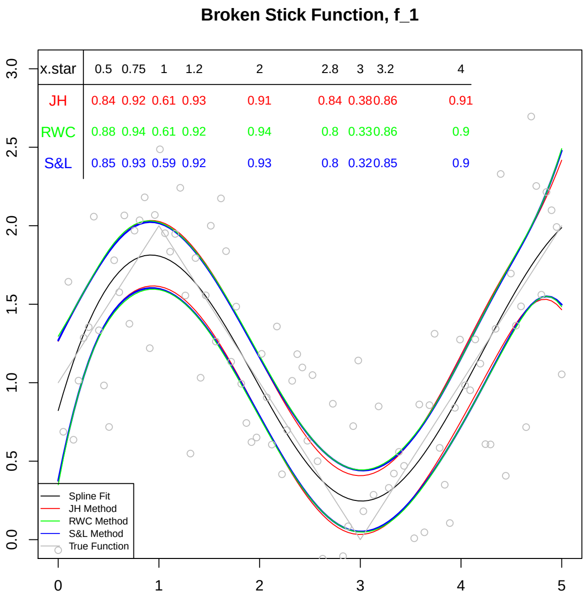
6 Appendix: Figures

References

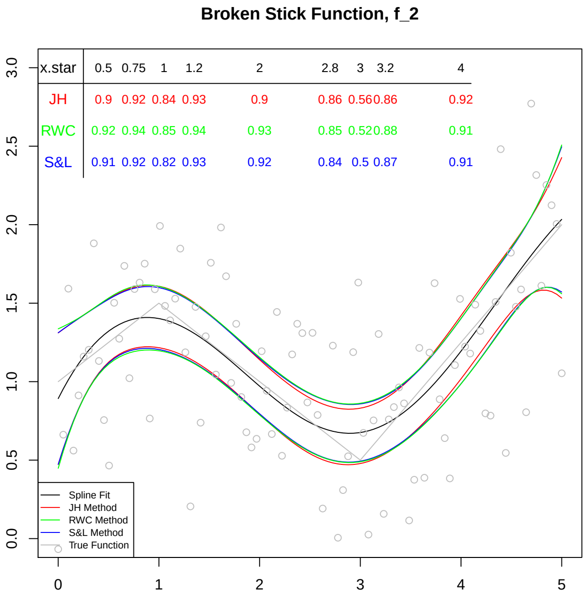
Hodges, J. (2014). *Richly Parameterized Linear Models: Additive, Time Series, and Spatial Models Using Random Effects*. Chapman and Hall.

Ruppert, D., Wand, M. P., and Carroll, R. J. (2003). *Semiparametric regression*. Cambridge University Press.

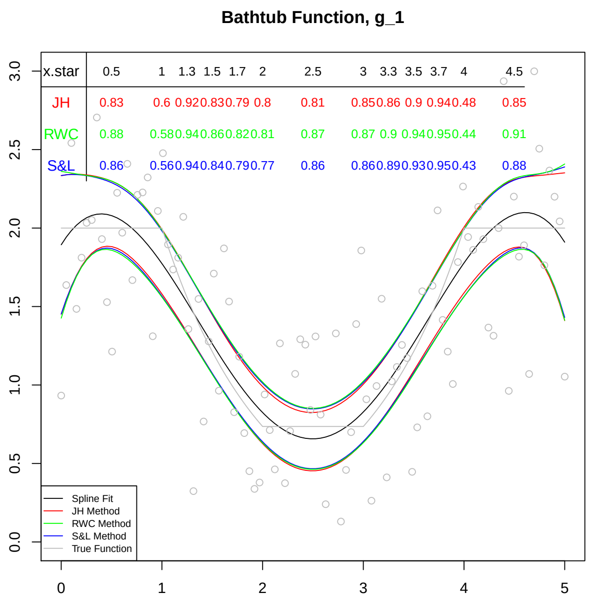
Sun, J. and Loader, C. R. (1994). Simultaneous confidence bands for linear regression and smoothing. *The Annals of Statistics* pages 1328–1345.

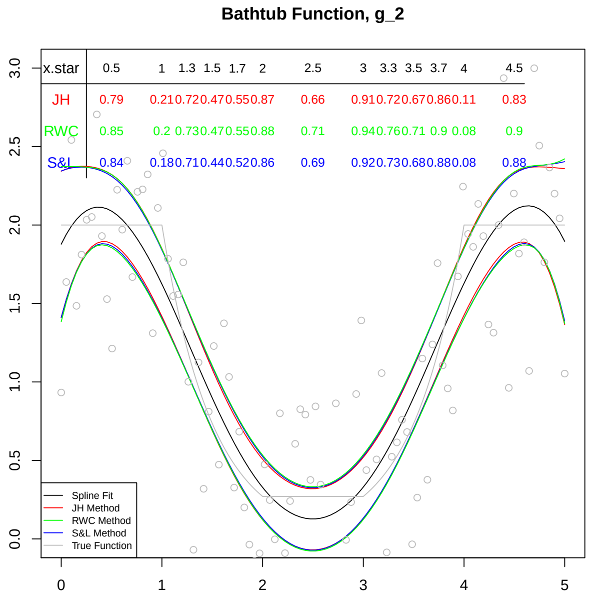


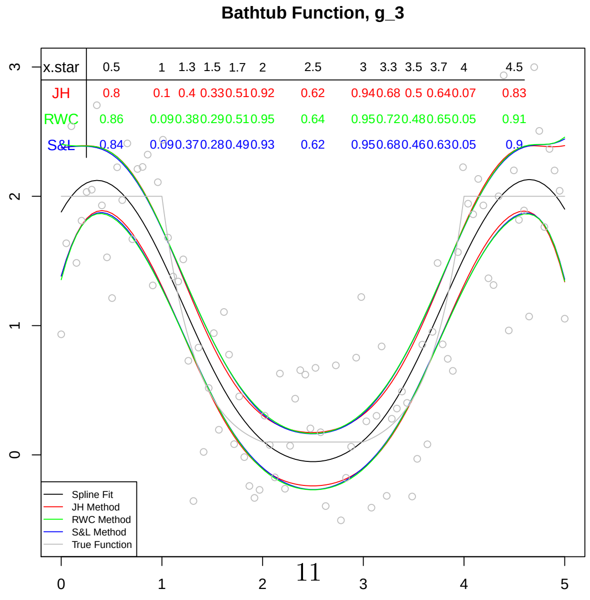
x.star	0.5	0.75	1	1.2	2	2.8	3	3.2	4
JH	0.84	0.92	0.61	0.93	0.91	0.84	0.38	0.86	0.91
RWC	0.88	0.94	0.61	0.92	0.94	0.8	0.33	0.86	0.9
S&L	0.85	0.93	0.59	0.92	0.93	0.8	0.32	0.85	0.9



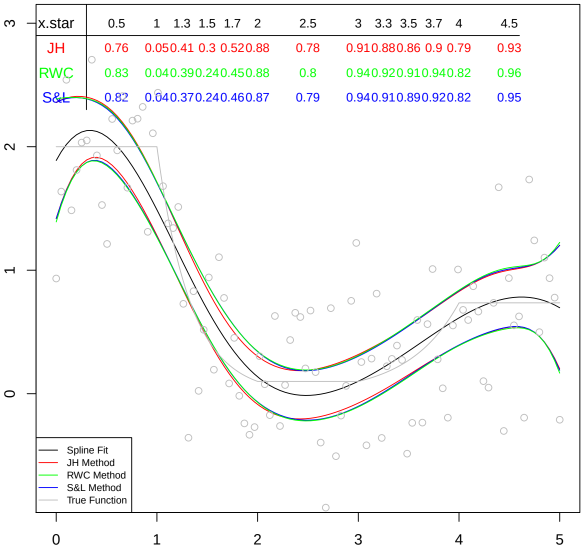
x.star	0.5	0.75	1	1.2	2	2.8	3	3.2	4
JH	0.9	0.92	0.84	0.93	0.9	0.86	0.56	0.86	0.92
RWC	0.92	0.94	0.85	0.94	0.93	0.85	0.52	0.88	0.91
S&L	0.91	0.92	0.82	0.93	0.92	0.84	0.5	0.87	0.91


Bathtub Function, g_1
x.star 0.5 1 1.3 1.5 1.7 2 2.5 3 3.3 3.5 3.7 4 4.5
JH 0.83 0.6 0.920.830.79 0.8 0.81 0.850.86 0.9 0.940.48 0.85
RWC 0.88 0.580.940.860.820.81 0.87 0.87 0.9 0.940.950.44 0.91
S&L 0.86 0.560.940.840.790.77 0.86 0.860.890.930.950.43 0.88
Spline Fit
JH Method
RWC Method
S&L Method
True Function
0.0
0.5
1.0
1.5
2.0
2.5
3.0
0
1
2
3
4
5

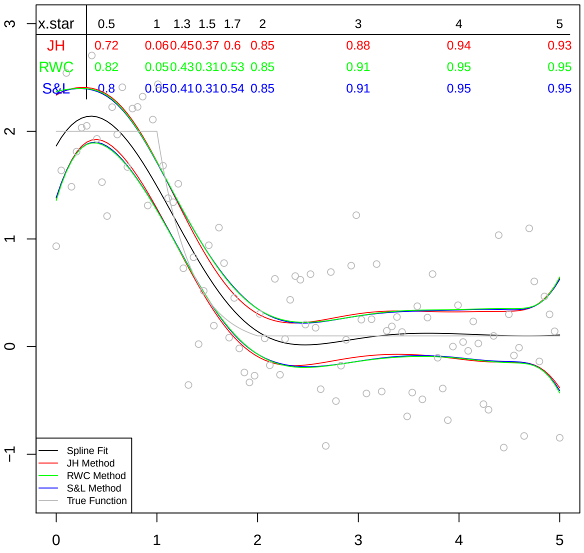

Bathtub Function, g_2
x.star 0.5 1 1.3 1.5 1.7 2 2.5 3 3.3 3.5 3.7 4 4.5
JH 0.79 0.210.720.470.550.87 0.66 0.910.720.670.860.11 0.83
RWC 0.85 0.2 0.730.470.550.88 0.71 0.940.760.71 0.9 0.08 0.9
S&L 0.84 0.180.710.440.520.86 0.69 0.920.730.680.880.08 0.88
Spline Fit
JH Method
RWC Method
S&L Method
True Function
0.0
0.5
1.0
1.5
2.0
2.5
3.0
0
1
2
3
4
5


Bathtub Function, g_3
x.star 0.5 1 1.3 1.5 1.7 2 2.5 3 3.3 3.5 3.7 4 4.5
JH 0.8 0.1 0.4 0.330.510.92 0.62 0.940.68 0.5 0.640.07 0.83
RWC 0.86 0.090.380.290.510.95 0.64 0.950.720.480.650.05 0.91
S&L 0.84 0.090.370.280.490.93 0.62 0.950.680.460.630.05 0.9
Spline Fit
JH Method
RWC Method
S&L Method
True Function
0
1
2
3
0
1
2
3
4
5

Bathtub Function, g_5



Bathtub Function, g_6


x.star
JH
RWC
S&L
Spline Fit
JH Method
RWC Method
S&L Method
True Function