Causal and Statistical Inference Kai Su, Qihang Chen

Poster Report: RDD

1 Analyzing Baysian Network

1.1 Introduction to RDD

When the running variable X takes value at different side of a threshold value t, there will be a treatment(marked $w = 1$) having an effect on the output Y on one side of $X = t$ and no treatment(marked $w = 0$) on the other side. Our goal is to measure the average effect τ solely caused by the treatment. That is, $\tau = E[Y|\operatorname{do}(W) = 1] - E[Y|\operatorname{do}(W) = 0]$.

Such effect appears as a sudden increase or decrease of y in the neighborhood of $X = t$(See Fig.1). Therefore, τ is traditionally estimated using the see effect of W. Formally put,

$$\hat{\tau} = \lim_{x \to t^+} E[Y|X = x] - \lim_{x \to t^-} E[Y|X = x]. \tag{1}$$

For simplicity, denote

$$E[Y|X = t_+] = \lim_{x \to t^+} E[Y|X = x], \tag{2}$$

$$E[Y|X = t_-] = \lim_{x \to t^-} E[Y|X = x]. \tag{3}$$

However, we are lack of data near the threshold in most cases, difficult for us to calculate these values on the threshold directly. Thus we need some regression to infer the relation between X and Y so as to predict $E[Y|X = t_+]$ and $E[Y|X = t_-]$.

A common practice is to apply linear regression on both sides. The data near the threshold is more valuable, thus we can set a bandwidth b, which is the largest distance where data are taken into account; and a kernel k, which assigns weights to data. Denote such an estimator as $\hat{\tau}(b, k)$ for simplicity.

1.2 Baysian Network

Basically, running variable X will decide W and affect Y; W will have effect on Y, as in Fig.2.

bandwidth

When the bandwidth is sufficiently small,

$$P(Y|W = 0, X = t) \approx P(Y|X = t_-), \tag{4}$$

$$P(Y|W = 1, X = t) \approx P(Y|X = t_+). \tag{5}$$

Then

$$\begin{aligned} \tau &= E[Y|\operatorname{do}(W = 1), X = t] - E[Y|\operatorname{do}(w = 0), X = t] \qquad (6)\\ &\approx E[Y|X = t_+] - E[Y|X = t_-] = \hat{\tau}, \qquad (7) \end{aligned}$$

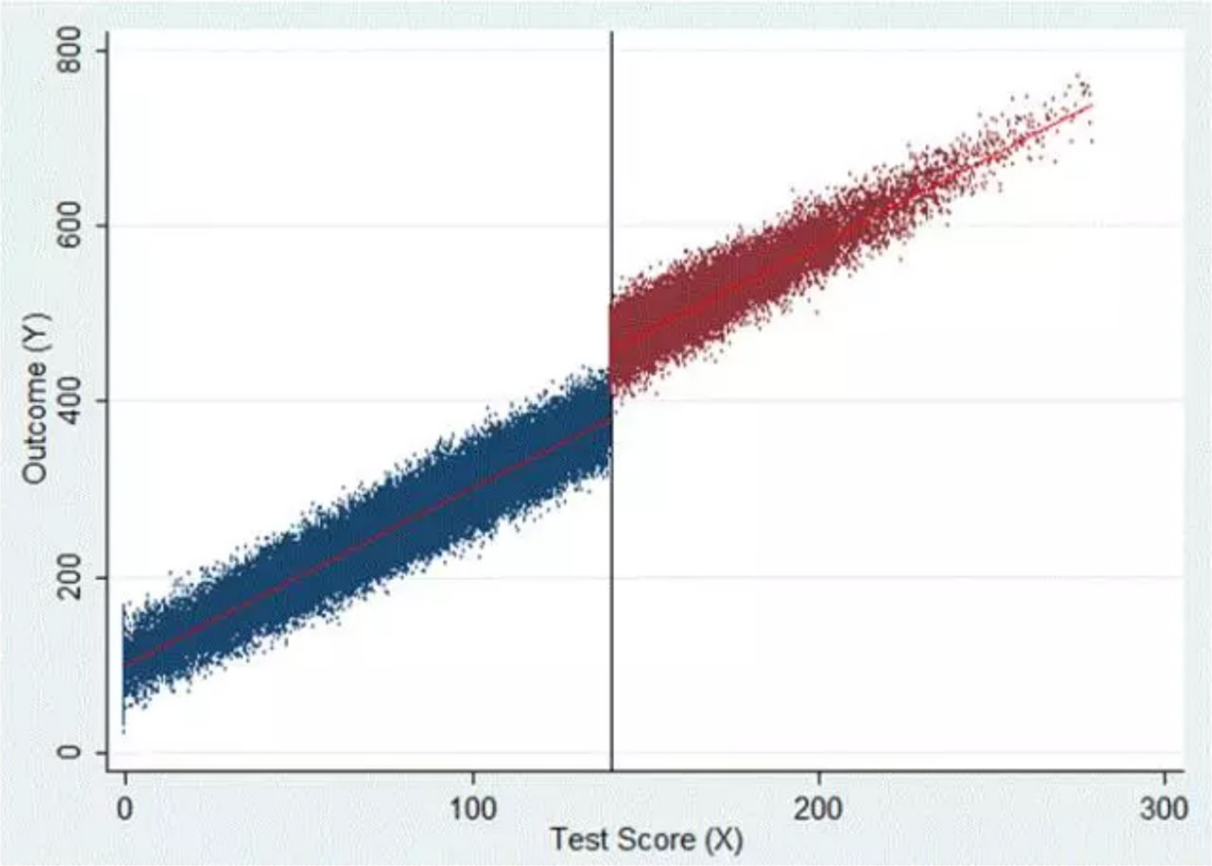


Figure 1: An example of dataset in RDD

proving that the traditional estimation in Eqn.(1) is unbiased.

In contrast, if the bandwidth b is non-negligible, this selection of data specified by b will cause a backdoor path between X and Y. (See Fig.3) Then the see effect observed by regression between X and Y is not the true causal effect, leading to a bias. To eliminate this bias, we need to adjust the way of sampling by changing the kernel.

covariates

Sometimes there are not only running variables X and Y, but also many other variables Z, called covariates, may have effect on Y. If they are independent with X, the regression still works. However in some cases, Z will affect both X and Y, creating a backdoor path between X and Y. (See Fig.4)

In order to eliminate this bias, we need to find a bandwidth, to make sure among the included samples, Z is independent with X. (See Fig.5)

1.3 Obstacles

Under the BN shown in Fig.3, the performance of the estimator $\hat{\tau}(b, k)$ is restricted by two factors:

(a) Large variance due to lack of data;

(b) Bias caused by fitting non-linear $X - Y$ relation with linear regression.

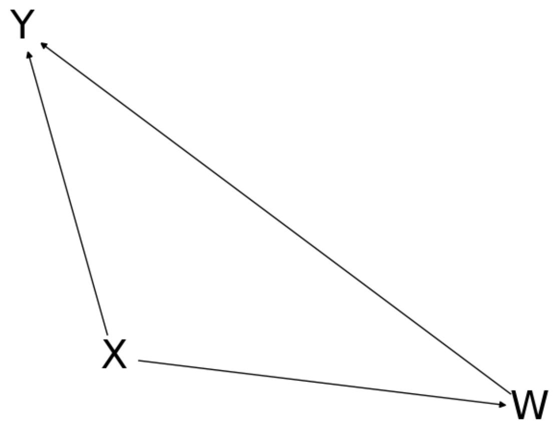


Figure 2: The basic BN for RDD

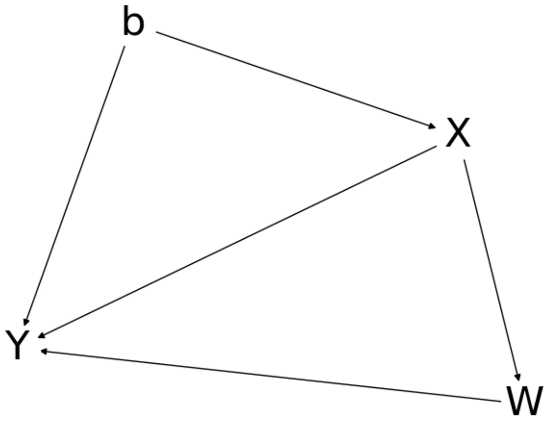


Figure 3: BN with bandwidth

When $b \to 0$, less and less data stay within the bandwidth, so the variance in (a) becomes significant. The bias in (b) converges to 0 under smoothness assumption on the $X - Y$ relation. The conclusion is opposite when $b \to +\infty$.

2 Experiment and Discussion

2.1 Using kernel to elimimate bias caused by bandwidth

In this section we do not consider covariates. We generate data by different $X - Y$ relation and different distribution of samples (See table 6). The cutoff is at $X = 59$.

We randomly generated 500 groups of data and 200 different types of kernels using bezier curves. For each kernel, we calculate the difference of Y at threshold compared with $Y(t)$ by True $X - Y$ relation in those 500 groups of data, getting the average to show the performance of the kernel. Thus we get the best kernel for different bandwidth.

Kernel varies with bandwidth

A typical result(case30) is shown in Fig.7. The x-axis is the distance from the cutoff, and the y-axis is the weight of the kernel.At a fixed bandwidth, the kernel is not just concentrating

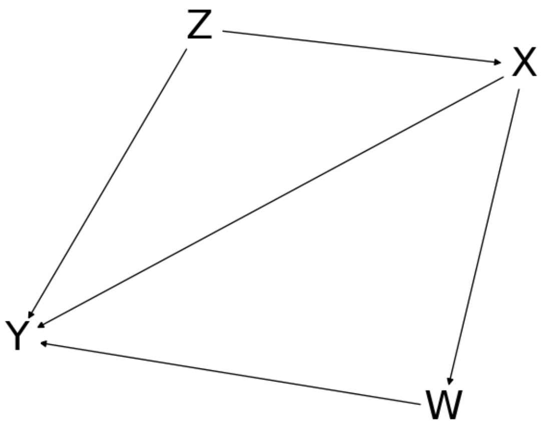


Figure 4: BN with covariates

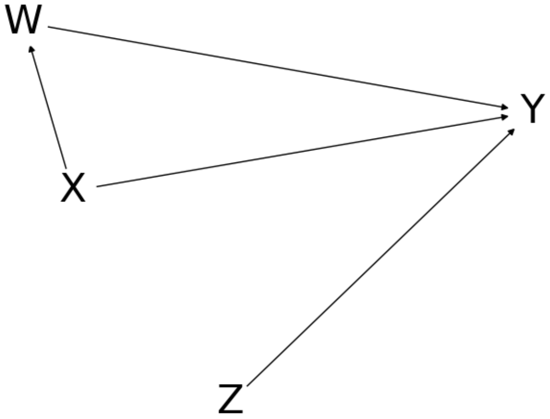


Figure 5: eliminate $Z \rightarrow X$ path

near the cutoff, but first chooses to spread out the weight to larger distance to the cutoff when bandwidth is small.

This consequence results from the randomness of data. Too few samples will cause large uncertainty. Giving more weight to those father samples has the similar effect of obtaining more samples, in order to lower the variance of estimation. As the bandwidth b becomes larger, the weights concentrate to the cutoff to lower the bias caused by the nonlinearity in $X - Y$ relation.

Kernel varies with sample distribution

Another observation is that, the distribution of samples along X axis really affect the performance of kernels. The reason is the distribution of samples affect the randomness of sampling, thus we need to use kernel to adjust the sampling method.

Comparing the result in different rows(in html files, can not be shown in pdf), we can find a regular pattern that adding weight on those samples where samples along the X axis are sparse will give better performance.

The reason is intuitive. If the samples are exactly randomly distributed, it will be easier for us to infer the information from data. But if some areas on X axis has larger density of samples, we shall give less weight to them in order to keep the sampling method "seems to be still random". Then the kernel will trend to samples in the sparse area.

$y =$	$0.2(x-59)^2+2$	$-0.2(x-59)^2-0.4(x-59)+2$	$0.1(x-59)^3+2$
Dense middle	case00	case01	case02
Sparse middle	case10	case11	case12
Dense cutoff	case20	case21	case22
Sparse cutoff	case30	case31	case32

Figure 6: $X-Y$ relations and data distributions

2.2 Differences after adding covariates

We generate data by different $X-Y$ relation and different ways Z affect X, Y (See table 9) where the samples are uniformly distributed.

After adding covariates, there will be another backdoor path caused by Z, and the perfomance of different kernels becomes different (See Fig.8). This tells us our selection of bandwidth should concern the effect of covariates, and it will be unbiased only when the BNS is the same as Fig.5.

3 Other Attempts

A better way to find best bandwidth

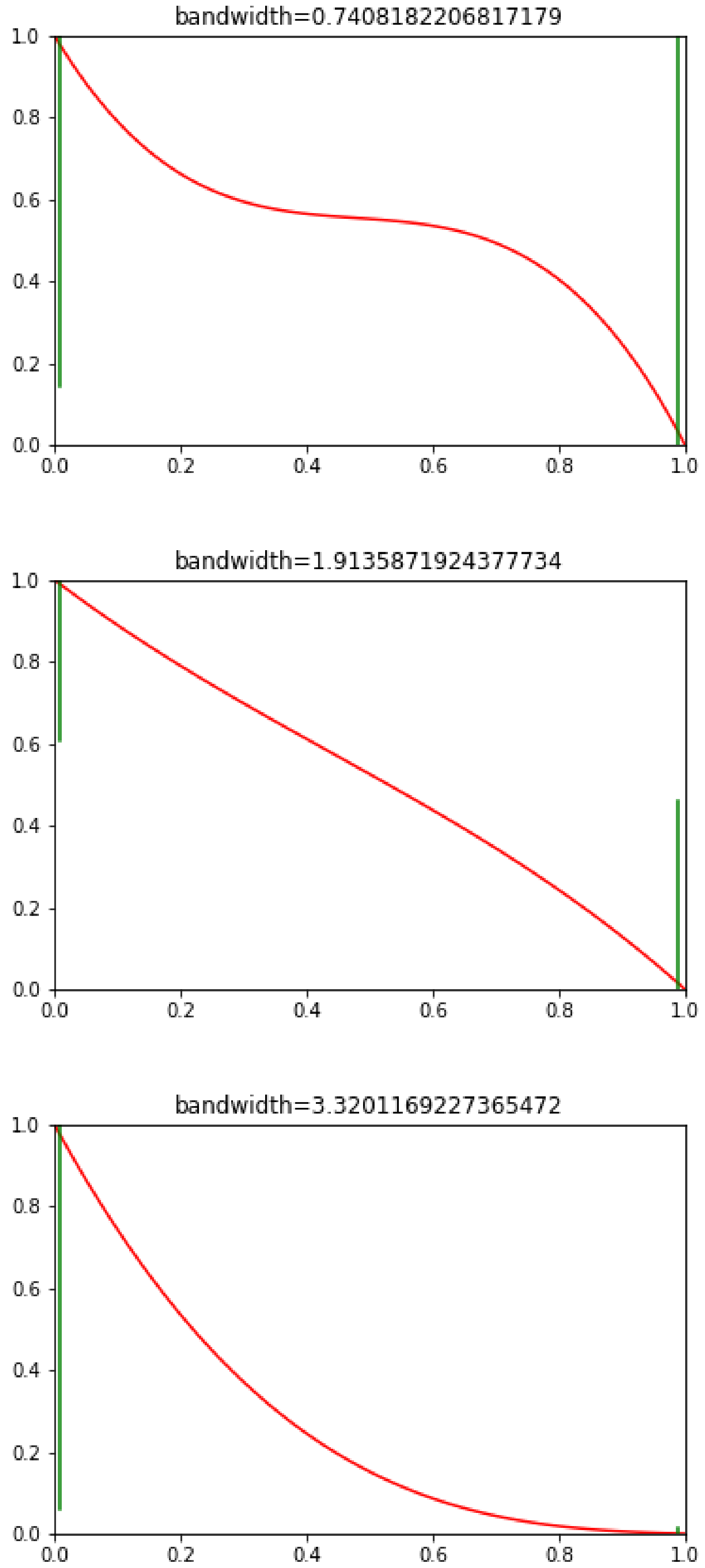
In some articles, cross validation method is used for selecting a optimal bandwidth. However this has not always accurate because samples near threshold may not have similar distribution and $X-Y$ relation.

We have these ways to quantify the expected error for different bandwidth caused by two aspects mentioned in section 1.3:

(a) Use bandwidth b, suppose the linear regression has the result $Y = a(X-t)+b$, we use a formula to calculate the standard derivation of b to represent the expected error caused by lack of data near threshold.

(b) First use a quadratic hypothesis function on regression to find out a curve showing approximate quadratic relationship between $X-Y$(do not use bandwitdh b). Then project all the samples onto the curve, use linear regression with bandwidth b and find the difference of value between the two ways of regression on the threshold.

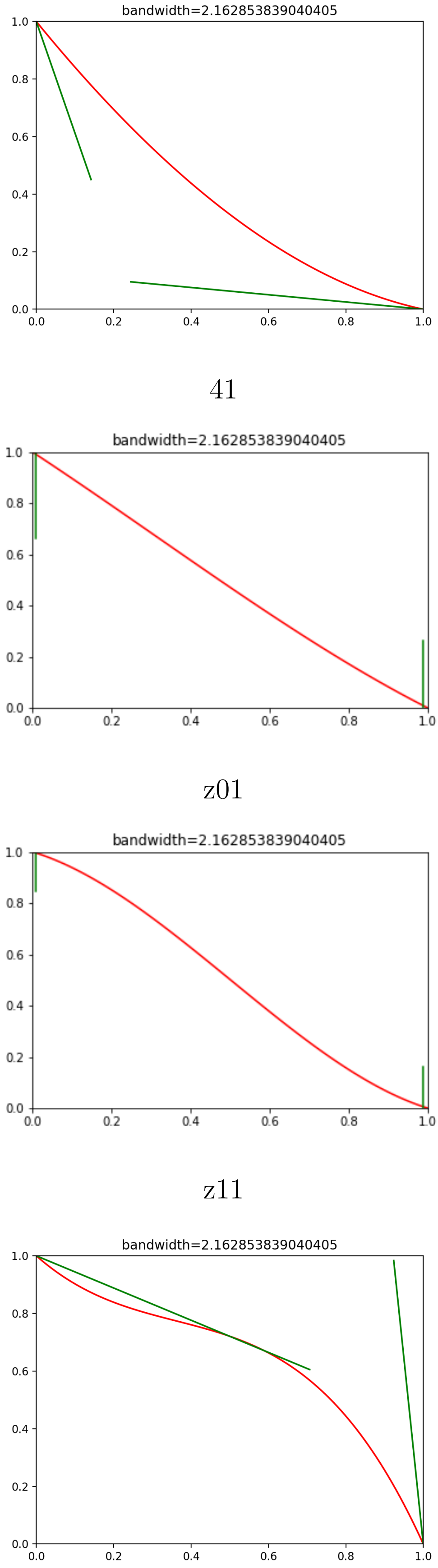
Then we find a bandwidth who has least sum of those two kind of errors, just the optimal bandwidth we need. (See table 10)


bandwidth=0.7408182206817179

bandwidth=1.9135871924377734

bandwidth=3.3201169227365472

Figure 7: Illustrated best kernels under different bandwidths



41

z01

z11

z21

Figure 8: Illustrated best kernels under different ways of adding covariates

$f(x) =$	$0.2(x-59)^2+2$	$-0.2(x-59)^2-0.4(x-59)+2$	$0.1(x-59)^3+2$
$x = x_0$ $y = f(x)$	case40	case041	case42
$x = x_0 + z$ $y = f(x) + 0.3z$	casez00	casez01	casez02
$x = x_0 + z^2$ $y = f(x) + 0.3z$	casez10	casez11	casez12
$x = x_0 + z$ $y = f(x) + 0.5Z^2$	casez20	casez21	casez22

Figure 9: $X - Y$ relations and ways of adding Z

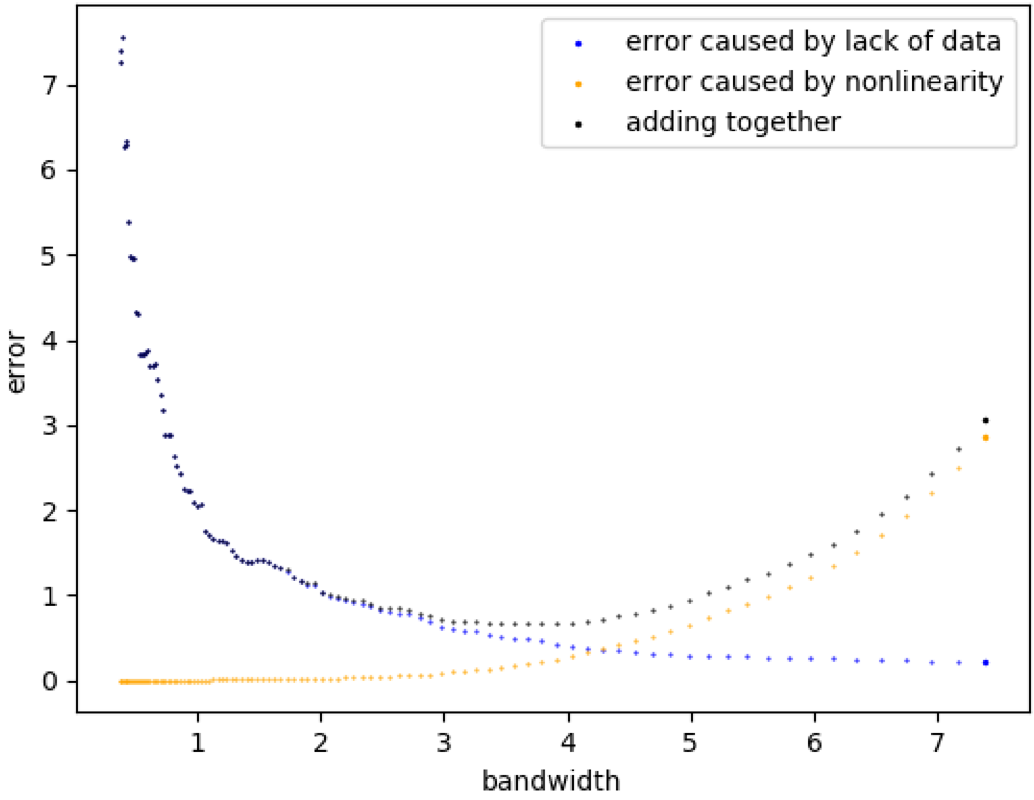


Figure 10: Evaluate error of different bandwidth (on the headstart dataset)