Localised Regression with Topological Data Analysis Ball Mapper

Paweł Dłotko*[1] and Simon Rudkin †[2]

[1]Dioscuri Centre for Topological Data Analysis, Polish Academy of Sciences,Poland
[2]Economics Department, Swansea University, United Kingdom

April 5, 2022

1 Introduction

Statistical modelling will take a considered functional form that maps a set of independent variables onto an outcome and seek to assign coefficients to the elements of the function to generate best fit. In the simplest form this means ordinary least squares (OLS) linear regression with no interactions between independent variables. However in so doing OLS makes a number of uncomfortable assumptions. Firstly, it presumes that each independent variable has a linear association with the outcome variable. Secondly, without interaction the model implicitly assumes that the relationship between a given independent variable and the outcome is not moderated by any other variable. Thirdly, and most implicitly, it is being assumed that one model is the most appropriate for all of the data available to the modeller. In this paper we step back to the data itself and evaluate how topological data analysis (TDA) can be used to understand relationships in data and how the models we apply to data truly fit the underlying data set.

[****] section to fill with bridge

Statistical models of outcomes assume that there is an underlying function $f(X)$ that fits the input variables in X to the outcome y. The model would be written in the form:

$$y = f(X) + \epsilon \tag{1}$$

The key addition is the ϵ term which captures any differences between the observed values and those fit by the model. If the model is to be assumed correct then ϵ will have no correlation with either the input variables or the outcome. Indeed residuals are typically assumed to have an identical independent distribution (iid) which has mean 0 and constant variance σ^2. In fitting the model we obtain an estimate for the parameters in the function $f(X)$ denoted by $\hat{f}(X)$ such that:

$$\hat{y} = \hat{f}(X) \tag{2}$$

Residuals are then $\hat{(e)} = y - \hat{(y)}$. Hats are used for fitted values and e is used as the estimate of ϵ. A perfect model will have residuals that are randomly distributed across the data points that are used in estimation, and which are as small in magnitude as possible. Where there is variation across the space there is a chance to generate useful local models to better explain that heterogeneity. Ours is an important contribution to that space.

Global models which use all of the data points within X have an inherent inability to identify patterns within subsets of X unless specific cutoffs are written into $f(X)$. Where the function is simply a multivariate linear model and there is no reason necessarily to impose breaks and the structure of $f(x)$ assumes a constant parameter on each x_k. However, global models do lend ready interperobility which any improved model must

*Full Address: TBC Email:pdlotko@impan.pl
†**Corresponding Author**. Full Address: Economics Department, School of Management, Swansea University, Bay Campus, Swansea, SA1 8EN, United Kingdom. Tel: +44 (0)1792 606325 Email:s.t.rudkin@swansea.ac.uk

also deliver. At the simplest this means that all $f(x)$ should be the same whatever subset of the data they are fitted to. This paper will assume a set of three data generating processes (DGP) but will proceed with the further assumption that $f(x)$ be an additively separable function in which each independent variable enters linearly.

Departure in this paper stems from the segmentation of the characteristic space using the the BM algorithm. Characteristics here are the K variables x_k, $k \in [1, K]$ variables that make up X. BM selects a point from X at random making this the landmark for a ball. Denote this point as l_1 and the ball of radius ϵ around that point as being ball 1. Any point within ball 1 is considered to be covered. A second landmark, l_2, is selected at random from the remaining uncovered points with ball 2 drawn around this point with the same radius ϵ. Coverage of the data proceeds until all data points are covered by at least one ball. A set of J landmarks, accompanied by the number of points within each ball, is generated. Because of the random nature with which points are selected it is possible to develop marginally different coverage by changing the seed used by the computer. The effect has limited qualitative impact on the interpretation of results, but is important here as there may be marginal changes to the fit quality. We return to the practical implications within the discussion of the localised modelling with BM.

Segmentation of space to improve model fit is not a new concept, researchers have been seeking optimal specifications to promote better fit of the data to hand. [****] Para on splines

Section 2 develops the algorithms through which localised regression is formally implemented. Artificial examples are showcased in Section **??**, with a real world example in Section **??** to show the posited approach in action. Section 6 concludes on the advantages offered and the scope to extend the primary contribution of this paper into wider applications.

2 Localised Linear Regression

BM constructs a coverage of the independent variable space in much the way that the splines approach segments the space based on thresholds of each individual independent variable. Setting out the algorithm employed, and the ways to avoid issues created by insufficient observations in a given locale, this section details the construction of the global (BMLRG) and local (BMLRN) approaches. First the comparator global model and linear splines are set out for reference.

2.1 Global Regression

Baseline to the localised approaches is the single model that makes use of all of the data within the sample. For each dataset the model estimated is simply equation (3), which contains a normally distributed error term ϕ_i and K independent variables. This function is additively separable in the x_k, with each entering the equation with a coefficient β_k.

$$y_i = \alpha + \sum_{k=1}^{K} \beta_k x_k + \phi_i \tag{3}$$

An identically independently distributed error term ϕ_i is introduced with mean 0 and constant variance.

Prediction from global regression generates $\hat{y}$ using the estimated coefficients $\hat{\beta}_k$. Hats here denoting the fitted values obtained in estimating equation (3). Residuals, r_i for the fit are simply $r_i = y_i - \hat{y}_i$. OLS estimation ensures that the sum of the r_i^2 is minimised. However, the minimisation employed in OLS is seen as an averaging across all observations meaning that particular patterns within subsets of the data are missed. This paper contributes to the literature that seeks to improve fit by working on segmenting the characteristic space.

2.2 Linear Splines

Maintaining the assumption that the relationship between input variables and the outcome variable should be linear in a neighbourhood in characteristic space we may adopt the approach of [****] and segment the space using a series of univariate cuts.

2.3 Ball Mapper Localised Regression (BMLR)

Intuition for a BM localised regression stems from the same belief in space segmentation that informs splines. However, rather than segment on a series of cuts BM is instead used to create a cover of all data points that group points within the multidimensional independent variable space. Once data points have been identified a linear model is fitted using OLS on the observations within a given ball.

A second feature of BM is that it captures connectivity in the data through points which sit in the intersection of multiple balls. Here the coefficient used for forecasting is taken as the mean average of the coefficients from each ball to which a data point belongs. Adopting a crude average removes some information from the available set, but is more faithful to the comparison with linear methods than introducing weightings based upon ball centres.

Denoting the betas from any given ball j on variable k as β_k^j then point i which exists in balls $j1$, ..., jJ will have an estimated coefficient:

$$\beta_{jk}^J = \frac{1}{J}(\beta_k^{j1} + ... + \beta_k^{jJ}) \tag{4}$$

Prediction for the BMLR model is thus as per equation (5):

$$\hat{y_i} = \hat{\alpha} + \sum_{k=1}^{K} \hat{\beta_{jk}^J} x_k \tag{5}$$

Fitted values, residuals and hence the quality of the model fit become reliant on the overlaps within the data and the choice of radius for the BM algorithm. As with the splines regression, choices of ϵ are made to achieve the best fit. BM selects points from the uncovered set at random such that there is a degree of variation in the resulting segmentation of the space. Some splits will produce better fit than others. Comparability necessitates taking the average over multiple repetitions, but the researcher targetting best fit may instead go forwards with the best fitting of the repetitions.

Unlike many clustering approaches BM produces groupings of fixed spatial extent, not considering how many points there are within a given ball. Consequently there may be balls which lack sufficient observations to perform OLS regression. When the dataset is large this may be less problematic. For smaller datasets, or in sparse regions of the characteristic space, there is a high probability that a ball will not meet the necessary requirement for model fitting. This paper adopts two approaches for dealing with prediction within these smaller balls.

BMLR may be summarised in the following steps. This is an iterative approach through which the data is covered and models estimated with a view to obtaining the best possible model.

1. Generate BM cover of dataset for given ball radius ϵ
2. Perform regression in all balls where the number of data points exceeds the cut value
3. Replace coefficients for smaller balls according to decision rule
4. Calculate model fit
5. Repeat steps 1 to 4 for a given number of times (r), taking the average coefficient for each data point
6. Loop steps 1 to 5 over different values of ϵ, the optimal model is then that ϵ which produces the best fit

In order to implement these steps the user must select the number of repetitions, r, to make at each epsilon to obtain best fit. An appropriate law of large numbers ensures that r does not need to be too high. This paper will use 20 repetitions in every case. Focus is on the extent to which the results that appear at step 6 outperform the most analogous approach.

2.3.1 Global Substitution (BMLRG)

Where there are balls with insufficient observations to estimate an OLS model it becomes necessary to substitute an alternative to predict outcomes for observations in that part of the space. As a base model the estimated parameters from the global model, equation (3), are used. Were all the balls to be fitted with the global model then we have the benchmark to the model. Within this BMLR global (BMLRG) model the aim is to identify the extent of the gain achieved from the use of the BMLR approach.

2.3.2 Local Substitution (BMLRN)

Assigning coefficients from the global model to those balls which lack sufficient size to have their own models estimated has computational motivation but will not accurately capture the behaviour of the outcome within the region of the space where the ball sits. As an extension the distances between the centre of a ball which is not sufficiently large and the balls for which estimation can be performed can be calculated. In the second stage the regression coefficients from the nearest large ball are used to form predictions within the small ball. Reliance on the nearest balls provides the acronym BMLRN.

2.4 Evaluating Effectiveness

Comparison between methodologies adopts the root mean squared error (RMSE) as the primary tool, reflecting the aim to have best possible fit to the data upon which the modelling is applied. Extension into out-of-sample performance continues to demand better fit so again the RMSE is an appropriate measure. Within the implementation of the BMLR methodologies the calculation of the RMSE informs the selection of the ϵ that is used in comparison with the other approaches. Formally the value is calculated as equation (6).

$$RMSE = \sqrt{1/n \sum_{k=1}^{n} (y_i - \hat{y}_i)^2} \tag{6}$$

3 In Sample Performance

3.1 Univariate Input

Recognising the simplicity with which predictions and actual values may be readily compared on two-dimensional graphs the first set of artificial examples used in this paper take from a single input and single output variable.

Let us consider two data generating processes, one based on a three part linear function and one which is derived from a quadratic input. In each case the inputs, x, will be random draws from a given distribution.

$$y = \begin{cases} 2x & for x \in [0, 0.3] \\ 0.6 & for x \in (0.3, 0.7) \\ 1.4 - x & for x \in [0.7, 1] \end{cases} \tag{7}$$

And to capture a process which has actually a curved form let us use:

$$y = (0.5 - x)^2 \tag{8}$$

To both DGP a noise term ν is added which has mean 0 and variance σ^2. The number of observations, n, and the noise variance will be varied to demonstrate the effect on the accuracy with which the BMLR models fit data.

In each case the data does not suit a global linear model making it straightforward for any localised approach to deliver lower RMSE. Hence although the imposed function remains $y = \alpha + \beta x$, the target here is to obtain the closest estimates to the true beta for every observation in the data set. Having different betas at different x values gives the local linear approximations to the outcome. As well as giving the best fit a good model of this data should deliver the closest possible values to the impact of x on y for a given x.

Table 1: Univariate Model Fit: Uniform Distribution of X

Sigma	n	OLS	Cut = 25			Cut = 50		
			Splines	BMLRG	BMLRN	Splines	BMLRG	BMLRN
Panel (a): Three Part Function								
0.02	200	0.158	0.038	0.024	0.020	0.027	0.037	0.029
	500	0.150	0.022	0.020	0.019	0.026	0.021	0.020
	1000	0.153	0.028	0.020	0.020	0.024	0.02	0.020
	2000	0.155	0.025	0.019	0.019	0.025	0.02	0.020
0.20	200	0.243	0.200	0.196	0.195	0.198	0.198	0.197
	500	0.257	0.199	0.195	0.194	0.2	0.197	0.196
	1000	0.253	0.204	0.2	0.199	0.205	0.202	0.201
	2000	0.249	0.199	0.195	0.192	0.199	0.197	0.195
0.50	200	0.508	0.496	0.49	0.488	0.495	0.492	0.492
	500	0.53	0.496	0.489	0.485	0.496	0.492	0.490
	1000	0.531	0.509	0.501	0.497	0.509	0.505	0.502
	2000	0.518	0.496	0.487	0.481	0.496	0.491	0.488
Panel (b): Quadratic Function								

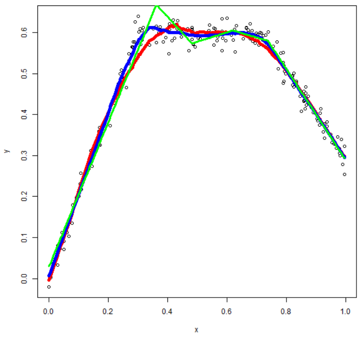
3.1.1 Uniform Distribution

Initially let us assume that x is drawn from a uniform distribution on $[0, 1]$, this ensures that ex-ante no region of the data is necessarily any more sparse than any other. Expositing the role of the parameters of the BMLR algorithms demonstration is given that BMLR can provide a more accurate fit than the splines regression[1].

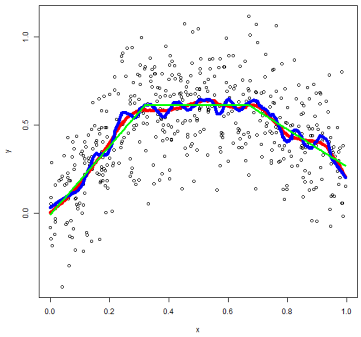
Consider equation (12) estimated for 1000 data points drawn randomly from the uniform [0,1] distribution. For the present examples 20 repetitions of the BMLR algorithms are used, higher degrees of accuracy can be obtained by using higher numbers. Panel (a) of Table 1 reports the RMSE for varying sample sizes and variances on the noise distribution. Whilst the OLS model is not affected by the cut size, the estimates of the splines, BMLRG and BMLRN must be reported for both cut sizes. A marginal improvement in fit is seen from giving splines the higher minimum knot size, whilst both BMLR have slightly worsened performance from the extra restriction on minimum ball size. Throughout BMLRN offers the smallest RMSE. Evaluation of the driving force behind these differentials is seen in Figure 1. Within these figures points generated by the DGP are shown as hollow, with the fitted splines line in green, the BMLRG line in red and the BMLRN in blue.

Plots selected in Figure 1 expand upon the issues raised by fitting localised models to noisy data. Panels (a) to (c) require that the minimum size of a segment in the splines regression is 25 observations, while (d) to (f) set that minimum at 50. Likewise a regression model may only be estimated for a ball that contains 25 observations in (a) to (c), that figure being 50 in (d) to (e). As the noise level increases the BMLR methods, especially BMLRN, start to pick up the noise and create a better fitting, but wavy, line. When the cut is increased to 50 a smoother plot is found. With low observation numbers the superior fit of the BMLRN model is clear, the splines model taking an extra kink around $x = 0.3$. When the cut is increased to 50 the BMLRN line is curved and so the splines fits better. With more observations BMLRN is able to pick the slopes at either end amongst the noise and quicker identifies the flat nature of the central section when $\sigma = 0.5$.

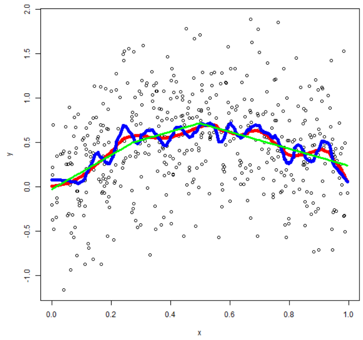
Figure 1: Univariate BMLR Fit Comparisons



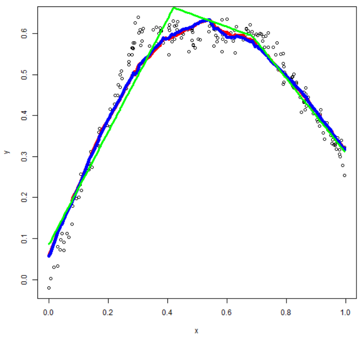
(a) Cut 25 $\sigma = 0.02$, $n = 200$



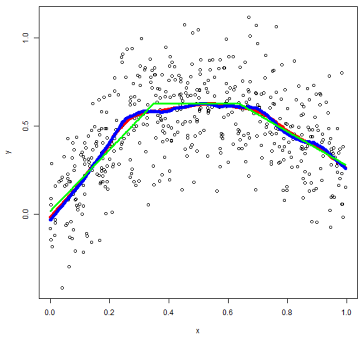
(b) Cut 25 $\sigma = 0.20$ $n = 500$



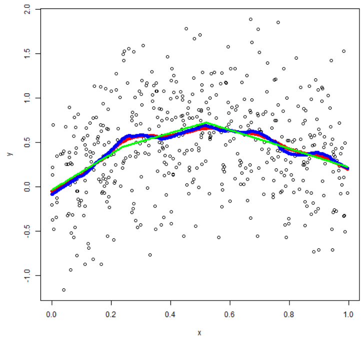
(c) Cut 25 $\sigma = 0.50$, $n = 500$



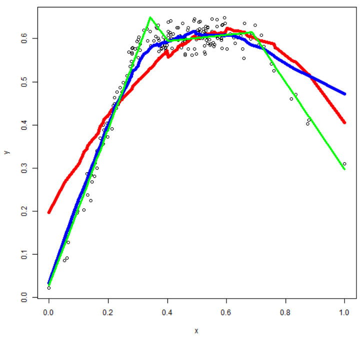
(a) Cut 50 $\sigma = 0.02$, $n = 200$



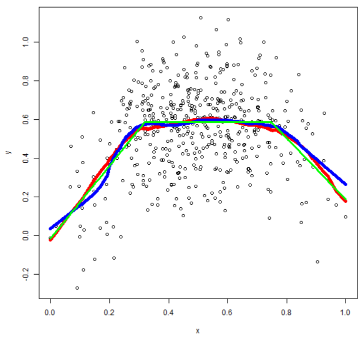
(b) Cut 50 $\sigma = 0.20$ $n = 500$



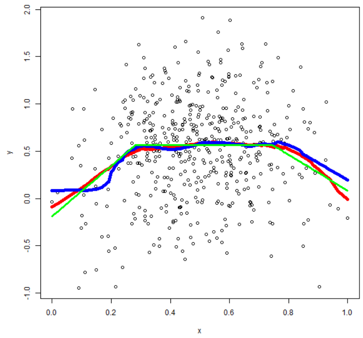
(c) Cut 50 $\sigma = 0.50$, $n = 500$



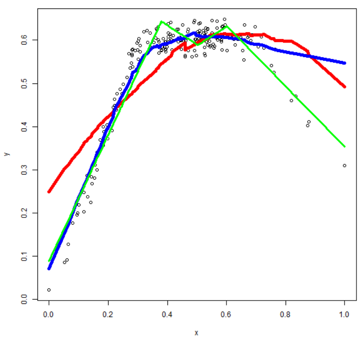
(a) Cut 25 $\sigma = 0.02$, $n = 200$



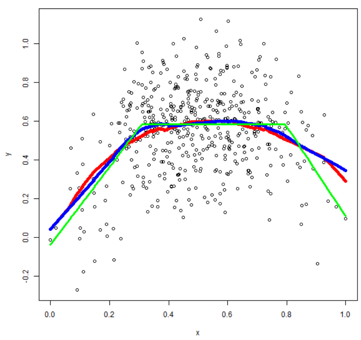
(b) Cut 25 $\sigma = 0.20$ $n = 500$



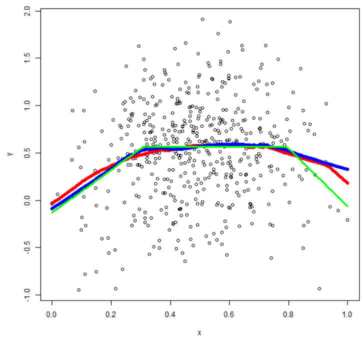
(c) Cut 25 $\sigma = 0.50$, $n = 500$



(a) Cut 50 $\sigma = 0.02$, $n = 200$



(b) Cut 50 $\sigma = 0.20$ $n = 500$



(c) Cut 50 $\sigma = 0.50$, $n = 500$

Table 2: Univariate Model Fit: Normal Distribution of X

Sigma	n	OLS	Cut = 25			Cut = 50		
			Splines	BMLRG	BMLRN	Splines	BMLRG	BMLRN
0.02	200	0.081	0.03	0.042	0.033	0.024	0.05	0.042
	500	0.107	0.022	0.03	0.021	0.021	0.039	0.025
	1000	0.082	0.021	0.027	0.020	0.022	0.034	0.021
	2000	0.047	0.021	0.025	0.020	0.021	0.027	0.021
0.20	200	0.213	0.197	0.199	0.197	0.197	0.201	0.201
	500	0.233	0.208	0.207	0.205	0.206	0.209	0.207
	1000	0.214	0.199	0.199	0.197	0.198	0.2	0.198
	2000	0.208	0.203	0.204	0.203	0.203	0.204	0.203
0.50	200	0.525	0.515	0.514	0.510	0.516	0.515	0.513
	500	0.529	0.516	0.516	0.513	0.516	0.517	0.516
	1000	0.503	0.496	0.496	0.493	0.496	0.496	0.494
	2000	0.51	0.507	0.507	0.506	0.507	0.508	0.506

3.1.2 Normal Distribution

3.2 Multivariate Examples

3.2.1 Basic Functions

Departure for linear approximation models is made from a desire to fit non-linear functions with more readily understood straight line segments. Therefore when faced with the linear regression model there would be no reason to include approximations. As a first case we seek to understand how the approximation methods perform when the true data generating process (DGP) is a single global linear function. For simplicity we assume that the DGP is given as equation (9)

$$y = x_1 + x_2 + x_3 + x_4 + x_5 + \psi \tag{9}$$

Where x_1 to x_5 are random draws from a uniform distribution on [0,1] and ψ is a noise term which has expected value 0 and constant variance σ^2. We may readily extend to the quadratic function. Here the origin produces a y value of zero and the outcome increases in all directions moving away from 0. With two axes the contour plot shows lines getting closer together. Note that the use of a uniform [0,1] for x means that only a part of the overall convex shape is plotted on our domain. Using a quadratic like equation (10) has the advantage of not being linear but is a good example of a case where linear approximation would be straightforward. The DGP employed is:

$$y = x_1^2 + x_2^2 + x_3^2 + x_4^2 + x_5^2 + \psi \tag{10}$$

Where again x_1 to x_5 are random draws from a uniform distribution on [0,1] and ψ is a noise term which has expected value 0 and constant variance σ^2.

Both (9) and (10) are written with coefficients of 1 on each term and are specified to be additively separable in x_k, $k \in [1, 5]$. Relaxing to allow alternative coefficients will cause the error associated with a particular x_k to have more, or less, importance to the overall results

We fit the following function on 5 variables. All variables are drawn from a random uniform distribution, whilst a noise term is added which is normally distributed with mean 0 and standard deviation σ. We fit

$$y = 10sin(\pi x_1 x_2) + 20(x_3 - 0.5) + 10x_4 + 5x_5 + \psi \tag{11}$$

[1]Note that we would expect both methods to produce a better fit than the simple OLS regression. This follows from the non-linearities in the respective DGPs.

3.3 Uniform Distribution Results

4 Cross Validation and Out of Sample Performance

Methodologies which offer improved in-sample fit face regular criticism for overfitting data and not offering sufficient inference out of sample. Evaluating out-of-sample performance a cross validation approach is applied to the previous analyses, in every case 10% of observations being held out to test the performance of the model.

5 Empirical Example

This will need a simple dataset where splines have been used.

6 Conclusions

This to consider as extensions for further work...

- Non-linearity
- Further smoothing from BM
- Dimensionality
- Model selection

Focus on linearity in this paper to match that in other work.... STRESS

A Univariate Examples Full Results

In this Appendix we offer fuller results and discussion to support the short tables and example graphs presented in the main paper. First we work through the four examined cases, comprising permutations two functions with both uniform and normally distributed points. We then discuss the role of the number of repetitions in the BM algorithm and the choice of minimum ball size.

A.1 Three Part Function: Uniform Distribution

A.2 Choice of Repetitions

In the main text we set the number of repetitions of the BMLR algorithms, r, to 20 appealing to a law of large numbers on the quality of fit. As a support to this decision within this subsection we present an example of the three part linear function used in the main text.

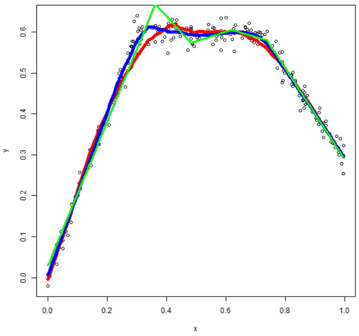
$$y = \begin{cases} 2x & for x \in [0, 0.3] \\ 0.6 & for x \in (0.3, 0.7) \\ 1.4 - x & for x \in [0.7, 1] \end{cases} \tag{12}$$

To y we add a noise that is normally distributed of mean 0 and $\sigma = 0.1$.

Table A.1: Three Part Function Fit

Sigma	n	OLS	Splines	BMLRG	BMLRN	Splines	BMLRG	BMLRN
0.02	200	0.158	0.022	0.02	0.0198	0.031	0.028	0.028
	300	0.146	0.031	0.02	0.020	0.041	0.023	0.020
	400	0.147	0.026	0.02	0.020	0.022	0.022	0.020
	500	0.15	0.022	0.020	0.019	0.026	0.021	0.020
	600	0.151	0.026	0.020	0.020	0.03	0.021	0.020
	700	0.152	0.030	0.020	0.020	0.023	0.021	0.020
	800	0.152	0.026	0.020	0.020	0.024	0.021	0.020
	900	0.154	0.029	0.020	0.020	0.027	0.02	0.020
	1000	0.153	0.028	0.020	0.020	0.024	0.02	0.020
	1100	0.155	0.028	0.020	0.020	0.025	0.02	0.020
	1200	0.155	0.026	0.020	0.020	0.029	0.02	0.020
	1300	0.155	0.026	0.020	0.019	0.024	0.02	0.020
	1400	0.155	0.027	0.019	0.019	0.025	0.019	0.019
	1500	0.155	0.025	0.019	0.019	0.024	0.019	0.019
	1600	0.155	0.026	0.019	0.019	0.025	0.019	0.019
	1700	0.154	0.026	0.019	0.019	0.027	0.02	0.019
	1800	0.154	0.026	0.019	0.019	0.027	0.02	0.020
	1900	0.155	0.024	0.019	0.019	0.027	0.02	0.020
	2000	0.155	0.025	0.019	0.019	0.025	0.02	0.020
0.20	200	0.243	0.200	0.196	0.195	0.198	0.198	0.197
	300	0.241	0.198	0.196	0.195	0.2	0.197	0.197
	400	0.254	0.202	0.2	0.198	0.202	0.201	0.200
	500	0.257	0.199	0.195	0.194	0.2	0.197	0.196
	600	0.258	0.206	0.199	0.197	0.205	0.201	0.201
	700	0.259	0.206	0.201	0.200	0.207	0.203	0.203
	800	0.254	0.205	0.201	0.199	0.204	0.203	0.202
	900	0.255	0.203	0.199	0.197	0.204	0.201	0.200
	1000	0.253	0.204	0.2	0.199	0.205	0.202	0.201
	1100	0.257	0.205	0.198	0.197	0.204	0.201	0.201
	1200	0.254	0.202	0.197	0.196	0.202	0.199	0.198
	1300	0.253	0.2	0.196	0.194	0.2	0.198	0.197
	1400	0.249	0.197	0.192	0.191	0.196	0.194	0.193
	1500	0.247	0.194	0.188	0.188	0.194	0.191	0.190
	1600	0.249	0.196	0.19	0.189	0.196	0.192	0.191
	1700	0.252	0.199	0.193	0.191	0.199	0.196	0.195
	1800	0.251	0.2	0.195	0.192	0.2	0.197	0.196
	1900	0.25	0.199	0.194	0.192	0.199	0.197	0.196
	2000	0.249	0.199	0.195	0.192	0.199	0.197	0.195
0.5	200	0.508	0.496	0.49	0.488	0.495	0.492	0.492
	300	0.511	0.494	0.491	0.489	0.494	0.493	0.492
	400	0.532	0.507	0.5	0.494	0.504	0.502	0.500
	500	0.53	0.496	0.489	0.485	0.496	0.492	0.490
	600	0.537	0.507	0.498	0.493	0.507	0.503	0.501
	700	0.54	0.512	0.503	0.500	0.513	0.507	0.507
	800	0.533	0.51	0.504	0.498	0.511	0.507	0.504
	900	0.53	0.505	0.498	0.492	0.504	0.502	0.499
	1000	0.531	0.509	0.501	0.497	0.509	0.505	0.502
	1100	0.534	0.509	0.496	0.494	0.507	0.503	0.501
	1200	0.528	0.504	0.491	0.489	0.504	0.498	0.496
	1300	0.523	0.499	0.489	0.486	0.499	0.494	0.491
	1400	0.514	0.49	0.48	0.477	0.49	0.485	0.483
	1500	0.507	0.483	0.47	0.47	0.484	0.477	0.474
	1600	0.511	0.487	0.475	0.472	0.487	0.481	0.479
	1700	0.52	0.496	0.483	0.476	0.496	0.49	0.487
	1800	0.521	0.5	0.487	0.480	0.5	0.493	0.490
	1900	0.519	0.497	0.485	0.479	0.496	0.492	0.489
	2000	0.518	0.496	0.487	0.481	0.496	0.491	0.488

Figure A.1: Univariate BMLR Fit Comparisons: Three Part Function Uniform Distribution

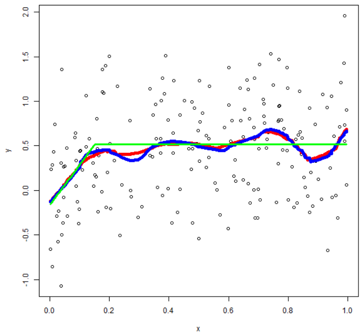


=

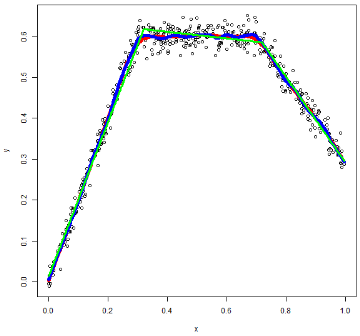
(a) Cut 25 $\sigma = 0.02$, $n = 200$

`varyneg1251200s2.png`

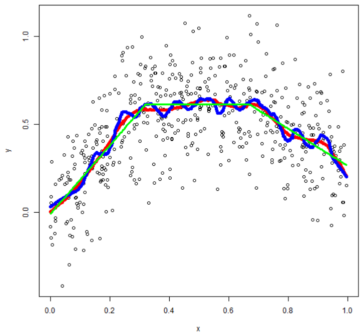
(b) Cut 25 $\sigma = 0.20$ $n = 200$



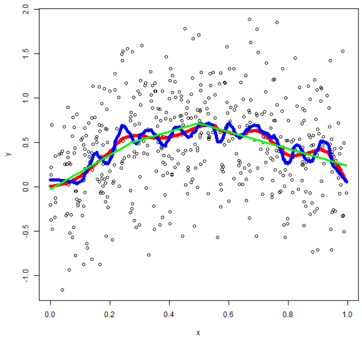
(c) Cut 25 $\sigma = 0.50$, $n = 200$



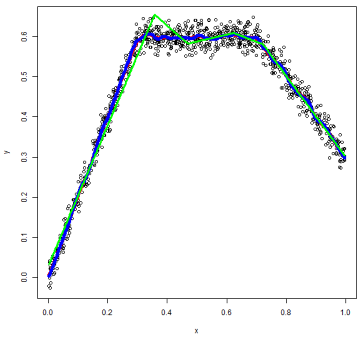
(d) Cut 25 $\sigma = 0.02$, $n = 500$



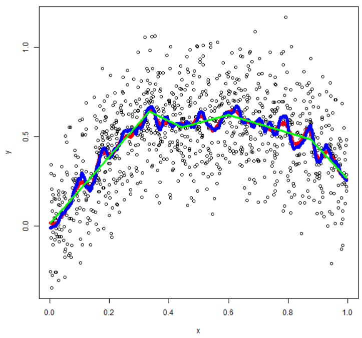
(e) Cut 25 $\sigma = 0.20$ $n = 500$



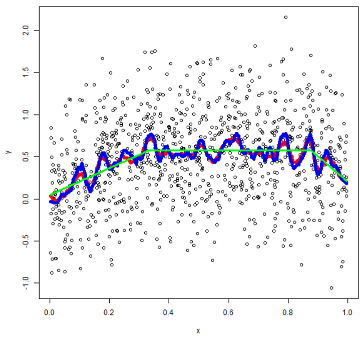
(f) Cut 25 $\sigma = 0.50$, $n = 500$



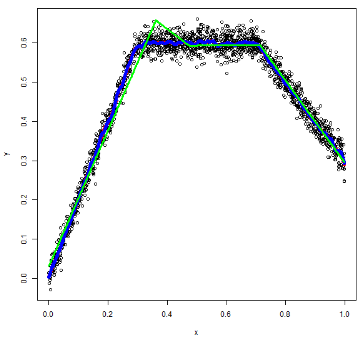
(g) Cut 25 $\sigma = 0.02$, $n = 1000$



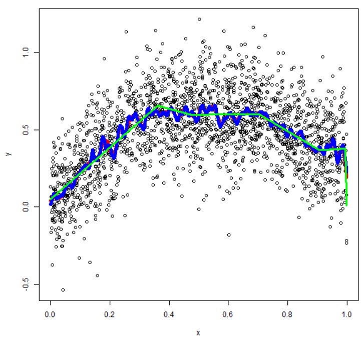
(h) Cut 25 $\sigma = 0.20$ $n = 1000$



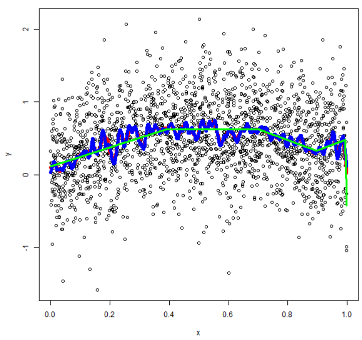
(i) Cut 25 $\sigma = 0.50$, $n = 1000$



(j) Cut 25 $\sigma = 0.02$, $n = 2000$

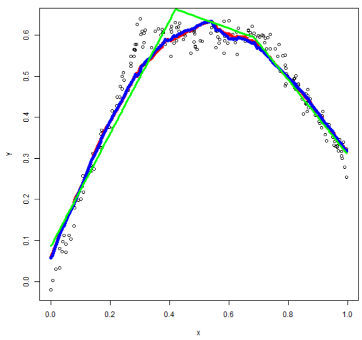


(k) Cut 25 $\sigma = 0.20$ $n = 2000$

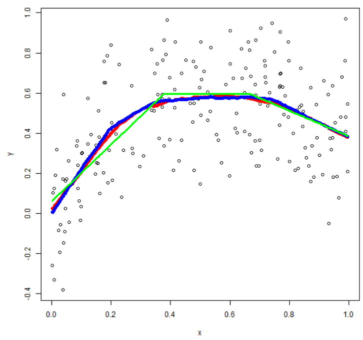


(l) Cut 25 $\sigma = 0.50$, $n = 2000$

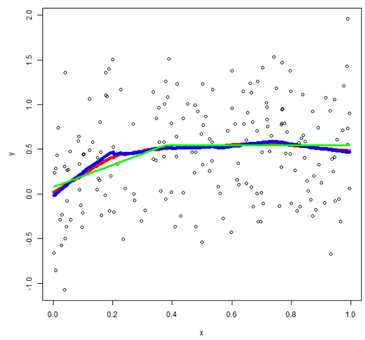
Figure A.2: Univariate BMLR Fit Comparisons: Three Part Function Uniform Distribution



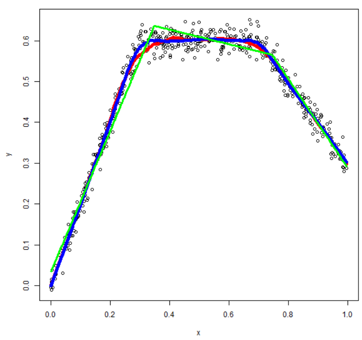
(a) Cut=50 $\sigma = 0.02$, $n = 200$



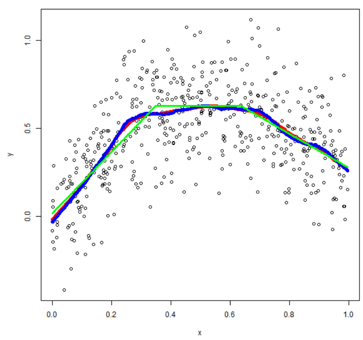
(b) Cut=50 $\sigma = 0.20$ $n = 200$



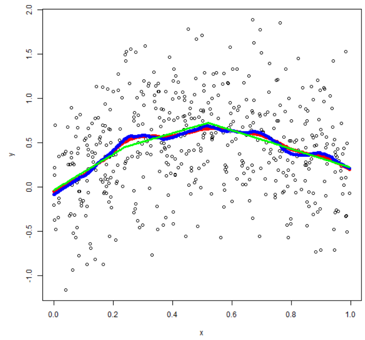
(c) Cut=50 $\sigma = 0.50$, $n = 200$



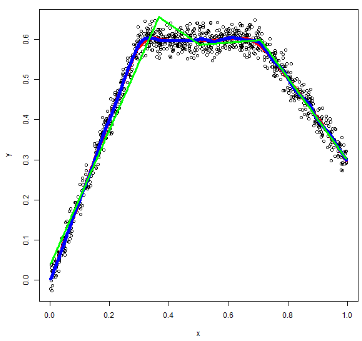
(d) Cut=50 $\sigma = 0.02$, $n = 500$



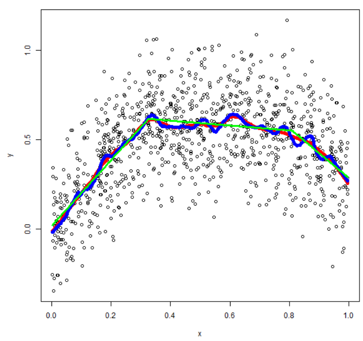
(e) Cut=50 $\sigma = 0.20$ $n = 500$



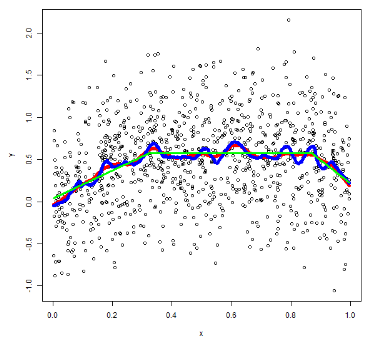
(f) Cut=50 $\sigma = 0.50$, $n = 500$



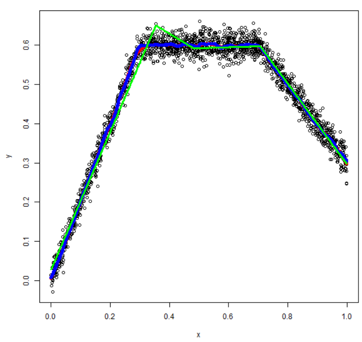
(g) Cut=50 $\sigma = 0.02$, $n = 1000$



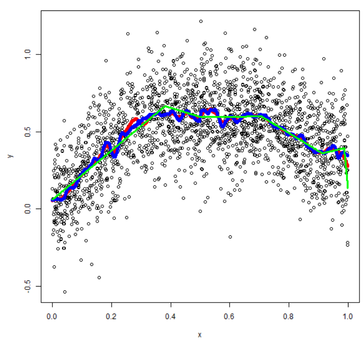
(h) Cut=50 $\sigma = 0.20$ $n = 1000$



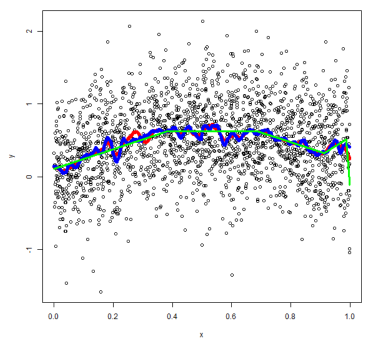
(i) Cut=50 $\sigma = 0.50$, $n = 1000$



(j) Cut=50 $\sigma = 0.02$, $n = 2000$



(k) Cut=50 $\sigma = 0.20$ $n = 2000$

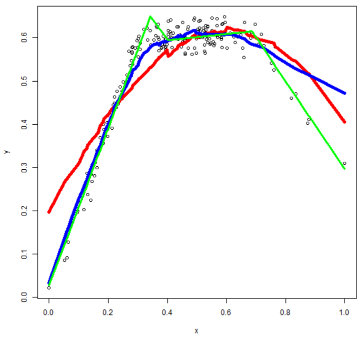


(l) Cut=50 $\sigma = 0.50$, $n = 2000$

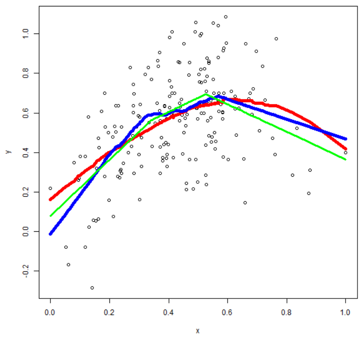
Table A.2: Three Part Function Fit: Normal Distribution

Sigma	n	OLS	Splines	BMLRG	BMLRN	Splines	BMLRG	BMLRN
0.02	200	0.081	0.03	0.042	0.033	0.024	0.05	0.042
	300	0.083	0.024	0.035	0.025	0.028	0.043	0.034
	400	0.101	0.022	0.032	0.02	0.021	0.04	0.026
	500	0.107	0.022	0.03	0.021	0.021	0.039	0.025
	600	0.092	0.023	0.026	0.021	0.024	0.031	0.023
	700	0.093	0.021	0.024	0.020	0.023	0.027	0.021
	800	0.097	0.021	0.023	0.020	0.025	0.025	0.020
	900	0.086	0.021	0.029	0.020	0.022	0.036	0.022
	1000	0.082	0.021	0.027	0.020	0.022	0.034	0.021
	1100	0.08	0.02	0.028	0.020	0.021	0.033	0.020
	1200	0.08	0.02	0.026	0.019	0.020	0.03	0.020
	1300	0.047	0.02	0.025	0.020	0.020	0.027	0.020
	1400	0.048	0.02	0.025	0.020	0.020	0.029	0.021
	1500	0.048	0.02	0.025	0.019	0.019	0.028	0.020
	1600	0.047	0.02	0.025	0.020	0.020	0.027	0.020
	1700	0.047	0.02	0.026	0.020	0.021	0.027	0.021
	1800	0.048	0.021	0.025	0.020	0.021	0.028	0.021
	1900	0.047	0.021	0.025	0.020	0.021	0.027	0.021
	2000	0.047	0.021	0.025	0.020	0.021	0.027	0.021
0.20	200	0.213	0.197	0.199	0.197	0.197	0.201	0.201
	300	0.214	0.2	0.202	0.197	0.201	0.204	0.202
	400	0.218	0.197	0.199	0.194	0.198	0.201	0.194
	500	0.233	0.208	0.207	0.205	0.206	0.209	0.207
	600	0.229	0.207	0.208	0.206	0.207	0.208	0.207
	700	0.227	0.203	0.203	0.201	0.204	0.204	0.202
	800	0.224	0.202	0.202	0.200	0.202	0.202	0.201
	900	0.219	0.204	0.205	0.203	0.204	0.206	0.204
	1000	0.214	0.199	0.199	0.197	0.198	0.2	0.198
	1100	0.211	0.196	0.197	0.194	0.196	0.198	0.195
	1200	0.211	0.194	0.194	0.192	0.193	0.195	0.193
	1300	0.2	0.196	0.196	0.195	0.196	0.196	0.196
	1400	0.204	0.197	0.198	0.197	0.197	0.198	0.197
	1500	0.198	0.194	0.195	0.193	0.194	0.195	0.193
	1600	0.202	0.198	0.198	0.198	0.198	0.198	0.198
	1700	0.203	0.199	0.200	0.198	0.199	0.2	0.199
	1800	0.206	0.201	0.202	0.201	0.201	0.202	0.201
	1900	0.206	0.202	0.202	0.201	0.202	0.202	0.201
	2000	0.208	0.203	0.204	0.203	0.203	0.204	0.203
0.50	200	0.525	0.515	0.514	0.510	0.516	0.515	0.513
	300	0.505	0.499	0.499	0.491	0.499	0.5	0.498
	400	0.501	0.488	0.493	0.485	0.489	0.495	0.485
	500	0.529	0.516	0.516	0.513	0.516	0.517	0.516
	600	0.529	0.518	0.517	0.515	0.518	0.518	0.516
	700	0.523	0.507	0.508	0.504	0.509	0.509	0.502
	800	0.514	0.504	0.503	0.499	0.504	0.504	0.503
	900	0.516	0.511	0.51	0.507	0.511	0.511	0.509
	1000	0.503	0.496	0.496	0.493	0.496	0.496	0.494
	1100	0.497	0.491	0.49	0.486	0.491	0.491	0.486
	1200	0.493	0.484	0.484	0.481	0.484	0.484	0.482
	1300	0.491	0.49	0.489	0.488	0.49	0.489	0.489
	1400	0.497	0.493	0.493	0.492	0.493	0.493	0.492
	1500	0.487	0.485	0.485	0.483	0.486	0.486	0.483
	1600	0.497	0.495	0.494	0.494	0.495	0.495	0.495
	1700	0.5	0.5	0.499	0.495	0.500	0.499	0.497
	1800	0.506	0.503	0.504	0.503	0.503	0.504	0.503
	1900	0.506	0.504	0.504	0.503	0.504	0.504	0.503
	2000	0.51	0.507	0.507	0.506	0.507	0.508	0.506

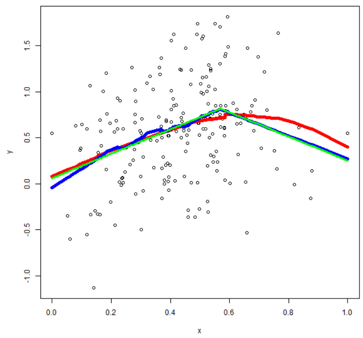
Figure A.3: Univariate BMLR Fit Comparisons: 3 Part Function with Normal Distribution



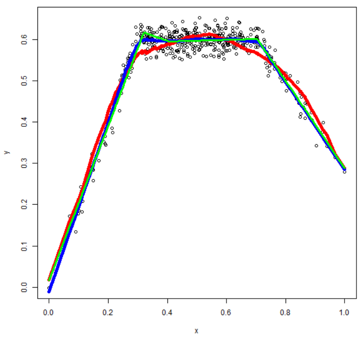
(a) Cut=25 $\sigma = 0.02$, $n = 200$



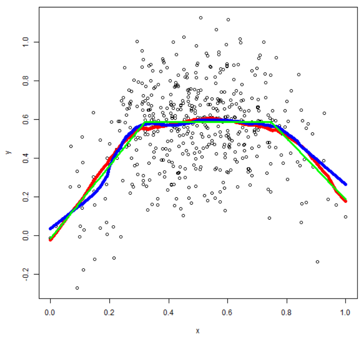
(b) Cut=25 $\sigma = 0.20$ $n = 200$



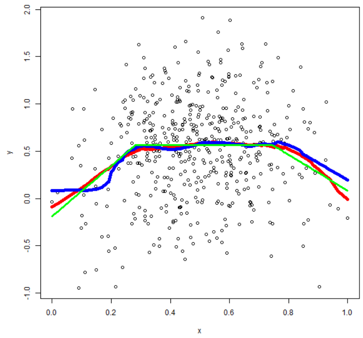
(c) Cut=25 $\sigma = 0.50$, $n = 200$



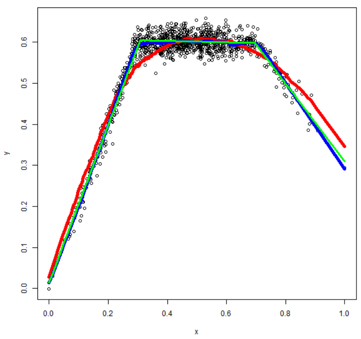
(d) Cut=25 $\sigma = 0.02$, $n = 500$



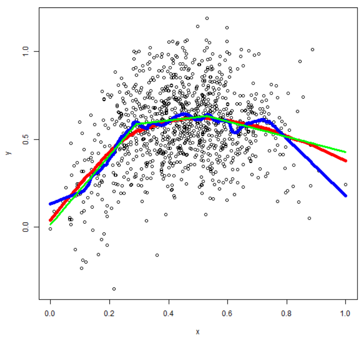
(e) Cut=25 $\sigma = 0.20$ $n = 500$



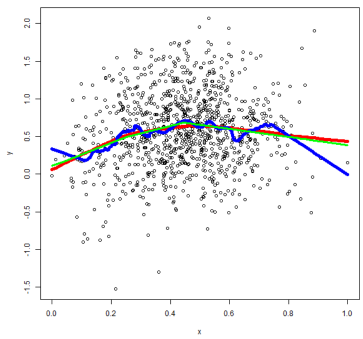
(f) Cut=25 $\sigma = 0.50$, $n = 500$



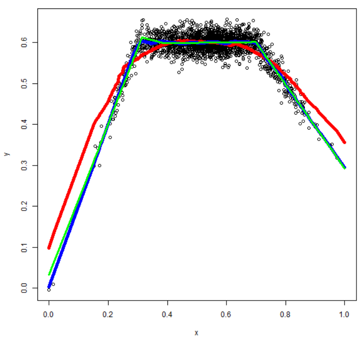
(g) Cut=25 $\sigma = 0.02$, $n = 1000$



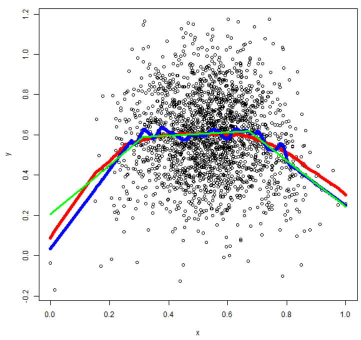
(h) Cut=25 $\sigma = 0.20$ $n = 1000$



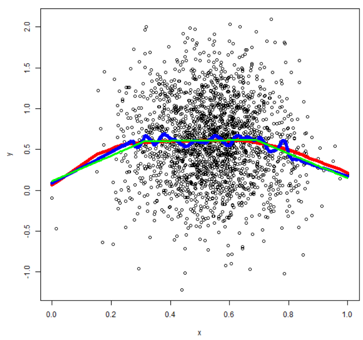
(i) Cut=25 $\sigma = 0.50$, $n = 1000$



(j) Cut=25 $\sigma = 0.02$, $n = 2000$

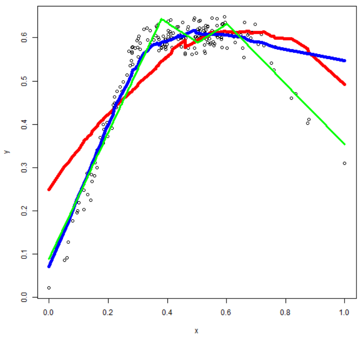


(k) Cut=25 $\sigma = 0.20$ $n = 2000$

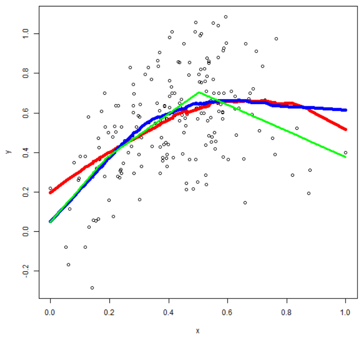


(l) Cut=25 $\sigma = 0.50$, $n = 2000$

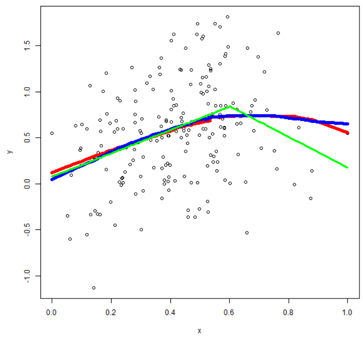
Figure A.4: Univariate BMLR Fit Comparisons: 3 Part Function with Normal Distribution



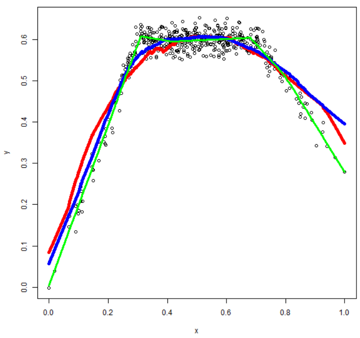
(a) Cut=50 $\sigma = 0.02$, $n = 200$



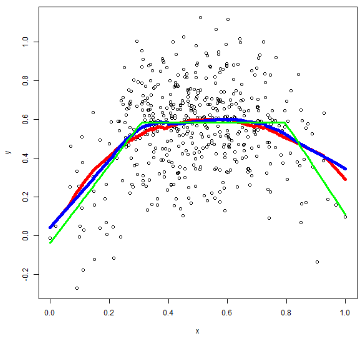
(b) Cut=50 $\sigma = 0.20$ $n = 200$



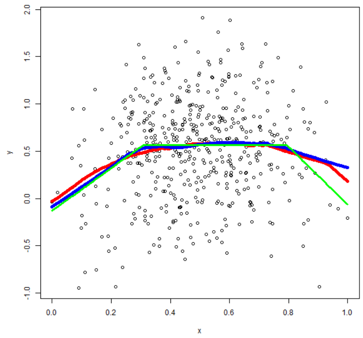
(c) Cut=50 $\sigma = 0.50$, $n = 200$



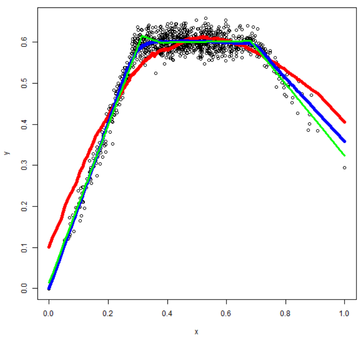
(d) Cut=50 $\sigma = 0.02$, $n = 500$



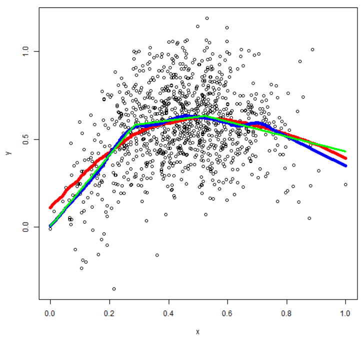
(e) Cut=50 $\sigma = 0.20$ $n = 500$



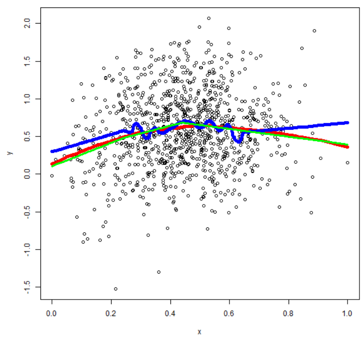
(f) Cut=50 $\sigma = 0.50$, $n = 500$



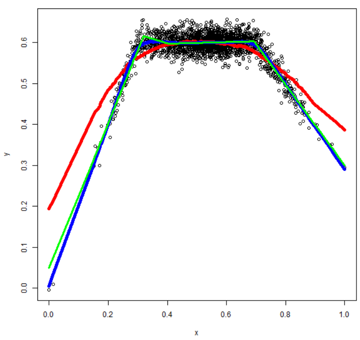
(g) Cut=50 $\sigma = 0.02$, $n = 1000$



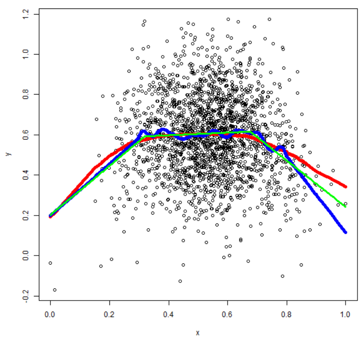
(h) Cut=50 $\sigma = 0.20$ $n = 1000$



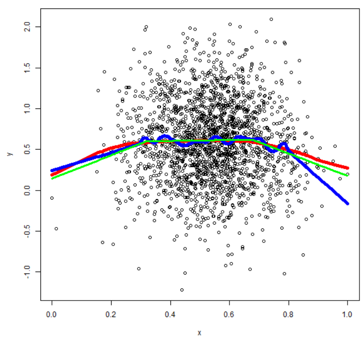
(i) Cut=50 $\sigma = 0.50$, $n = 1000$



(j) Cut=50 $\sigma = 0.02$, $n = 2000$



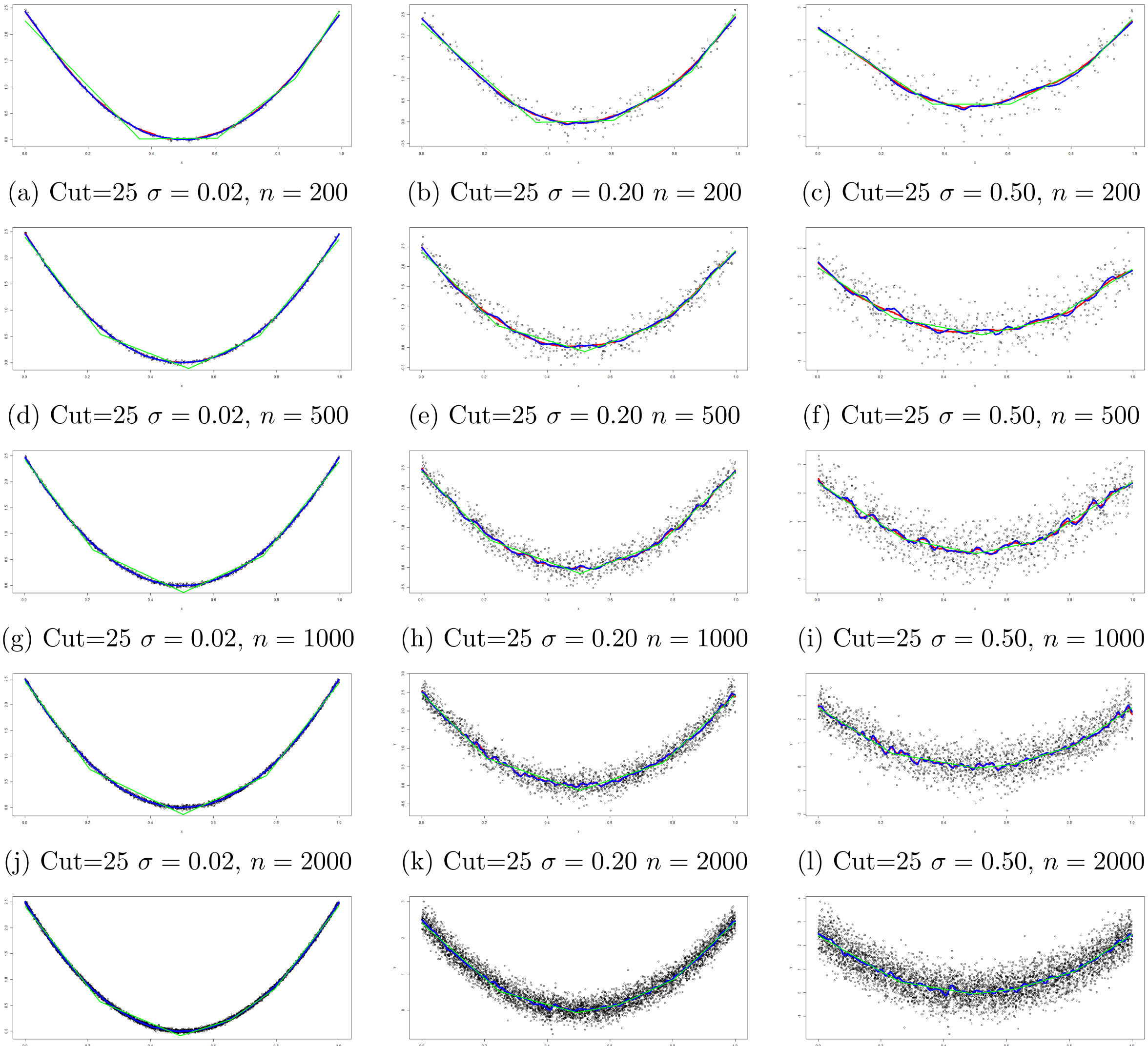
(k) Cut=50 $\sigma = 0.20$ $n = 2000$



(l) Cut=50 $\sigma = 0.50$, $n = 2000$

Notes: Data points generated using $y = f(x) + \nu$, where x is drawn at random from a standard normal distribution and then the realised set of x is normalised onto the range $[0, 1]$. ν is $f(x) = \begin{cases} 2x & for x \in [-\infty, 0] \\ 1 & for x \in (0, 1) \\ 2 - x & for x \in [1, \infty] \end{cases}$ Three values of σ are represented across the three columns, whilst rows vary by number of data points, n. Cut here relates to the minimum number of observations in a knot of the splines regression or in a ball of the BMLR regressions. Green lines represent the splines fit, red lines BMLRG and blue lines show the fit from BMLRN.

Figure A.5: Univariate BMLR Fit Comparisons: Uniforim Distribution



(a) Cut=25 $\sigma = 0.02$, $n = 200$

(b) Cut=25 $\sigma = 0.20$ $n = 200$

(c) Cut=25 $\sigma = 0.50$, $n = 200$

(d) Cut=25 $\sigma = 0.02$, $n = 500$

(e) Cut=25 $\sigma = 0.20$ $n = 500$

(f) Cut=25 $\sigma = 0.50$, $n = 500$

(g) Cut=25 $\sigma = 0.02$, $n = 1000$

(h) Cut=25 $\sigma = 0.20$ $n = 1000$

(i) Cut=25 $\sigma = 0.50$, $n = 1000$

(j) Cut=25 $\sigma = 0.02$, $n = 2000$

(k) Cut=25 $\sigma = 0.20$ $n = 2000$

(l) Cut=25 $\sigma = 0.50$, $n = 2000$

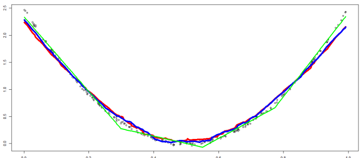
(m) Cut=25 $\sigma = 0.02$, $n = 5000$

(n) Cut=25 $\sigma = 0.20$ $n = 5000$

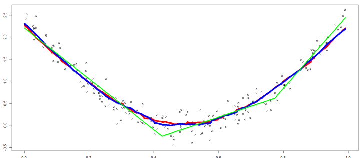
(o) Cut=25 $\sigma = 0.50$, $n = 5000$

Notes: Data points generated using $y = (0.5 - x)^2 + \nu$, where x is drawn at random from a standard normal distribution and then the realised set of x is normalised onto the range $[0, 1]$. ν is a noise term of mean 0 and standard deviation σ. Three values of σ are represented across the three columns, whilst rows vary by number of data points, n. Cut here relates to the minimum number of observations in a knot of the splines regression or in a ball of the BMLR regressions. Green lines represent the splines fit, red lines BMLRG and blue lines show the fit from BMLRN.

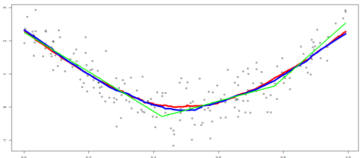
Figure A.6: Univariate BMLR Fit Comparisons: Quadratic Function Uniform Distribution



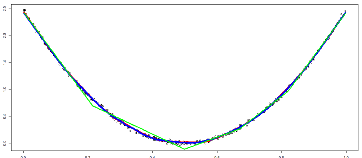
(a) Cut=50 $\sigma = 0.02$, $n = 200$



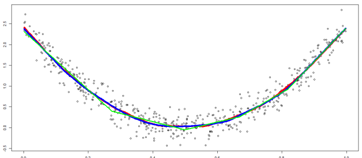
(b) Cut=50 $\sigma = 0.20$ $n = 200$



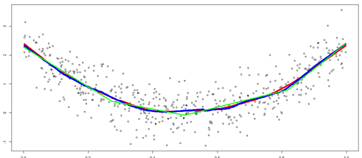
(c) Cut=50 $\sigma = 0.50$, $n = 200$



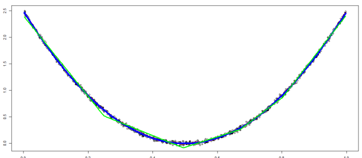
(d) Cut=50 $\sigma = 0.02$, $n = 500$



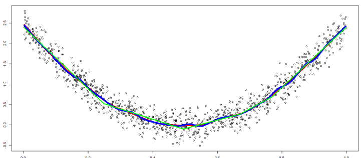
(e) Cut=50 $\sigma = 0.20$ $n = 500$



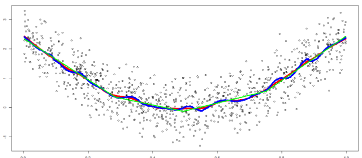
(f) Cut=50 $\sigma = 0.50$, $n = 500$



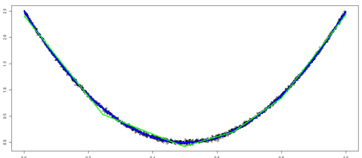
(g) Cut=50 $\sigma = 0.02$, $n = 1000$



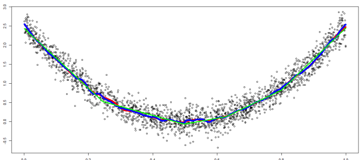
(h) Cut=50 $\sigma = 0.20$ $n = 1000$



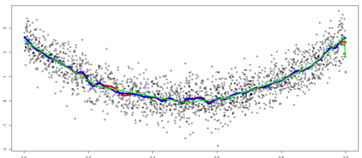
(i) Cut=50 $\sigma = 0.50$, $n = 1000$



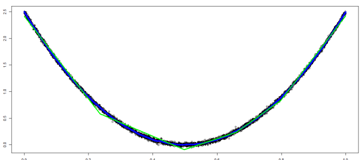
(j) Cut=50 $\sigma = 0.02$, $n = 2000$



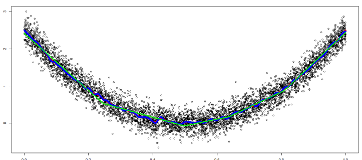
(k) Cut=50 $\sigma = 0.20$ $n = 2000$



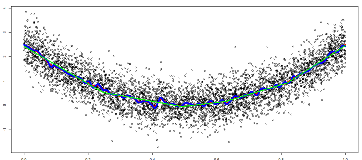
(l) Cut=50 $\sigma = 0.50$, $n = 2000$



(m) Cut=50 $\sigma = 0.02$, $n = 5000$



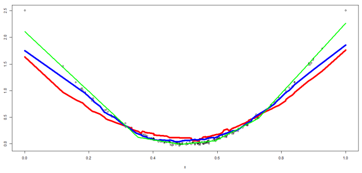
(n) Cut=50 $\sigma = 0.20$ $n = 5000$



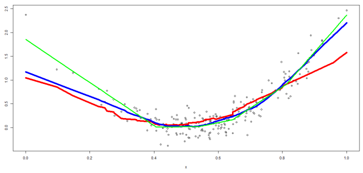
(o) Cut=50 $\sigma = 0.50$, $n = 5000$

Notes: Data points generated using $y = (0.5 - x)^2 + \nu$, where x is drawn at random from a standard normal distribution and then the realised set of x is normalised onto the range $[0, 1]$. ν is a noise term of mean 0 and standard deviation σ. Three values of σ are represented across the three columns, whilst rows vary by number of data points, n. Cut here relates to the minimum number of observations in a knot of the splines regression or in a ball of the BMLR regressions. Green lines represent the splines fit, red lines BMLRG and blue lines show the fit from BMLRN.

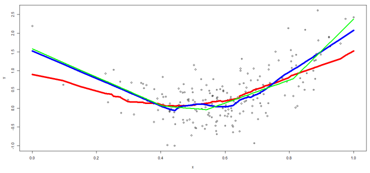
Figure A.7: Univariate BMLR Fit Comparisons: Normal Distribution



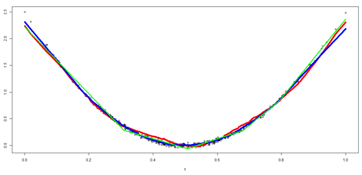
(a) Cut=25 $\sigma = 0.02$, $n = 200$



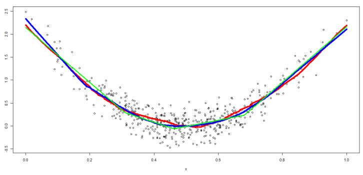
(b) Cut=25 $\sigma = 0.20$ $n = 200$



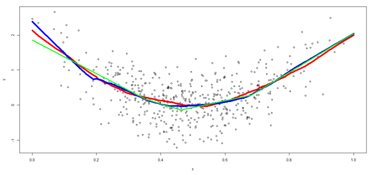
(c) Cut=25 $\sigma = 0.50$, $n = 200$



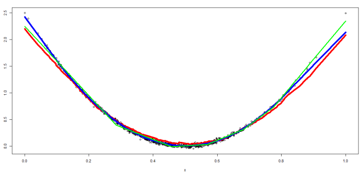
(d) Cut=25 $\sigma = 0.02$, $n = 500$



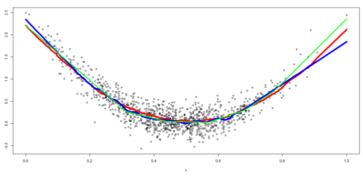
(e) Cut=25 $\sigma = 0.20$ $n = 500$



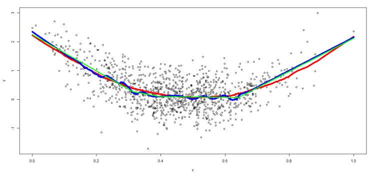
(f) Cut=25 $\sigma = 0.50$, $n = 500$



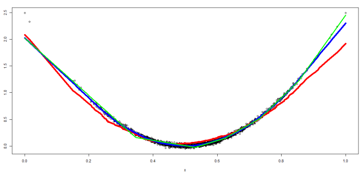
(g) Cut=25 $\sigma = 0.02$, $n = 1000$



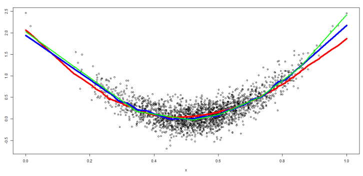
(h) Cut=25 $\sigma = 0.20$ $n = 1000$



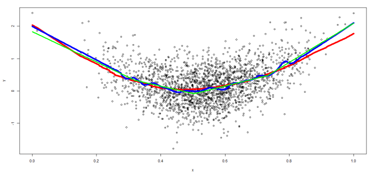
(i) Cut=25 $\sigma = 0.50$, $n = 1000$



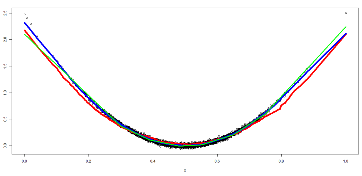
(j) Cut=25 $\sigma = 0.02$, $n = 2000$



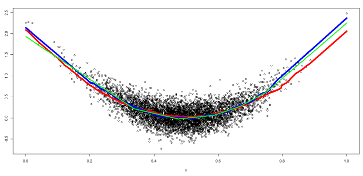
(k) Cut=25 $\sigma = 0.20$ $n = 2000$



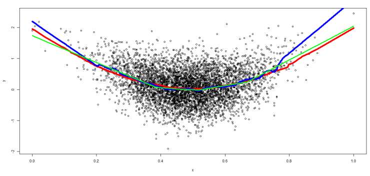
(l) Cut=25 $\sigma = 0.50$, $n = 2000$



(m) Cut=25 $\sigma = 0.02$, $n = 5000$



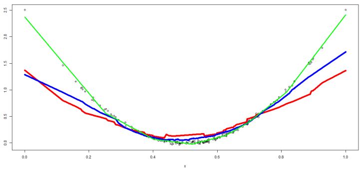
(n) Cut=25 $\sigma = 0.20$ $n = 5000$



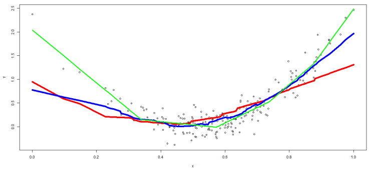
(o) Cut=25 $\sigma = 0.50$, $n = 5000$

Notes: Data points generated using $y = (0.5 - x)^2 + \nu$, where x is drawn at random from a standard normal distribution and then the realised set of x is normalised onto the range $[0, 1]$. ν is a noise term of mean 0 and standard deviation σ. Three values of σ are represented across the three columns, whilst rows vary by number of data points, n. Cut here relates to the minimum number of observations in a knot of the splines regression or in a ball of the BMLR regressions. Green lines represent the splines fit, red lines BMLRG and blue lines show the fit from BMLRN.

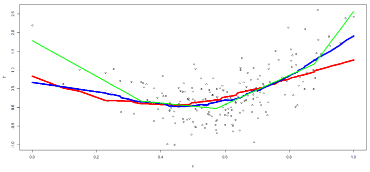
Figure A.8: Univariate BMLR Fit Comparisons: Quadratic Function Normal Distribution



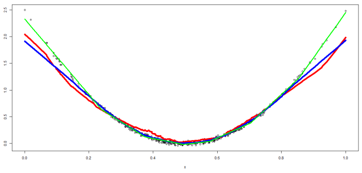
(a) Cut=50 $\sigma = 0.02$, $n = 200$



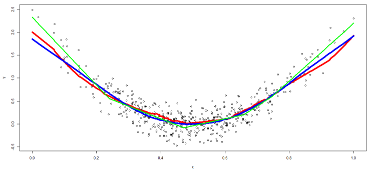
(b) Cut=50 $\sigma = 0.20$ $n = 200$



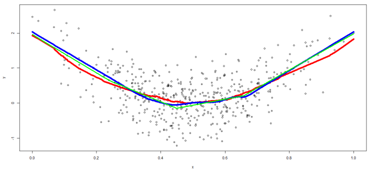
(c) Cut=50 $\sigma = 0.50$, $n = 200$



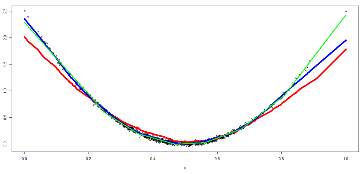
(d) Cut=50 $\sigma = 0.02$, $n = 500$



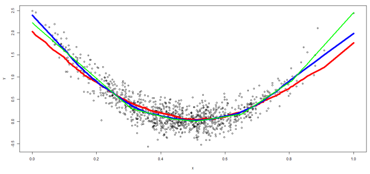
(e) Cut=50 $\sigma = 0.20$ $n = 500$



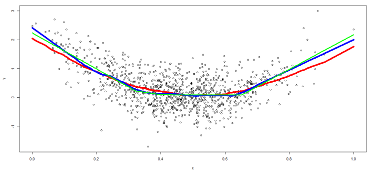
(f) Cut=50 $\sigma = 0.50$, $n = 500$



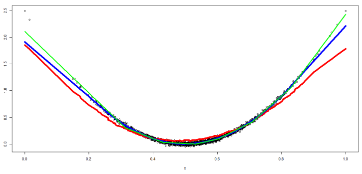
(g) Cut=50 $\sigma = 0.02$, $n = 1000$



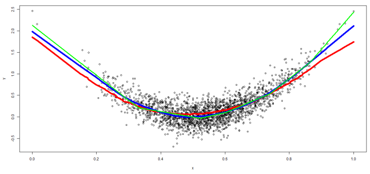
(h) Cut=50 $\sigma = 0.20$ $n = 1000$



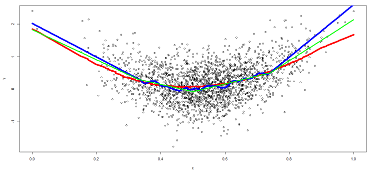
(i) Cut=50 $\sigma = 0.50$, $n = 1000$



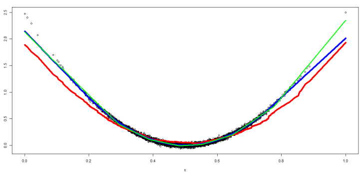
(j) Cut=50 $\sigma = 0.02$, $n = 2000$



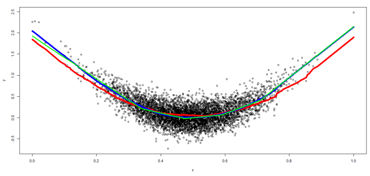
(k) Cut=50 $\sigma = 0.20$ $n = 2000$



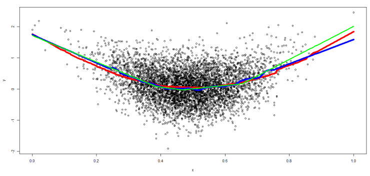
(l) Cut=50 $\sigma = 0.50$, $n = 2000$



(m) Cut=50 $\sigma = 0.02$, $n = 5000$



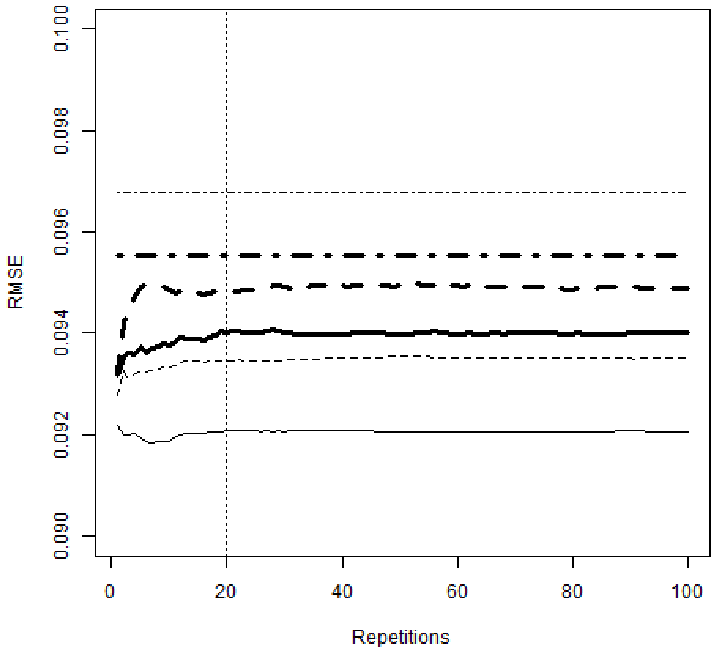
(n) Cut=50 $\sigma = 0.20$ $n = 5000$



(o) Cut=50 $\sigma = 0.50$, $n = 5000$

Notes: Data points generated using $y = (0.5 - x)^2 + \nu$, where x is drawn at random from a standard normal distribution and then the realised set of x is normalised onto the range $[0, 1]$. ν is a noise term of mean 0 and standard deviation σ. Three values of σ are represented across the three columns, whilst rows vary by number of data points, n. Cut here relates to the minimum number of observations in a knot of the splines regression or in a ball of the BMLR regressions. Green lines represent the splines fit, red lines BMLRG and blue lines show the fit from BMLRN.

Figure A.9: Repetitions and Fit: Three Part Function Uniform Distribution



Notes: Solid lines denote BMLRN, dashed lines show BMLRG, dotdash lines show splines. In each case a thicker line is used for the cut size of 50 and a thinner line for a cut size of 25. $n = 500$.