Econ 589 Econometrics: Problem set 2

Ste Rose

November 14, 2019

1 Question One.

1.1 Part (a).

We consider a nonparametric kernel density estimator with standard normal kernel and fixed n. The bandwidth h is effectively a measure of how close we want random points x_i to be to the point for which we're estimating the density, x (of course in a more sophisticated way via the use of the kernel; rather than a simple binary choice of close enough to count as the same, the kernel with the bandwidth determines a weight).

If the bandwidth h tends to infinity, then every point in the support is considered to be "close enough" to count as x, and in fact the argument of every entry in the sum tends to 0. Since k is continuous, the numerator in the estimator tends to $k(0)$, but the denominator h tends to infinity, so that the estimator tends to 0.

If the bandwidth h tends to 0, then every point in the support other than x (if it occurs in the random sample) is considered to be too far away to count at all as a point close to x, in which case both the numerator and the denominator of the estimator tend to 0. Intuitively, we expect that as the kernel is a transformation of an exponential, that the numerator tends to 0 faster than the denominator, so that the limit is 0. For all $x_i \neq x$, we can use the Taylor expansion of ϕ the normal pdf to find the limit

$$\lim_{h\to 0} \frac{1}{h} k\left(\frac{x - x_i}{h}\right) = \frac{1}{\sqrt{2\pi}} \lim_{h\to 0} \exp\left(-\frac{(x-x_i)^2}{h^2}\right) \frac{1}{h}. \tag{1}$$

By definition of the exponential function,

$$e^{\frac{(x-x_i)^2}{h^2}} = 1 + \left(\frac{x - x_i}{h}\right)^2 + O(h^{-4}), \tag{2}$$

where the remainder term here is strictly positive, so that

$$e^{\frac{(x-x_i)^2}{h^2}} > \left(\frac{x - x_i}{h}\right)^2, \tag{3}$$

and hence

$$\frac{h^2}{(x-x_i)^2} > e^{-\left(\frac{x-x_i}{h}\right)^2} > 0. \tag{4}$$

Clearly the limit of the left hand side divided by h tends to 0, and hence our limit of interest tends to 0. In the case where the random sample does not contain a j such that $x_j = x$, this implies that, given n is fixed, the estimator tends to 0.

If one is unfortunate enough to have such an $x_j = x$, then all terms in the sum tend to 0 except the term corresponding to x_j, which is

$$\frac{k(0)}{h}, \tag{5}$$

which tends to infinity.

1.2 Part (b).

We now consider the same question for the regression density estimator, which is

$$\hat{m}(x) = \frac{\hat{a}(x)}{\hat{f}(x)}, \tag{6}$$

where $\hat{f}(x)$ is our kernel density estimator and

$$\hat{a}(x) = \frac{1}{nh} \sum_{i \leq n} k\left(\frac{x - x_i}{h}\right) y_i. \tag{7}$$

This implies that

$$\hat{m}(x) = \frac{\sum_{i \leq n} k\left(\frac{x-x_i}{h}\right) y_i}{\sum_{i \leq n} k\left(\frac{x-x_i}{h}\right)}. \tag{8}$$

Annoyingly, it is clear that both $\hat{a}(x)$ and $\hat{f}(x)$ will tend to 0 as h tends to infinity, and 0 as h tends to zero if x is not in the sample, and infinity if x is in the sample. However, given our expression for the regression estimator, we can write

$$\hat{m}(x) = \sum_{i \leq n} y_i \frac{k\left(\frac{x-x_i}{h}\right)}{\sum_{j \leq n} k\left(\frac{x-x_j}{h}\right)}. \tag{9}$$

As h goes to infinity, each of these terms tend to

$$y_i \frac{k(0)}{nk(0)}, \tag{10}$$

so that the estimator tends to

$$\frac{1}{n} \sum_{i \leq n} y_i, \tag{11}$$

which is the sample mean of y_i.

We now consider $h \to 0$. The reciprocal of the factor multiplying y_i is

$$\sum_{j \leq n} \frac{k\left(\frac{x-x_j}{h}\right)}{k\left(\frac{x-x_i}{h}\right)} = 1 + \sum_{j \neq i} \frac{k\left(\frac{x-x_j}{h}\right)}{k\left(\frac{x-x_i}{h}\right)}. \tag{12}$$

We consider one of the terms in this sum,

$$\frac{k\left(\frac{x-x_j}{h}\right)}{k\left(\frac{x-x_i}{h}\right)} = \exp\left(\frac{1}{2h^2}(x_j - x_i)(2x - x_i - x_j)\right). \tag{13}$$

If the coefficient of h^{-2} is positive for any j, then the expression in equation (12) tends to infinity. If this is the case for each x_i, the terms multiplying y_i in our regression estimator all tend to 0, and hence $\hat{m}(x)$ tends to 0.

The coefficient is positive if

$$x_i < x_j < 2x - x_i, \tag{14}$$

or

$$2x - x_i < x_j < x_i. \tag{15}$$

If this isn't true for any x_j, then the expression in equation (12) tends to 1, in which case y_i is not killed by the limit, and we have

$$\hat{m}(x) \to \frac{1}{n} \sum_{i \leq n} y_i \mathbb{1}\{\forall j : x_j < x_i \iff 2x \geq x_j + x_i\}. \tag{16}$$

2 Question Two.

2.1 Part (a).

We conduct a simulation exercise whereby we draw random numbers from a standard Gumbel distribution, and compute the nonparametric kernel density estimates using the following kernels:

$$k_1(t) = \phi(t), \tag{17}$$

$$k_2(t) = \phi(t)\frac{3 - t^2}{2}, \tag{18}$$

$$k_3(t) = 2\phi(t) - \frac{1}{\sqrt{2}}\phi\left(\frac{t}{\sqrt{2}}\right), \tag{19}$$

$$k_4(t) = \frac{\sin(\pi t)}{\pi t}. \tag{20}$$

We consider $n = 10$, $n = 100$, and $n = 1000$, and replicate each 1000 times, so that we can plot the root mean square error of $\hat{f}(1)$ against the bandwidth h, which we take from 0.05 to 10, in increments of 0.05. The following plots are plots of each estimator corresponding to the kernels with each sample size n on a different graph. This is shown below.

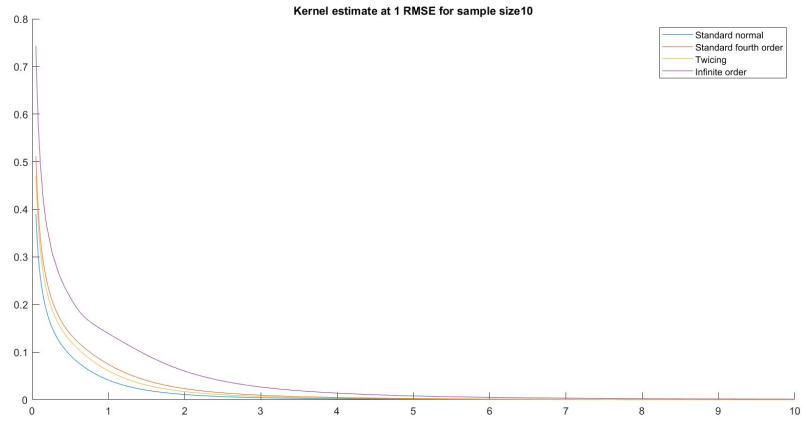

Figure 1: Root mean square error of $\hat{f}(1)$ as a function of bandwidth for sample size 10.

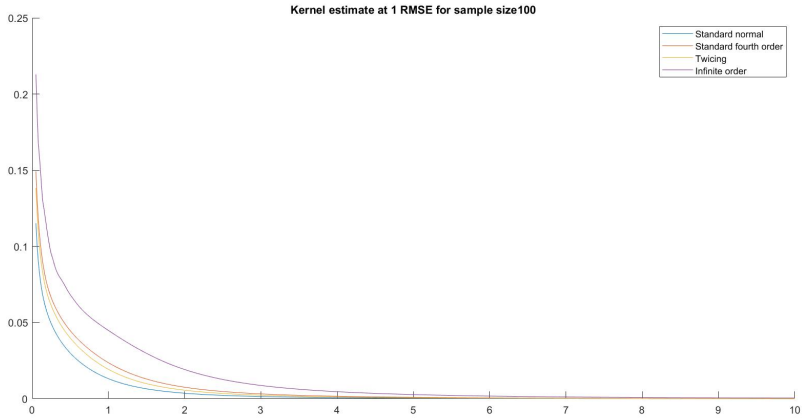

Figure 2: Root mean square error of $\hat{f}(1)$ as a function of bandwidth for sample size 100.

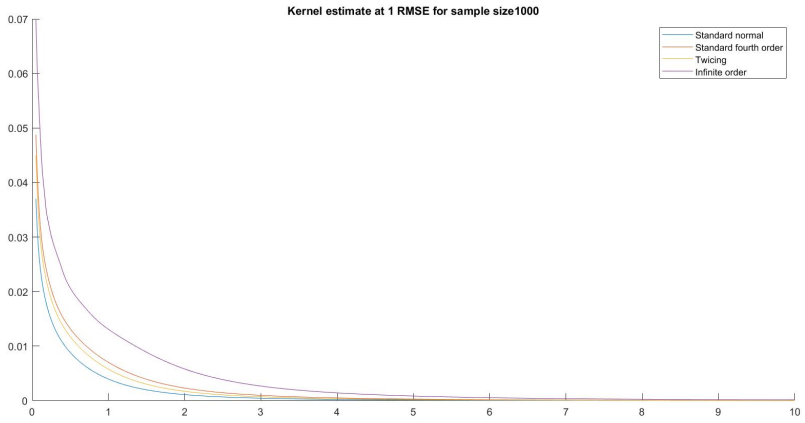

Figure 3: Root mean square error of $\hat{f}(1)$ as a function of bandwidth for sample size 1000.

2.2 Part (b).

We subsequently use the previously generated data and the regression function

$$m(x) = \log(1 + x^2), \tag{21}$$

to generate the data

$$y_i = m(x_i) + u_i, \tag{22}$$

where we choose u_i to be distributed as a standard normal. We again plot the root mean squared error against h, shown below. We note that this seems to be monotonic in h for both the regression and density estimation, which I think indicates that I've not considered a wide enough range, or that I've done something wrong in the code. Either way, I have graphs and a short life, so, here we are.

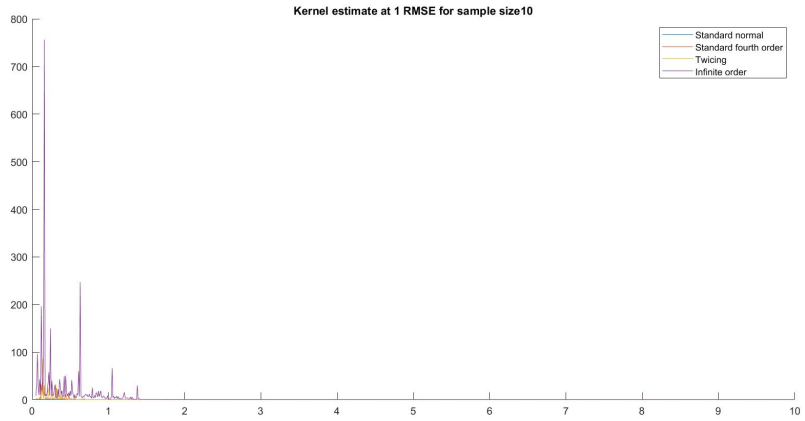

Figure 4: Root mean square error of $\hat{m}(1)$ as a function of bandwidth for sample size 10.

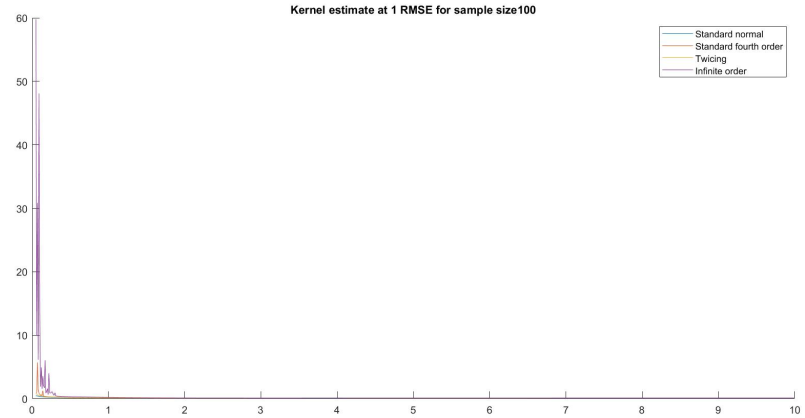

Figure 5: Root mean square error of $\hat{m}(1)$ as a function of bandwidth for sample size 100.

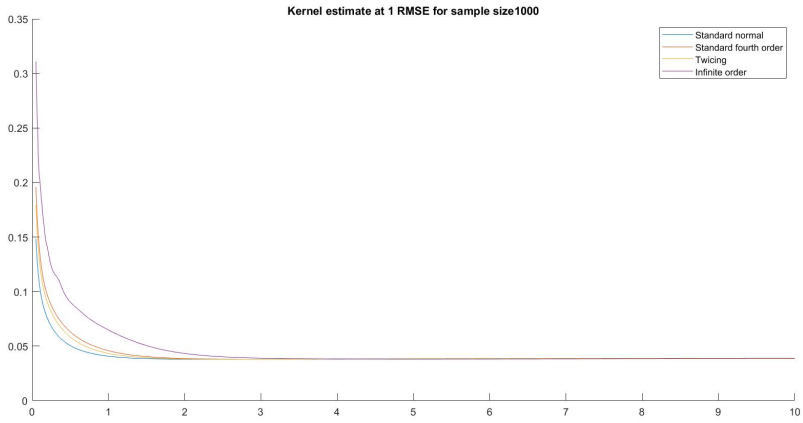

Figure 6: Root mean square error of $\hat{m}(1)$ as a function of bandwidth for sample size 1000.

3 Question Three.

3.1 Part (a).

We consider again a simulation using the standard Gumbel distribution, and consider two extremes for bandwidths, $h_1 = 10^{-6}$, and $h_2 = 10^6$. We use the standard normal kernel, presumably since there is evidence that the shape of the kernel is somewhat unimportant in determining the accuracy of the statistic [10]. Using the same data set consisting of $n = 1000$ observations, we plot the estimated distribution below for each bandwidth in Figure 7.

This is a mess, which I can only attribute to the absurd choices of bandwidths. I assume the spikes occur where the x value is included in the sample.

For context, the Silverman bandwidth here is 0.3372.

3.2 Part (b).

As before, we do the same thing with the regressor m. This is plot in Figure 8.

This, clearly, is another mess.

3.3 Part (c).

Finally, we consider using Silverman's rule of thumb:

$$h_s = \hat{\sigma} n^{-\frac{1}{5}}. \tag{23}$$

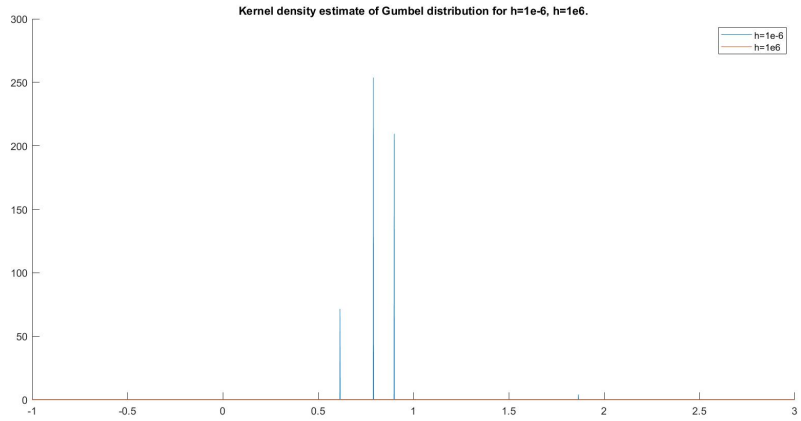

Figure 7: Apparently the kernel estimator for the density of a standard Gumbel distribution. The kernel is the standard normal.

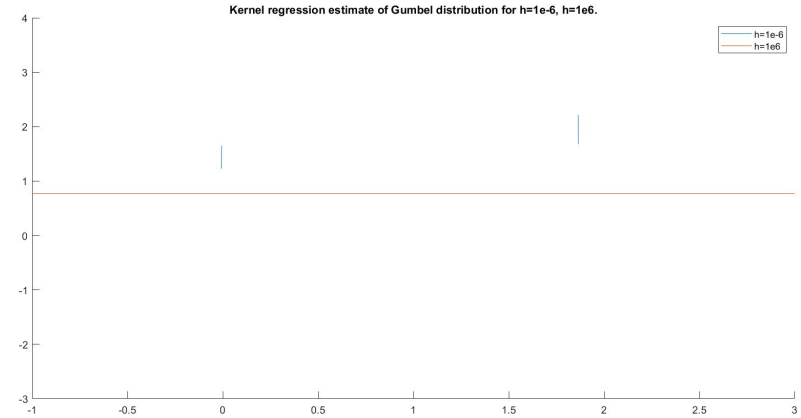

Figure 8: Apparently the kernel estimator for the regressor of a standard Gumbel distribution. The kernel is the standard normal.

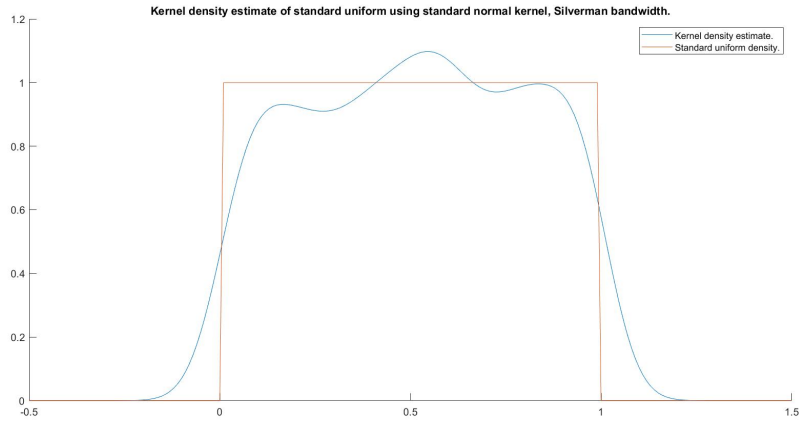

Figure 9: The kernel density estimator of the standard uniform distribution with standard normal kernel and Silverman bandwidth.

For our data set here, we get that $h_s = 0.0725$. The plot of the density is given in Figure 9.

We see that this fits much better, which is unsurprising, given the far more appropriate bandwidth.

4 Question Four.

4.1 Part (a).

My name is Ste. It begins with the letter 'S.' So does the name of the state of South Carolina. We therefore look at income microdata from South Carolina in 2010 provided by the US census bureau. Throughout we use the standard normal kernel, and investigate options for bandwidth selection to construct a nonparametric estimation of the income distribution.

We begin using Silverman's rule of thumb: $h_s = \hat{\sigma} n^{-\frac{1}{5}}$. We implement this on Matlab, which runs quickly enough that I don't have time to procrastinate somewhere else. The kernel estimator is plot in Figure 10. We note that the values are absurdly small (of the order 10^{-6}) but this is not too concerning since the support of the distribution is very large.

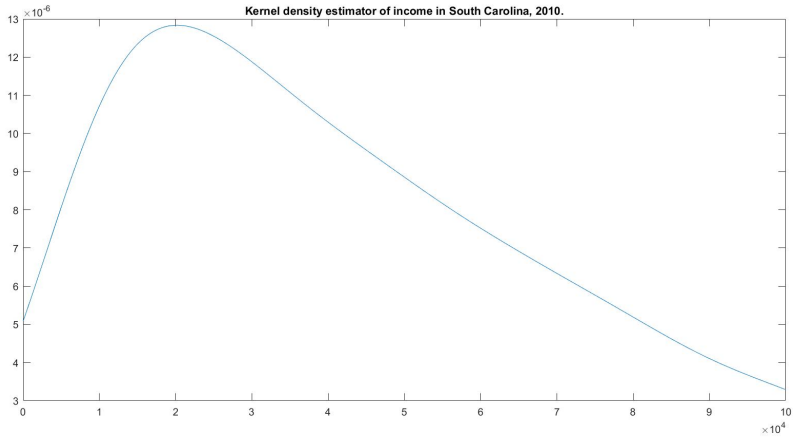

Figure 10: The kernel density estimator of household income in South Carolina, 2010. The standard normal kernel is used, and the Silverman rule of thumb bandwidth, 0.0725 is used.

4.2 Part (b).

We now use cross validation, which is to say, we choose h to maximise

$$\sum_{i \leq n} \log \left\{ \frac{1}{(n-1)h} \sum_{j \neq i} k\left(\frac{x_j - x_i}{h} \right) \right\}. \tag{24}$$

Our dataset (which we clean up with a few lines of code in Matlab), is somewhat large, consisting of over 19,000 observations. When we tried running this code in Matlab, we found that it ran far too slowly.

We therefore migrated to c++ for the calculation of the cross-validated bandwidth. Despite the suggestion that classes would not be needed, it turned out that the easiest way to construct the objective that we need to maximise, it made sense to create a class consisting of a public array containing the data, and a public function that is the derivative of the objective (the derivative since we wrote non-linear solvers, a secant method and a bisection method.

When we ran the secant algorithm, we found that while we could get the first order condition close to 0 (both positive and negative either side, so that, given the continuity of the function, we knew there was some finite root, and in fact at least two), the algorithm did not converge sufficiently quickly. We therefore identified a range on which a root had to exist by the intermediate value theorem, $[914, 1200]$ (excluding negative bandwidths, though it's worth noting in passing that for the starting value -2871.93, the secant algorithm did actually converge fairly quickly). The bisection algorithm (set to a tolerance of 10^{-6}) yielded the cross-validation bandwidth value 1059.4696. I'm surprised at

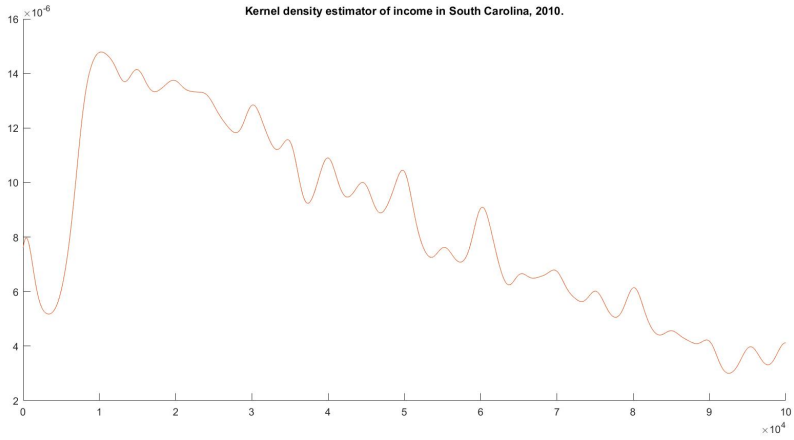

Figure 11: The kernel density estimator of household income in South Carolina, 2010. The standard normal kernel is used, and the cross-validation bandwidth 1059.4696 is used.

how large this is, but it gives a somewhat reasonable curve, so fine[1]. The added lumpiness is an effect of the larger bandwidth, since the estimator is relying on data points further out from each x. The kernel estimator is plot in Figure 11.

5 Question Five.

We are asked to find three examples of applied papers employing kernel regression. Because I'd rather our world isn't such that my children or grand children will have to kill someone for a bottle of water, I have somewhat of an interest in economics relating to climate change, so that's the bent with which I took this task. An exceedingly brief summary of each paper is provided.

5.1 Nonparametric modeling of carbon prices, J. Chevallier (2011)[2].

Chevallier applies a standard local linear regression approach to the problem of carbon prices (for example, the EU ETS), with a multiplicatively separable and symmetric kernel and scalar bandwidth. He does this by using the model

$$y_t = \mu(x_{t-1}) + \sigma(x_{t-1})\varepsilon_i, \tag{25}$$

[1] Sadly, I think this means I've wasted my time and the FOC that I wrote down is incorrect. Perhaps over the Fall break I'll have a second crack at it because I'm curious as to whether we'd expect the Silverman to perform better than the cross validation - I imagine not given that the latter is so much more expensive.

where x_t consists of an appropriate number of lagged y values, y_t is the carbon price at time t, and ε_t is some shock. He contrasts this to the approach common in the literature which is to assume μ and σ are linear functions.

It is notable that this model is nonparametric both in the lack of assumptions on μ and σ, but also that the number of lags contained in x, denoted by m, is unknown. This is chosen as a tuning parameter along with (as is of course standard) the bandwidth of the kernel. The lack of constraints on m (barring a maximum value M) is interesting in this case since this affects the dimensionality of the estimator. To approach this, Chevallier creates an error function which one optimises over h **and** m. This error function (the "asymptotic final prediction error," as the paper is constructed with a view to forecasting) involves coefficients that depend on unknown densities, and hence must also be approximated nonparametrically.

The take home of the paper is that the nonparametric model offers greater forecasting and expository power than does the linear models standardly assumed in the literature.

5.2 The polarization of American environmental policy: A regression discontinuity analysis of Senate and House votes, 1971-2013, S.E. Kim & J. Urpelainen (2017)[6].

For all it's pros and cons, the two-party system leads to an environment suitable for regression discontinuity design, since the binary independent variable for the purpose of much analysis - the party of the legislator in a given constituency - is an indicator function:

$$\mathbb{1}\{V_p > 0.5\}, \tag{26}$$

where V_p denotes the share of votes for party p.

Kim and Urpelainen exploit this fact to use an RDD approach to examine whether party membership is an explainer of voting on environmental legislation independent of electoral conditions in the United States. They do this in two ways: a parametric approach that uses lags on electoral results on an assumed functional form for votes on evironmental bills (not of interest for our purposes), and a nonparametric RDD approach.

The key insight (taken from other authors) is that for close races - 'close' eventually being defined in the paper as within 0.5%, a value that is able to be so small due to the massive amount of observed votes (319,258) - the selection of a legislator is effectively random, and hence is a quasi-experimental environment. Intuitively, this is explained by the fact that in such seats, the decision is made by a handful of votes, and an individual voter's turnout is influenced by random variables - the example in the paper being the weather, though the argument is more convincing when one considers individual heterogeneity in voter turnout, such as how ill an individual may be feeling, the amount of fuel in the car and the price of fuel on the day, or any other such real-world concern. The authors provide evidence that districts that are marginal are often similar in terms of explanatory variables such as income, thus strengthening their "almost-random"

(their words) assumption. One should note - and they do - that a possible driver for elections in marginal seats is in fact heterogeneity in the candidates that is independent of political affiliation, such as their career history, but the authors control for this by excluding samples where there is a large discrepancy in the characteristics of the candidates in a given election. They have an absolutely massive data set which I guess makes the decision to do that a fairly easy one.

The ultimate conclusion of this paper is that the polarisation of party elites on environmental policy is precisely that, and not reflective of the degree of public polarisation. That is, all things equal, a Republican will vote to subsidise polluters to protect local industry, but will vote against subsidies to green energy because that's 'interfering with the market.' This is tested for a variety of subsamples differentiated by factors such as geography, and in all cases, the House is separated from the Senate, due to the perception of the Senate as acting in a less partisan manner, presumably due to less frequent electoral pressures.

5.3 Carbon dioxide emissions and governance: A nonparametric analysis for the G-20, G.E. Halkos, G.N. Tzemerez (2013)[4].

Halkos and Tzemerez present a somewhat rudimentary analysis that links World Bank indicators of governance (of which there are six, concentrating on different aspects of governance, such as citizen participation, rule of law, etc.) and carbon emissions for the G20 countries, from 1996 to 2010. This is done using a Nadaraya-Watson estimator for

$$g(x) = \mathbb{E}[y_i | X_i = x], \tag{27}$$

where x is the vector of governance indices and y_i is CO_2 emissions. The units are countries and year, which I would have thought threw up some fixed effects issues but those don't seem to be discussed.

I won't deny the similarity of the approach in this paper to the approach in the first paper I've offered, but I believe the differences are significant enough. For one, the first paper is more sophisticated in that it offers the nonparametric kernel treatment to functions in a stochastic process in discrete time rather than simply a regressor function, and also in that by incorporating uncertainty in the appropriate number of lags m, they extend the nonparametric analysis in a novel way, including the construction of a new process for selecting bandwidth. By contrast, this paper simply applies the NW estimator, and uses cross-validation to determine the bandwidth. However, to be fair to this paper, while the first paper offered uses a scalar bandwidth, this paper uses a vector bandwidth for multivariate kernel estimation (though it is just a diagonal matrix).

Looking at the graphs, it's not clear that they draw any useful conclusions. They observe that the regressions they obtain are highly nonlinear, but they are also not monotonic or really immediately recognisable as a function driving data as opposed to just noisy waves. No offence to Halkos and Tzemerez, but I'd be highly surprised if they ended up reading my assignment for a second year class anyway.

6 Question Six.

We run a simulation exercise on a binary choice model:

$$y_i = \mathbb{1}\{G(x_i'\theta_0) + u_i \geq 0\}, \tag{28}$$

where G is a linking function, and u is noise. For simplicity, we take G to be the identity mapping. This actually doesn't help our results however - it would seem I suck at coding. We consider a sample of 1000 individuals with five characteristics, the first two of which are discrete, the final three of which are continuous.

Not that I think it should matter, but, the first is distributed Poisson with parameter 3, the second is uniform on $\{-10, -9, \cdots, 9, 10\}$, the third is an exponential with parameter 3, the fourth is uniform on $[-10, 10]$, and the final is distributed as a standard normal.

Normal errors are distributed either homoscedastic with variance 1, and normal errors with homoscedastic variance have variance e^{x_5}. Non-normal errors are distributed as a fifty-fifty chance of being drawn from a normal with mean -1 and a normal with mean 1, where for the homoscedastic errors, these normals have variance 1, and for heteroscedastic errors, they have variance e^{x_5}. While this is built from normals, we feel that the bimodal nature of the distribution makes it suitably different from a normal for the purpose of adding errors.

We run each of these possibilities with $n = 100$ and $n = 1000$. We consider the probit estimator, the average derivative estimator, the Klein-Spade estimator, and the maximum score estimator. For the Klein-Spade estimator, we use a standard normal kernel with the Silverman bandwidth for $\mathbf{X}\theta$, where θ is the parameter of the objective. For the average derivative estimator, we use a product of standard normal kernels, and for the bandwidth, select the Silverman bandwidth fro $\mathbf{X}\hat{\theta}_{KS}$, where $\hat{\theta}_{KS}$ is the Klein-Spade estimator.

The squared error (the difference with the true value of θ_0, which is $(-16, -8.8, 1.87, 18.3, 1)'$) is reported for each simulation. I'll be honest, the main two take aways from this is that they all suck (I think computing a decent bandwidth is tricky but I'm not sure why probit is so lame when it's exactly the probit model), and that you should never do anything with kernels in Matlab, and if I ever program a nonparametric thing again, you can bet your bottom dollar that I'll be writing it in c++. The table below is the squared errors divided by 1000 (note then that 0s represent only that the error was less than 1), though that's a hell of a lot better than most.

	Klein-Spady	Probit	Maximum score	Average derivative
1	0.0233	0.0250	0.0261	0.0578
2	0.0137	0.0041	0.0214	0.5009
3	0	0	0.0261	0.0594
4	0	0	0.0214	1.0804
5	0	0	0.0261	0.0594
6	0	0	0.0214	0.6824
7	0	0	0.0261	0.0578
8	0	0	0.0214	0.8278

7 Question Seven.

7.1 Part (a).

We consider an RDD estimator using standard kernel regression estimation, so that the parameter is an estimate of the difference of the regression estimators either side of the treatment cutoff. Since the Nadaraya-Watson estimator is upward biased if the conditional expectation m is upward sloping at the boundary [9][1], we have that the difference would be exaggerated, and hence the RDD estimate is upward biased.

7.2 Part (b).

Beyond new kernel's and new estimators, the only change we can make is to the data, and hence we consider transformations of the data that will reduce the boundary bias.

Our intuition as to why the boundary bias occurs for the kernel distribution estimator is that the estimator "expects" there to be data to both sides of the boundary point, and finding only half as many points as it expects, and hence introduces bias. One method to deal with this could be to add points to the other side of the estimator to make up for this, with the most intuitive choice of points being a reflection of the existing data. In fact, such an approach is taken by Hall and Wehrly (1991), who find using a simulation study that an estimator that reflects data across boundaries and runs as a standard kernel estimator outperforms common boundary-corrected kernels, and performs approximately as well as estimators proposed by Gasser-Muller [5].

8 Question Eight.

We consider the partially linear model

$$y_i = x_i'\theta_0 + g(z_i) + u_i. \tag{29}$$

The Robinson estimator replaces the structural equation with

$$y_i - \bar{y} = (x_i - \bar{x})'\theta_0 + g(z_i) - \overline{g(z)} + u_i - \bar{u}, \tag{30}$$

and runs OLS on this. The other estimator, which shall hereafter be the Wren estimator (named after my cat, because it's late, and I need to give this thing a name), replaces the structural equation with

$$y_i - y_{i-1} = (x_i - x_{i-1})'\theta_0 + (g(z_i) - g(z_{i-1})) + (u_i - u_{i-1}), \tag{31}$$

where observations are ordered by z_i, and runs OLS on this. It can be shown that the Robinson estimator, $\hat{\theta}_R$ is

$$\hat{\theta}_R = \theta_0 + (\Delta\mathbf{X}'\Delta\mathbf{X})^{-1}\Delta\mathbf{X}\Delta g(\mathbf{Z}) + (\Delta\mathbf{X}'\Delta\mathbf{X})^{-1}\Delta\mathbf{X}'\mathbf{u}, \tag{32}$$

where

$$\Delta\mathbf{X} = \mathbf{X} - \bar{\mathbf{X}}\mathbf{1}, \tag{33}$$

where $\mathbf{1}$ is a vector of 1s; and that the Wren estimator $\hat{\theta}_W$ is

$$\hat{\theta}_W = \theta_0 + (\mathbf{X}_{-1}'\mathbf{X}_{-1})^{-1}\mathbf{X}_{-1}g(\mathbf{Z})_{-1} + (\mathbf{X}_{-1}'\mathbf{X}_{-1})^{-1}\mathbf{X}_{-1}\mathbf{u}, \tag{34}$$

where subscript -1 denotes a lagged value.

Note that we can consider Δ to be a matrix

$$\mathbf{I} - \frac{1}{n}\mathbf{1}_{n\times n}, \tag{35}$$

so that

$$\Delta\mathbf{X} = \Delta\mathbf{X}. \tag{36}$$

Similarly, if we write

$$\mathbf{L} = \mathbf{I} - \mathbf{I}_{-1}, \tag{37}$$

where $\mathbf{I}_{-1}$ is the identity matrix with all entries shuffled 1 to the left, we can write

$$X_{-1} = \mathbf{L}\mathbf{X}. \tag{38}$$

The intuition of the success behind each of these estimators is that in the structural equations, the bigger n is, the smaller the differenced $g(z)$ term is. For the Wren estimator, we see that as n grows larger, since the observations are ordered by z, the closer each z_i is to its neighbour, and hence, for sufficiently large n and not awful g, the closer each $g(z_i)$ is to its neighbour, and hence the lower the error added to the estimator. It is however very hard to determine as a function of n what the average distance between neighbour points is, or maybe I just don't know enough about order statistics, but trying to integrate directly from the joint pdf of order statistics the expectation of $|z_i - z_{i-1}|$ is very, very hard. I am going to point out this note by Lopez (2011), which indicates the problem is certainly not a trivial one, as work in 2011 was still being done to find upper bounds on expectations of differences between order statistics even given strict assumptions regarding the support of the distribution [8].

We also note here, while we're discussing the errors of $\hat{\theta}_W$ that the errors in the structural equation inherit the iid property from the underlying data, and hence is simply white noise.

It is not actually clear to me why the second term in $\hat{\theta}_R$ would disappear now that I'm looking at it, except perhaps if there are hopes of orthogonality between x and z.

9 Question Nine.

We are asked to consider the advantages of the median absolute deviation as a measure of the success of an estimator in simulation studies over the root mean squared error. The definition of the median absolute deviation (hereafter MAD) is defined[7] as

$$\text{MAD} = \text{median}\left(x_i - \text{median}(x_i)\right). \tag{39}$$

That is, it is the median of the distribution of the data minus the median of the data. The first thing to note of this estimator is that the population analogue of this exists for all distributions (since every distribution has a median), which allows an economist to choose simulation data from a wider variety of distributions, no longer restricted to moment concerns.

Perhaps more importantly is the robustness of the measure to outliers. This is obvious intuitively, since individual observations cannot effect the median, which enters the formula twice (offering two layers of 'protection' from rogue observations). If we define the notion of a breakdown point as the number of observations that are allowed to go to infinity before a statistic describing spread goes to infinity, it is clear that the root mean square error has a breakdown value of 0[7], whereas the median has a breakdown value of $\frac{n}{2} - 1$, which is a significant improvement, and indicates much greater robustness to outliers, which are often not of interest in simulation exercises.

10 Question Ten.

We consider x_i drawn i.i.d. from $N(\theta_0, 1)$ with $\theta_0 \in \{0, 1, \cdots, 10000\}$, and consider the rates of convergence of the sample mean and the MLE. We presume that the intention of this is for the sample mean to be simply the sample mean (i.e. so that $\bar{\theta} \notin \Theta$), whereas the MLE is the maximiser of the log likelihood over $\Theta = \{0, 1, \cdots, 10000\}$. Note that since in a continuous parameter space we have

$$\hat{\theta}_{MLE} = \bar{\theta}, \tag{40}$$

this is the only way to consider them as separate estimators, and because the log likelihood for a normal distribution of known variance is nicely behaved, we have that for the discrete (in fact finite, which is important in the paper I'll reference) parameter space, that

$$\hat{\theta}_{MLE} = [\bar{\theta}], \tag{41}$$

where $[\cdot]$ is the rounding operator which maps $\cdot$ to the nearest integer.

The rate of convergence of the MLE comes from the argument from Choirat et al (2012) which demonstrates for m-estimators in general - with MLE given as a particular example - have convergence rate

$$\frac{e^{-cn}}{n^{\frac{1}{2}}}, \tag{42}$$

where c is some positive constant expression of parameters introduced in the proof [3]. The convergence rate of the sample mean is, as is standard since we are calculating it as though the parameter space is continuous, $\sqrt{n}$.

We also conduct a simulation exercise, generating 1000 data points from $N(326, 1)$ (the number of seats technically needed for a majority in the House of Commons, excluding seats held by the Speaker and their Deputies, and seats formally held by abstaining Sinn Fein members) and compare the sample average (continuous) to the MLE (discrete) by plotting the estimates against number of observations to observe how much better the MLE performs. We do this rather than report the root mean squared error of multiple simulations as we expect that for any reasonable n, the value of $\hat{\theta}_{MLE}$ will be θ_0 and return a root mean squared value of 0 (of course indicative of greater performance, but not particularly worth running simulations for). Since the variance is small compared to the gap between possible θs in the question given, we increase the variance to 100 for the purpose of our simulation, and to 10.

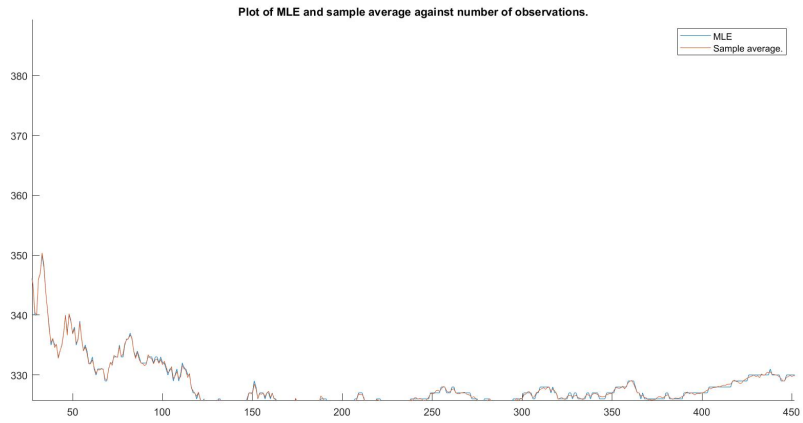

Figure 12: The estimators over a run of 1000 data points with variance 100.

For variance equal to 100, we find that the estimators in fact track each other very highly - the benefit of rounding the sample average is low when the variance disturbs the data enough. However, decreasing the variance to 10 indicates the advantages of the MLE, and decreasing to 1 shows immediate benefits to the estimator, indicating that if gaps between points in the parameter space are on the order of the variance, this method should be preferred. The simulation

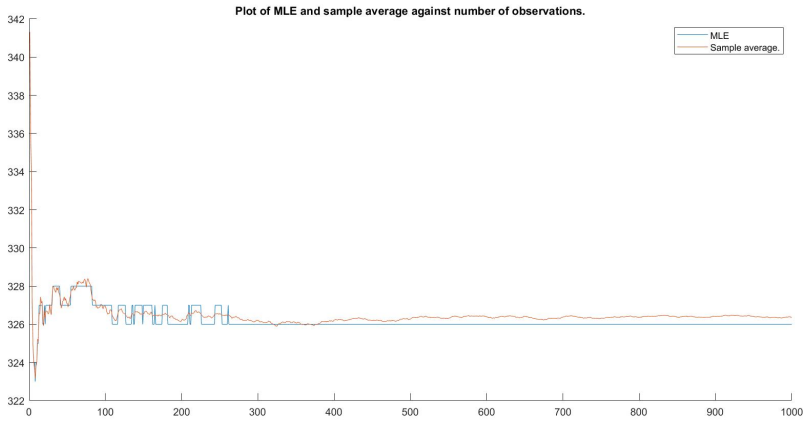

Figure 13: The estimators over a run of 10000 data points with variance 10.

exercise would seem to to suggest that the coefficient c in equation (42) is highly sensitive to the ratio of the variance and the spacing within the parameter space, as for higher variances, it's not at all clear that the performance is outperforming the sample average.

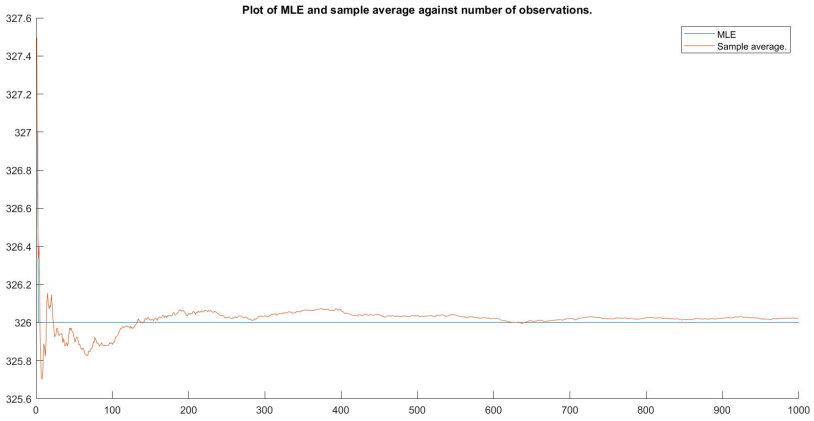

Figure 14: The estimators over a run of 1000 data points with variance 1.

References

[1] Anon, *Lecture 12 nonparametric regression.* https://www.bauer.uh.edu/rsusmel/phd/ec1-27.pdf.

[2] J. Chevallier, *Nonparametric modeling of carbon prices*, Energy Economics, 33 (2011), pp. 1267–1282.

[3] C. Choirat, R. Seri, et al., *Estimation in discrete parameter models*, Statistical Science, 27 (2012), pp. 278–293.

[4] G. E. Halkos and N. G. Tzeremes, *Carbon dioxide emissions and governance: A nonparametric analysis for the g-20*, Energy Economics, 40 (2013), pp. 110–118.

[5] P. Hall and T. E. Wehrly, *A geometrical method for removing edge effects from kernel-type nonparametric regression estimators*, Journal of the American Statistical Association, 86 (1991), pp. 665–672.

[6] S. E. Kim and J. Urpelainen, *The polarization of american environmental policy: A regression discontinuity analysis of senate and house votes, 1971–2013*, Review of Policy Research, 34 (2017), pp. 456–484.

[7] C. Leys, C. Ley, O. Klein, P. Bernard, and L. Licata, *Detecting outliers: Do not use standard deviation around the mean, use absolute deviation around the median*, Journal of Experimental Social Psychology, 49 (2013), pp. 764–766.

[8] M. Lopez and J. Marengo, *An upper bound for the expected difference between order statistics*, Mathematics Magazine, 84 (2011), pp. 365–369.

[9] J. Racine, *Bias-corrected kernel regression*, Journal of Quantitative Economics, 17 (2001), pp. 25–42.

[10] I. Trapp and A. Francis, *Applications of non-parametric kernel smoothing estimators in monte carlo risk assessments*, (2012).