# Efficient estimation in regression discontinuity designs via asymmetric kernels

Fe, Eduardo

24 February 2012

# Efficient estimation in regression discontinuity designs via asymmetric kernels*

Eduardo Fé †
*University of Manchester*

February 24, 2012

**Abstract**

Estimation of causal effects in regression discontinuity designs relies on a local Wald estimator whose components are estimated via local linear regressions centred at an specific point in the range of a treatment assignment variable. The asymptotic distribution of the estimator depends on the specific choice of kernel used in these nonparametric regressions, with some popular kernels causing a notable loss of efficiency. This article presents the asymptotic distribution of the local Wald estimator when a gamma kernel is used in each local linear regression. The resulting statistics is easy to implement, consistent at the usual nonparametric rate, maintains its asymptotic normal distribution, but its bias and variance do not depend on kernel-related constants and, as a result, is becomes a more efficient method. The efficiency gains are measured via a limited Monte Carlo experiment, and the new method is used in a substantive application.

**Key Words**: Regression Discontinuity, Asymmetric Kernels, Local Linear Regression.

**JEL Classification**: *C13, C14, C21.*

*An earlier version of this paper was titled "An application of Local Linear Regression with Asymmetric Kernels to Regression Discontinuity designs"

†Health Economics. Room 4.305, Jean McFarlane Building. University of Manchester. Manchester, M13 9PL, UK. email:e.fe-rodriguez@manchester.ac.uk.

# 1 Introduction.

This article explores the use of asymmetric kernels when estimating treatment effects in a Regression Discontinuity (RD) design. RD (Thistlethwaite and Campbell, 1960) is a program evaluation technique applicable when the probability distribution describing assignment to treatment has a discontinuity at a known cut-off level of certain continuous *running* or *assignment* variable. Hahn et al. (1999) and Lee (2008) show that, under certain weak regularity conditions, a comparison of the conditional means of outcomes between recipients and non-recipients located in a neighbourhood of the cut-off level has a causal interpretation. Identification holds even when treatment effects vary across individuals or there is selection into treatment (due to, for instance, anticipated gains), provided that individuals cannot perfectly control their score of the running variable (Lee, 2008). RD designs seem to occur frequently in real life and, given that estimation in this setting is straightforward[1], the technique has become very popular among researchers. Representative applications are in van der Klaauw (1996), Angrist and Lavy (1999) Ludwig and Miller (2007), Lee (2008), Battistin et al. (2009), Almond et al. (2010) or Almond and Doyle (2011) to mention but a few.

In RD designs, the cut-off point splits the region of estimation in two (or more) intervals with known bounds. It is at these bounds where the treatment effect is identifiable. Therefore valid inference of the causal effect requires two ingredients: firstly, a consistent, distribution-free estimator of the conditional mean of the outcome and assignment variables in a neighbourhood of the cut-off points and, secondly, estimators with good boundary properties. To reduce the likelihood of model misspecification, Hahn et al. (1999) and Porter (2003) proposed the use of nonparametric regression methods. Among these estimators, Local Linear Regression (LLR) with symmetric kernels has been widely used. This estimator is consistent, asymptotically normal distributed and

[1] Researchers only need to compute a local version of Wald's estimator (Wald, 1940) centred at the point where the discontinuity in the likelihood of treatment takes place.

it circumvents the so called boundary bias problem: its bias vanishes at the boundary and interior of the regions of estimation at similar rates[2].

The boundary bias is an asymptotic problem. In finite samples, symmetric kernels typically used in LLR still have part of the kernel window devoid of data, leading to the introduction of a small sample bias. Kernels with a bounded support (such as uniform, triangular or Epanechnikov kernels) are unlikely to ameliorate the problem because the LLR estimator with a compact-support kernel has unbounded unconditional variance in finite samples (Seifert and Gasser, 1996), what suggests that these weighting functions can lead to fairly imprecise estimates.

In RD designs the small sample bias translates into kernel-dependent constants affecting, not only the expression of the small sample bias, but also the (asymptotic) variance of the local Wald estimator (Wald, 1940) of the causal effect of interest. Given that this estimator is consistent, researchers need not give any special consideration to the exact expression of its bias. However, the kernel-dependent constant affecting the asymptotic variance is greater than one[3] for the most popular symmetric kernels, reflecting the loss of precision implied by using kernels that do not match the area of integration.

In a series of papers, Brown and Chen (1999) and Chen (2000, 2001) study kernel selection when estimating conditional moments over a bounded support (see also Scaillet, 2004). These authors note that gamma (and beta) densities can be used as kernels. Doing this one can match the support of the regression function to the support of the kernel what ensures, not only that all weight is allocated within the boundaries of the support (thus eliminating small sample biases), but also that the whole sample is effectively used by the estimator, thus increasing the stability of the statistic. Unlike LLR with kernels of bounded support, the LLR estimator with beta and gamma kernels has a bounded

[2]In contrast, the bias of the popular Nadaraya-Watson estimator disappears at a slower rate at the boundary and so larger sample sizes are required in this region to obtain a given level of accuracy (Fan, 1993, Wand and Jones, 1994 or Ruppert and Wand, 1994).

[3]When using a Gaussian kernel that constant takes a value equal to 1.78, but when using a uniform or Epanechnikov kernel, that constant equals or exceeds 4.

finite sample variance. At the boundary, the rate of convergence to the true regression equals to that exhibited by standard kernel estimators. In addition to this, gamma and beta kernels are parameterised so that their shape and scale depends on a smoothing and on the point at which estimation is taking place, thus allowing a sort of adaptive local smoothing. Chen (2001) shows through simulations that, at the boundary, these new methods have biases at least comparable to those obtained with a Gaussian kernels while the variance is substantially smaller across the domain of the curve and so the new estimators are preferred in accordance to a mean square error criterion.

Given the boundary, support and locally-adaptive properties of gamma and beta kernels, this article explores theoretical and empirical consequences of replacing traditional kernels by asymmetric densities. The focus of this paper is on the gamma distribution[4]. Section 2 presents the asymptotic distribution of the local Wald estimator in Hahn et al. (1999) when LLR with a gamma kernel is used to estimate each of its components. The asymptotic distribution of the statistic is particularly simple and the asymptotic variance does not depend on kernel-related constants, what brings an efficiency gain with respect to estimators based on traditional kernels. The asymptotic mean square error of the estimator is used to obtain the optimal value of the bandwidth parameter, which tuns out to exhibit a rate of convergence to zero identical to that derived from LLR with symmetric kernels. However, the optimal bandwidth is not directly comparable to those suggested by other (symmetric) kernels, and for this reason bandwidth selection must be done carefully in order to be able to draw comparisons among the different methods discussed in this article. We suggest the use of a plug-in method firstly introduced by Ruppert et al. (1995). This section also extends the results in Chen (2001), by establishing the asymptotic normality of the LLR with gamma kernel, although given the type

[4]The working paper version of this article treats the Beta case as well. It turns out that the asymptotic distribution of the estimator thus obtained is identical to the gamma-based estimator discussed here. However, the use of beta kernels is bound to the selection of two bandwidths: one to select the data included in estimation and another to allocate weights to the data. This makes this choice of kernel less attractive.

of problem under study, our results specialise to the case of estimation at the boundary, however they can be modified in a straightforward fashion to obtain the asymptotic normality of the estimator elsewhere in the domain or the regression function[5]. Our claims of efficiency gains are confirmed in section 3, where we study the small sample performance of the new estimator via a Monte Carlo experiment. Section 4 revisits a study by Lee et al. (2004) who try to identify whether voters elect or affect policy. This is a substantive question that sits at the core of Political Economy. Once again, our results emphasise the advantages of the gamma kernel when estimating treatment effects in RD designs. Section 5 concludes.

## 2 Estimation in a Regression Discontinuity Design.

A researcher has a sample of $i = 1, \ldots, n$ independent observations $(Y_i, Z_i, X_i)_{i=1}^n$ where $X_i$ is a binary indicator of treatment such that $X_i = 1$ if the individual has received treatment, while $X_i = 0$ otherwise. Treatment allocation is a function of $Z$, the *running* variable, so that $P(X_i = 1|Z_i = z_i)$ is discontinuous at $z_o$, a known threshold or cut-off point in the range of $Z$. A particular case of this setting (the so called Sharp RD design) happens when the probability distribution of $X_i$ is degenerated at $z_o$ so that $P(X_i = 1|Z_i)$ jumps from 0 to 1 at the threshold. Here the focus is on the general case

$$\lim_{z \to z_o^+} E(X_i = 1|Z_i = z) \neq \lim_{z \to z_o^-} E(X_i = 1|Z_i = z).$$

The goals is to estimate the causal effect of $X$ on the outcome variable, $Y_i \in \mathbb{R}$.

Each individual in the sample has two potential responses to treatment status: $Y_{i1}$ if the individual receives treatment and $Y_{i0}$ otherwise. Both outcomes are not observed

[5]Note, however that the rate of convergence of the estimator in the interior is likely to differ from that obtained at the boundary (see Chen, 2001).

simultaneously. Instead one observes

$$Y_i = Y_{i0} + X_i(Y_{i1} - Y_{i0}) = \alpha_i + X_i\beta_i$$

where $\beta_i = (Y_{i1} - Y_{i0})$ captures the effect of the intervention. Hahn et al. (1999) and Hahn et al. (2001) show that

$$\tau = \frac{m^+ - m^-}{p^+ - p^-} \tag{2.1}$$

where

$$m^+ = m^+(z_o) = \lim_{z \to z_o^+} E(Y_i|Z_i = z) \quad m^- = m^-(z_o) = \lim_{z \to z_o^-} E(Y_i|Z_i = z) \tag{2.2}$$

$$p^+ = p^+(z_o) = \lim_{z \to z_o^+} E(X_i|Z_i = z) \quad p^- = p^+(z_o) = \lim_{z \to z_o^-} E(X_i|Z_i = z) \tag{2.3}$$

identifies a causal effect, although the interpretation of $\tau$ depends on what assumptions one is willing to include in estimation. For example, if $E(\alpha_i|Z = z)$ is assumed to be continuous at $z_o$ and $\beta_i = \beta$ for all $i$, $\tau$ identifies $\beta$, while under the additional assumptions that $E(\beta_i|Z = z_o)$ is continuous at $z_o$ and $X$ is independent of $\beta_i$ given $Z$ near $z_o$, then $\tau = E(\beta_i|Z = z_o)$.

Calculation of the $\tau$ requires estimates of the limits $m^+$, $m^-$, $p^+$ and $p^-$. Local Linear Regression (LLR) has become the staple estimator in empirical work and therefore $m^+(z_o)$ is the value of $a$ solving the weighted least squares problem[6],

$$\arg\min_{a,b} n^{-1} \sum_{i=1}^{n} \left( Y_i - a - b(Z_i - z_o) K_h \left( \frac{Z_i - z_o}{h} \right) \right)^2 \mathbb{I}(Z_i > z_o) \tag{2.4}$$

where $K_h = h^{-1}K(.)$ is a kernel function distributing weights across the sample points, $h = h(n)$ is a bandwidth parameter regulating the width of the kernel and such that $h \to 0$ as $n \to \infty$, and $\mathbb{I}(.)$ is an indicator function taking the value 1 when the condition

[6]The estimators of $m^-, p^+$ and $p^-$ are defined similarly.

inside the brackets is true. Estimates of $m^+$ and the remaining conditional moments can be combined in the following *local* Wald estimator of $\tau$,

$$\hat{\tau} = \frac{\hat{m}^+ - \hat{m}^-}{\hat{p}^+ - \hat{p}^-}. \tag{2.5}$$

Under the certain regularity conditions, Hahn et al. (1999) show that, $\hat{\tau} \sim N(\lambda, \varphi)$, where

$$\begin{aligned}
\lambda &= \frac{\varrho}{2}\left(\frac{1}{p^+ - p^-}\left(m''^+\omega^+ - m''^-\omega^-\right) - \frac{\hat{m}^+ - \hat{m}^-}{(p^+ - p^-)^2}(p^+\omega^+ - p^-\omega^-)\right) && (2.6)\\
\varphi &= \frac{1}{f}\left(\frac{1}{(p^+ - p^-)^2}(\sigma_+^2 k^+ - \sigma_-^2 k^-)\right.\\
&- 2\frac{\hat{m}^+ - \hat{m}^-}{(p^+ - p^-)^3}(\eta^+ k^+ - \eta^- k^-)\\
&+ \left.\frac{\hat{m}^+ - \hat{m}^-}{(p^+ - p^-)^4}(p^+(1-p^+)k^+ - p^-(1-p^-)k^-)\right) && (2.7)
\end{aligned}$$

Here $\varrho = \lim_{n\to\infty} h^2\sqrt{nh}$ and[7],

$$\begin{aligned}
\omega^+ &= \frac{\left(\int_0^\infty u^2K(u)du\right)^2 - \int_0^\infty uK(u)du\int_0^\infty u^3K(u)du}{\int_0^\infty u^2K(u)du\int_0^\infty K(u)du - \left(\int_0^\infty uK(u)du\right)^2} && (2.8)\\
k^+ &= \frac{\int_0^\infty\left(\int_0^\infty u^2K(u)du - s\int_0^\infty uK(u)du\right)^2 K^2(s)ds}{\left(\int_0^\infty u^2K(u)du\int_0^\infty K(u)du - \left(\int_0^\infty uK(u)du\right)^2\right)^2}. && (2.9)
\end{aligned}$$

In practice, most applications use symmetric kernels, so that $\omega^+ = \omega^-$, $k^+ = k^-$, thus simplifying the expression of $\lambda$ and $\varphi$.

The constants $\omega$ and $k$ play an important role in the discussion that follows. They form part of the asymptotic distribution of $\hat{\tau}$ because the kernels used in LLR place part of their weight beyond the cut-off point, where the sample is devoid of data. This is the *small sample* implication of the so called boundary bias problem (Wand and Jones, 1994 or Ruppert and Wand, 1994). Clearly, $\omega$ is less troublesome than $k$ because one can

[7]The constants $\omega^-, k^-$ are defined similarly but the limits of integration are from $-\infty$ to 0.

| | Type of Kernel | | |
|---|---|---|---|
| | Gaussian | Uniform | Epanechnikov |
| $\omega$ | -0.75193 | -0.16 | -0.115789 |
| $k$ | 1.785961 | 4.0 | 4.497981 |

Table 1: Contants $\omega$ and $k$ for popular symmetric kernels.

mitigate the effect of the bias term by choosing $h$ appropriately. However the effect of $k$ is permanent, and can be considerable. In table 1 we collect the value of these constants for the three most popular kernels used in empirical applications (uniform, Gaussian and Epanechnikov). For the selected kernels, $-1 < \omega < 0$, so that it operates by reducing the magnitude of the bias term. On the contrary, $k > 1$ and so it inflates the asymptotic variance of the estimator. In the particular case of uniform and Epanechnikov kernels $k$ is a large as 4, thus reducing the attractiveness of these choices of kernel. The question is whether one can find a kernel such that $k$ can be reduced so as to provide a more efficient estimator.

### 2.1 The Gamma Kernel

Consider now the solution of (2.4) but where $K$ is replaced by the following gamma kernel

$$K^G_{z_o,b}(Z_i) = \frac{Z_i^{z_o/b} e^{-Z_i/b}}{b^{\frac{z_o}{b}+1}\Gamma\left(\frac{z_o}{b}+1\right)}. \tag{2.10}$$

In the above expression $\Gamma(.)$ is the gamma function, and $b = b(n) > 0$ is a smoothing parameter satisfying $b \to 0$ as $n \to \infty$. Here $K^G(.)$ is a gamma density with parameters $r = b$ and $s = z_o/b + 1$. This parameterisation places the mode of the density at $z_o$, while the shape of (2.10) depends on the values of the parameters $r$ and $s$. These are ultimately determined by $z_o$ and the smoothing parameter $b$. Therefore, for fixed $b$, the gamma kernel allows a type of adaptive smoothing in nonparametric regression. Note that, although the domain of the gamma kernel is $[0, \infty)$, in practice the regions generated by the cut-off point in a RD design will rarely equal this interval, however, since the cut-off values are known to the researcher, it is always possible to map the different regions onto the interval $[0, \infty)$ (leaving $y$ unaltered).

To proceed with the analysis, we need to introduce new definitions, properties and assumptions. Following Chen (2001), the point $z$ in the support of $z_i$ is a boundary point if:

1. $z/b \to \kappa$, when using gamma kernel

for some $\kappa \geq 0$. The following properties of the gamma kernel will be necessary to obtain the asymptotic distribution of the estimator of $\tau$

**Property 2.1** *(Chen, 2001). Let $\xi$ be a random variable and let $z$ be a point in the range of $Z_i$. Define $p_l(z) = E(\xi - z)^l$. If $\xi$ has the density (2.10), $p_2(z) = bz + 2b^2$ and $p_l(z) = O(b^2)$ for $l \geq 3$. In particular, if $z$ is a boundary point, $p_2(z) = b^2(2 + \kappa)$. Furthermore, $K^G_{z,b}(Z_i)^2 = A_b(z)K_{z\prime,b'}(Z_i)$. If $z_o$ is a boundary point, then $A_b(z) = b^{-1}\Gamma(2\kappa + 1)/2^{2\kappa+1}\Gamma^2(\kappa + 1) + o(b^{-1})$.*

Finally, in order to introduce the main result of the paper, the following set of assumptions is needed.

**Assumption 2.1** *Let $\mathcal{V}$, be a neighbourhood of $z_o$ and $0 < M < \infty$,*

1. *The running variable, $z$, has marginal density $f(z)$, bounded, positive and twice continuously differentiable in $\mathcal{V}$.*

2. *The conditional expectations $m(z)$, $p(z)$, $\sigma^2(z) = var(Y|Z = z)$ and $\eta(z) = cov(Y, X|Z = z)$ are twice continuously differentiable at $\mathcal{V} \setminus z_o$. Their left limits and first and second derivatives exist and are uniformly bounded on $[z_o - M, z_o)$,*

*and similarly, the right limits of $m(z)$, $p(z)$ and its first and second derivatives exist and are uniformly bounded on $(z_o, z_o + M]$.*

3. *$m(Z)$, $\sigma^2(Z)$ and $\eta(Z)$ satisfy the following one-sided expansion,*

$$\zeta(Z) = g(Z) - g^+(z) - g'^+(z) - \frac{1}{2}m''^+(z)(Z-z)^2$$

*where $\sup_{z<Z<z+Mb} |\zeta(Z)| = o(b^2)$, for $0 < M < \infty$ and where $g(.)$ stands for $m$, $\sigma^2$ or $\eta$.*

4. *$E\left[(Y - m(Z))^{2+\delta}|Z = z\right]$ and $E\left[(X - p(Z))^{2+\delta}|Z = z\right]$ are uniformly bounded on $\mathcal{V}$, for $\delta > 0$.*

5. *$b \to 0$, $nb \to \infty$ as $n \to \infty$.*

The conditions above are standard in the literature (Hahn et al., 1999; Porter, 2003). Assumption 2.1.4 ensures the satisfaction of a suitable Lyapounov condition, and is weaker than the condition found in Hahn et al. (1999). Unlike in other sources continuous differentiability of second order is imposed on $\eta$, $\sigma^2$ and $m$ (through assumption 2.1.3 above), in order to approximate these quantities on a neighbourhood of the cut-off point. The results in Hahn et al. (1999) only require the dominated convergence theorem.

**Lemma 2.1** *Let $\hat{m}^+(z_o)$ be a LLR estimator of $m^+(z_o)$ with the gamma kernel defined in equation (2.10), let the point $z_o$ be the boundary point of the support of the regression function and assume that $b^2\sqrt{nb} \to \varrho > 0$. Then, under assumption 2.1,*

$$\sqrt{nb}\left\{\hat{m}^+(z_o) - m^+(z_o)\right\} - \varrho m''^+(z_o) \sim N\left(0, \frac{\sigma^{2+}(z_o)}{f(z_o)}\right). \tag{2.11}$$

The proof is given in the Appendix. The expression of the asymptotic bias and variance is particularly simple, because the results are circumscribed to an specific kernel at the boundary point.

From equation (2.11) one can easily derive the expression of the mean square error of this estimator. The bandwidth parameter that minimises the mean square error of the estimator at the boundary is:

$$b^{*+} = \left\{ \frac{\sigma^{+2}(z_o)}{f(z_o)m''^{+2}(z_o)} \right\}^{1/5} n^{-1/5} = Cn^{-1/5} \tag{2.12}$$

The optimal smoothing parameter converges to 0 at the same rate than $h$ in the standard nonparametric regression with symmetric kernels. The lemma also establishes that, as in the case of LLR with symmetric kernels, the bias of the estimator depends on the curvature of the (limit) of the regression function in a neighbourhood of the boundary point. If $m(.)$ is linear in this neighbourhood, the estimator will be unbiased and the optimal bandwidth will tend to $\infty$; as the complexity of the design increases and the bias increases, the optimal value of the bandwidth gets progressively smaller. However, the magnitude of the constant $C$ will in general be distinct to that accompanying the theoretically optimal value of $h$ in equation (2.4), so that estimations based on equal $h$ and $b$ are not directly comparable. In particular, for symmetric kernels, the bandwidth that minimises the mean square error at the boundary equals

$$h^{*} = \left\{ k \frac{\sigma^{+2}(z_o)}{\omega^{2}(z_o)m''^{+2}(z_o)} \right\}^{1/5} n^{-1/5} = Cn^{-1/5} \tag{2.13}$$

where $\omega$ and $k$ were defined in equations (2.8) and (2.9) respectively. In particular, for the Gaussian kernel, $k/\omega^2 = 2.37516923$, while for the uniform kernel $k/\omega = 144$. Thus $h^*$ is largest for the uniform kernel, while for the Gaussian kernel, $h^* > b^*$ but both quantities will be relatively close. When $f(z_o) = 0.38$ (approximately the value of the Gaussian density at its mode), $m''(z_o) = 1$, $\sigma^2(z_o) = 1$, and $n = 1000$, $h^*_{Gaussian} = 0.342079683$, $h^*_{Uniform} = 0.82359894$ and $b^* = 0.304819875$.

Finally, since $b^*$ depends on unknown moments, the usual problems associated to bandwidth selection apply here as well (see, for example, Hart, 1997). Ludwig and Miller

(2007) and Imbens and Lemieux (2008) discuss several data-driven methods devised with a RD design in mind, all of which are applicable here. In our simulations below we will suggest the use of the plug-in method discussed in Ruppert et al. (1995)[8], as it is simple to compute, has been showed to have good coverage properties and is more stable than bandwidths calculated via standard cross-validation methods. The method requires the following steps:

- Calculate $\hat{m}'' = 2\mathbf{e}_2' \left(\mathbf{R}'\mathbf{R}\right)^{-1} \mathbf{R}'\mathbf{Y}$ where $\mathbf{R}$ has rows $\mathbf{r} = (1, (Z - z_o), ..., (Z - z_o)^4)$ and $\mathbf{e}_2$ is a vector with a one in the second row and zeros elsewhere.

- $\hat{\sigma}^2 = \sum_{i=1}^{n}(Y_i - \tilde{m})^2/(n-5)$, where $\tilde{m}$ is the least squares fit for the the above regression.

- $f(z_o)$ can be approximated with a kernel density estimator. In this paper, for simplicity, we used a Gaussian kernel. Note that because $Z$ is assumed continuous, $f(z_o)$ is continuous, so there is not boundary bias problem here. The bandwidth in this stage equals Silverman's rule of thumb Silverman (1986), $h = 1.06 * \hat{\sigma_z} n^{-1/5}$, where $\hat{\sigma}_z$ is $Z$'s sample standard deviation.

With assumptions 1 and 2 in place, it is now possible to establish the main result of the paper.

**Theorem 1** *Consider the estimator (2.4) at $z_o$ with the kernel (2.10). Under assumption 2.1 and with $b^2\sqrt{nb} \to \varrho$, $0 \leq \varrho < \infty$,*

$$\sqrt{nb}\left(\frac{\hat{m}^+ - \hat{m}^-}{\hat{p}^+ - \hat{p}^-} - \frac{m^+ - m^-}{p^+ - p^-}\right) \to \mathcal{N}\left(\lambda, \varphi\right) \tag{2.14}$$

*with,*

$$\lambda \;=\; \varrho\left(\frac{1}{p^+ - p^-}(m''_+ - m''_-) - \frac{m^+ - m^-}{(p^+ - p^-)^2}(p''_+ - p''_-)\right) \tag{2.15}$$

[8]This method has also been suggested by Frandsen et al. (2011) for empirical applications in discontinuity designs.

*and*

$$
\begin{aligned}
\varphi &= \frac{1}{f}\left(\frac{1}{(p^+-p^-)^2}(\sigma_+^2+\sigma_-^2)\right. \\
&- 2\frac{m^+-m^-}{(p^+-p^-)^3}(\eta^+ + \eta^-) \\
&+ \left.\frac{(m^+-m^-)^2}{(p^+-p^-)^4}(p^+(1-p^+)+p^-(1-p^-))\right) \qquad (2.16)
\end{aligned}
$$

*where $f$ and all the limits of the conditional expectations are evaluated at the cut-off point $z_o = 0$.*

The above theorem allows us to establish some predictions regarding the behaviour of the estimator of $\tau$ that uses the gamma kernel. Consider expressions (2.6) and (2.7) for the cases of Gaussian and Epanechnikov kernels (so that $\omega^+ = \omega^-$ and $k^+ = k^-$). Assume, further, that the limits of conditional moments and second derivatives were known. Then, (2.15) and (2.16) coincide with (2.6) and (2.7) a up to a kernel-dependent constant. The bias and variance of the estimator that uses symmetric kernels is of the form $Bias = \varrho\omega C/2$, $var = kC'/f$ for constants $C, C'$, while for the estimator that uses asymmetric kernels, $Bias = \varrho C$, $var = C'/f$. Gaussian, Epanechnikov and uniform kernels have $\omega < 1$ but $k > 1.7$. This implies that the estimator that uses the gamma kernel will exhibit a larger bias but a much smaller variance than estimators using any of the other three symmetric kernels. In view of this results, we conclude that using a gamma kernel improves the efficiency of the local Wald estimator of causal effects in a RD design.

Note, finally, that the gamma kernel is particularly simple to implement in our setting because the boundary point of interest (after transformation of the running variable) is 0, so that the gamma kernel in (2.10) reduces to an exponential distribution,

$$K^G_{0,b}(Z_i) = \frac{e^{-Z_i/b}}{b}$$

Then, each of the four LLR estimations can be performed via weighted least squares with $K^G_{0,b}$ as weights.

## 3 Small Sample Performance.

This section explores the extent to which the theoretical advantages brought about by the gamma kernel translate into small sample improvements. For this task we borrowed the Monte Carlo setting described in Feir et al. (2011)[9]. Samples of $n = 1000$ observations were drawn from the following data generating process $R = 20.000$ times,

$$Y_i = Y_{i0} + X_i\beta \tag{3.1}$$

$$X_i = \begin{cases} \mathbb{I}(u_i < 0) & \text{if } Z_i \leq 0 \\ \mathbb{I}(u_i < c) & \text{if } Z_i > 0 \end{cases} \tag{3.2}$$

where

$$\begin{pmatrix} Y_{i0} \\ u_i \end{pmatrix} \sim N\left(\mathbf{0}, \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}\right). \tag{3.3}$$

The running variable was allocated a standard normal distribution, $Z_i \sim N(0, 1)$. The parameter $\rho$ controls the correlation between the outcome and assignment variables, and was set to $\rho = 0.5$ (moderate endogeneity) and $\rho = 0.9$, high endogeneity. $\beta$ is the parameter of interest, and was set to 0 (not causal effect) and 2. The parameter $c$ controls the magnitude of the discontinuity in the distribution of treatment. Too small values of $c$ imply a small discontinuity in the distribution of treatment, in which case the instrument $\mathbb{I}(Z_i \geq z_o)$ becomes a poor predictor of allocation into treatment. Therefore, small $c$ would imply a problem of weak instruments (Bound et al., 1995, Staiger and

[9] Feir et al. (2011) were originally interested in evaluating the effect of weak identification on a test of the statistical significance of the estimated causal effect. The results in their paper apply to our discussion verbatim. In particular, if researchers suspect that their discontinuity design is weakly identified, adjustments must be implemented on the estimator of the asymptotic variance in order to *robustify* inferential procedures.

Stock, 1997, Feir et al., 2011). To avoid this type of situation, we set $c = 2$, so that the probability of receiving treatment jumps from 0.5 to 0.977 at the threshold (which in this design is set to $z_o = 0$).

To estimate $\beta$ we used the statistics $\hat{\tau}$ in equation (2.5). Each of the conditional moments in $\hat{\tau}$ were calculated via local linear regression with a Gaussian, uniform of gamma kernels. As discussed in the previous section, each of these kernels requires different bandwidth parameters. Therefore, to be able to compare the results, we used the Ruppert-Sheather-Wand procedure described in subsection 2.1.2.

The results of the simulation are in table 2. It is apparent that the gamma kernel contributes a notable gain in terms of efficiency. The average standard error of the simulation is around 0.34 when LLR was used with a Gaussian or uniform kernels, while it decreases to about 0.28 when using the gamma kernel. In these simulations, the average bias of all three estimators is comparable and it is not possible to establish which method exhibits the largest bias overall. These results are independent of the values of $\beta$ and $\rho$. The results also confirm the expectations about the magnitude of the bandwidth. This parameter is largest when using the uniform kernel and smallest when using the gamma kernel. Therefore, this simulation confirms the theoretical results introduced in section 2.

## 4 Application: Do voters affect or elect policies?

In an influential paper, Lee, Moretti and Butler (LMB, hereafter) study in what way voters influence the formation of policies, a question that lies at the core of political economy. The view that has dominated economics since the late 1950s is that competition for votes leads to *full/partial policy converge*, so that opposing parties adopt similar policies in an attempt to reflect voters' favourite policies (Downs, 1957). This downsian paradigm assumes that politicians always implement the policies they com-

$\beta = 0,\ \rho = 0.5$

| | $\hat{\tau}$ | Bias | S. D. | MSE | $h_{Y^-}$ | $h_{Y^+}$ | $h_{X^-}$ | $h_{X^+}$ |
|---|---|---|---|---|---|---|---|---|
| Gaussian | 0.013504 | 0.013504 | 0.34593 | 0.61390 | 0.36956 | 0.37676 | 0.42072 | 0.43399 |
| Uniform | 0.021221 | 0.021221 | 0.34709 | 0.68169 | 0.84853 | 0.83587 | 0.95449 | 0.94993 |
| Gamma | 0.024719 | 0.024719 | 0.28210 | 0.52429 | 0.31378 | 0.31153 | 0.35163 | 0.35419 |

$\beta = 0,\ \rho = 0.9$

| | $\hat{\tau}$ | Bias | S. D. | MSE | $h_{Y^-}$ | $h_{Y^+}$ | $h_{X^-}$ | $h_{X^+}$ |
|---|---|---|---|---|---|---|---|---|
| Gaussian | 0.038579 | 0.038579 | 0.34919 | 0.62865 | 0.36800 | 0.36939 | 0.42585 | 0.42113 |
| Uniform | 0.042739 | 0.042739 | 0.35106 | 0.70724 | 0.84537 | 0.83809 | 0.94056 | 0.96576 |
| Gamma | 0.035515 | 0.035515 | 0.28338 | 0.52580 | 0.31210 | 0.31339 | 0.35409 | 0.35384 |

$\beta = 2,\ \rho = 0.5$

| | $\hat{\tau}$ | Bias | S. D. | MSE | $h_{Y^-}$ | $h_{Y^+}$ | $h_{X^-}$ | $h_{X^+}$ |
|---|---|---|---|---|---|---|---|---|
| Gaussian | 2.0239 | 0.023868 | 0.34641 | 0.60387 | 0.37889 | 0.36620 | 0.41904 | 0.42583 |
| Uniform | 2.0314 | 0.031442 | 0.34774 | 0.68537 | 0.86577 | 0.82130 | 0.96968 | 0.95950 |
| Gamma | 2.0229 | 0.022918 | 0.28081 | 0.51033 | 0.31934 | 0.31060 | 0.35342 | 0.35474 |

$\beta = 2,\ \rho = 0.9$

| | $\hat{\tau}$ | Bias | S. D. | MSE | $h_{Y^-}$ | $h_{Y^+}$ | $h_{X^-}$ | $h_{X^+}$ |
|---|---|---|---|---|---|---|---|---|
| Gaussian | 2.0226 | 0.022571 | 0.35329 | 0.61850 | 0.33868 | 0.37094 | 0.42191 | 0.42215 |
| Uniform | 2.0057 | 0.005724 | 0.35089 | 0.66684 | 0.76745 | 0.82471 | 0.94961 | 0.95850 |
| Gamma | 2.0100 | 0.010037 | 0.28421 | 0.50646 | 0.28825 | 0.30853 | 0.35267 | 0.35555 |

Table 2: Monte Carlo Experiment based on a design by Feir et al. (2011). The sample size was 1000 observations while the experiment was repeated 20.000 for each combination of $\beta$ and $\rho$. Bandwidths were chosen as suggested in Ruppert et al. (1995).

mitted to during the elections, so that credibility is taken for granted. However, when politician's credibility is questionable, not only are partisan policies unlikely to converge, but also winning candidates have an incentive to simply pursue their most preferred policy (Alesina, 1988, Osborne and Slivinski, 1996, Besley and Coate, 1997). Therefore, if politicians' promises are credible, different candidates will moderate their discourse in order to reflect the preferences of the electorate (voters affect policies). If candidates' policies are not credible, then elections become a mechanism to decide a candidate's political programme (voters elect policies).

To assess which setting is more relevant in practice, Lee et al. (2004) devised an identification strategy that relies on Alesina's observations regarding the relationship between relative electoral strength and convergence/divergence of policies. In a context of full/partial political convergence, a candidate with a strong political support can afford to implement more radical policies at a small cost, while weaker candidates will have to moderate their policies. However, if politicians cannot credibly promise to implement moderate policies, the relative electoral strength of the candidates is irrelevant, since the purpose of elections is to decide which programme is implemented. To study which stream is more relevant in practice, Lee et al. (2004) focus on elections to the U.S. House of Representatives. These elections are generally a competition between a Republican and Democrat whose bliss points or preferred policies are normally in conflict regarding important policy issues. Given the bipartisan nature of the electoral race, a Democrat candidate is elected if he/she has above 50% of the electoral support. There is no term limits, and so it is possible to develop a whole career as a Representative, so if credibility matters, it is likely that convergence will be observed among the policies of Representatives, what makes these elections an interesting case study to evaluate voters' role in influencing policy.

Electoral support is endogenous to the type of policies supported by a candidate. However it is possible to devise an identifying strategy as follows. Let $t$ denote an electoral cycle. For instance, when $t = 1992$ it includes the 1992 electoral campaign, the November 1992 election and the 1993-1994 Congressional session. In each Congressional session, Representatives vote sets of legislation, and their voting record reflects how liberal each candidate's stand on policy is. LMB use the American Democratic Action's (ADA) voting score to describe each Representative's policy stance. The ADA score is an index between 0 and 100 with higher scores denoting more liberal voting record[10]. In what follows, a representative's ADA score in the Congressional session $t$ will be denoted by $ADA_t$. Let $P_t^*$, $P_t$ be a Democrat candidate's electoral strength (which is unobservable) and probability of victory at $t$ (vote share) respectively, $P^{*D}$ and $P^{*R}$ are a Democrat candidate's electoral strength given Democrat and Republican incumbency respectively, $D_t$ is an indicator of a Democrat victory, and $\varepsilon_t$ is a random variable reflecting heterogeneity among candidates' preferred policies. Lee et al. (2004) show that,

$$ADA_t = \alpha + \pi_o P_t^* + \pi_1 D_t + \varepsilon_t \tag{4.1}$$

In the above equation the parameter $\pi_o$ measures the sensitivity of a party's optimal policy to changes in electoral strength, while $\pi_1$ emphasises that voting scores depend on party affiliation.

The key observation to identify the effect of interest is that "... *incumbents are known to possess an electoral advantage...*"(Lee et al., 2004, pg. 815) and therefore are likely to have a stronger electoral strength. Assume by now that the winner of an election at time $t$ could be decided randomly, so that $D_t$ is independent of $P^*$ and $\varepsilon_t$ (and so on $\varepsilon_{t+1}$). Random assignment ensures that constituencies are comparable. Then, the difference between the $ADA_{t+1}$ scores of the winners of the election at $t+1$ where the Democarats had held the seat during the $t+1$ campaign and the $ADA_{t+1}$ scores of the winners of the election at $t+1$ where the Republicans had held the seat during the $t+1$ campaign

[10]Further details about roll-call voting records and ADA's voting scores can be found in Groseclose et al. (1999) or at ADA's website, `http://www.adaction.org/`.

is a valid measure of the causal effect of incumbency on voting records. Therefore, under the assumption of random allocation of winners in electoral cycle $t$ one can estimate the causal effect of electoral strength,

$$\gamma = E(ADA_{t+1}|D_t = 1) - E(ADA_{t+1}|D_t = 0).$$

Here $ADA_{t+1}$ and $D_t$ are observable in the sample, so they can be estimated via a consistent method. The parameter $\gamma$ is interesting in itself, but it does not answer the question of whether voters affect or elect policies. However, under the assumption of random allocation of winners at time $t$ and equation (4.1), one can decompose $\gamma$ in two terms,

$$E(ADA_{t+1}|D_t = 1) - E(ADA_{t+1}|D_t = 0) = \pi_o \left(P^{*D}_{t+1} - P^{*R}_{t+1}\right) + \pi_1 \left(P^{D}_{t+1} - P^{R}_{t+1}\right) \quad (4.2)$$

where

$$E(ADA_t|D_t = 1) - E(ADA_t|D_t = 0) = \pi_1 \quad (4.3)$$

$$E(D_{t+1}|D_t = 1) - E(D_{t+1}|D_t = 0) = P^{D}_{t+1} - P^{R}_{t+1} \quad (4.4)$$

The first term in equation (4.2), $\pi_o \left(P^{*D}_{t+1} - P^{*R}_{t+1}\right)$, is the *affect* component; it measures the sensitivity of parties/candidates to political strength, that is, how candidates may moderate their policies depending on the likelihood of winning the election. The second term in equation (4.2), $\pi_1 \left(P^{D}_{t+1} - P^{R}_{t+1}\right)$, is the *elect* component; it reflects that policies will tend to be more liberal wherever a Democrat candidate held the sit during the $t+1$ election because of the higher electoral strength of incumbents. Equations (4.3) and (4.4) show how to estimate the elect component directly from the data -using a consistent estimator of the conditional moments. Once this is done, the affect component can be computing as the difference of (4.2) and the product of (4.3) with (4.4).

The decomposition of the overall effect $\gamma$ is feasible under the assumption of randomisation. Random allocation of power is unrealistic, however Lee et al. (2004) note that quasi-experimental variation can be assumed by focusing attention on elections that were decided by a very narrow margin (say 2%) because the winner in these elections is decided *virtually at random* (Lee, 2008), and the distribution of the corresponding constituencies is similar is all respects except in the sign of the election winner[11]. This argument justifies the existence of a Regression Discontinuity design with electoral share at time $t$ acting as the running variable and 50% as the threshold at which potential discontinuities should take place.

We replicate the study in Lee et al. (2004)[12] using their data set which contains $ADA$ scores for all Representatives in the U.S. House from 1946 to 1995, as well as effective voting share for each party in the elections. There are in total 18842 elections with about 1200 of these being decided by a narrow margin of $\pm 2\%$ of the vote share. We use all the sample available, and estimate equations (4.2)-(4.4) via nonparametric (local linear) regression using three different kernels (Gaussian, Uniform and Gamma). Different bandwidths are calculated for the regressions to the left and right of the 50% vote-share cut-off point using the method proposed by Ruppert et al. (1995) which was described in section 2.

The results are given in table 3. The first row in the table reproduces the results in table I of LMB. In accordance to these estimates, the overall effect of Democrat incumbency on roll-vote is an increase of about 21 ADA points. The estimates of $\pi_1$ and $P^D_{t+1} - P^R_{t+1}$ are 47.6 and 0.48 respectively, so that the elect effect equals $47.6 \times 0.48 = 22.82$, while the estimated affect component equals $21.2 - 22.84 = -1.64$. In accordance to these results, the elect effect "*overwhelmingly*"dominate the affect component.

The results obtained with each of the three nonparametric methods are similar.

[11] An excellent and more detailed discussion is provided in Lee et al. (2004).

[12] The data set for this study is available at Enrico Moretti's website, `http://elsa.berkeley.edu/ moretti/`

Close Election Sample

| | $\hat{\gamma}$ | $\hat{\pi}_1$ | $P^D_{t+1} - P^R_{t+1}$ |
|---|---|---|---|
| LMB | 21.2 | 47.6 | 0.48 |
| | (1.9) | (1.3) | (0.02) |
| Gaussian | 21.055 | 47.727 | 0.40965 |
| | (1.6862) | (1.2168) | (0.030754) |
| Uniform | 21.137 | 48.011 | 0.39357 |
| | (1.7680) | (1.3086) | (0.035747) |
| Gamma | 21.032 | 48.016 | 0.41913 |
| | (1.3915) | (1.0238) | (0.024029) |

Table 3: Estimated treatment effects with standard errors (in parenthesis). LMB are the results reported in Lee et al. (2004) using observations between for which the Democrat vote share at time $t$ is between 0.48 and 0.52 (915 observations). For the remaining estimators we use the whole sample and bandwidth is selected in accordance to the procedure in Ruppert et al. (1995).

The elect effects estimated when using Gaussian, Uniform and Gamma kernels were 19.568, 18.9163 and 23.52 respectively. The standard errors of all four sets of results are comparable. However, although it is not possible to establish an efficiency ranking among the LMB, Gaussian and Uniform methods, the Gamma method typically dominates all others in terms of efficiency. For instance, the standard errors estimated by the LMB, Gaussian and Uniform methods for $\hat{\pi}_1$ are 1.3, 1.2 and 1.3 respectively, while the standard error obtained using the Gamma kernel was 1.023. This implies an efficiency gain of about 25% with respect to the LMB/Uniform methods. These results simply corroborate what was found in section 2 and section 3, namely that by avoiding the boundary bias, the gamma kernel can produce substantial efficiency gains in estimation of causal effects in regression discontinuity designs.

# 5 Conclusion

This article has considered the use of Local Linear Regression (LLR) with asymmetric kernels in order to estimate causal effects in a Regression Discontinuity design. At the centre of the RD methodology sits a local Wald estimator (Wald, 1940), whose components are typically estimated via LLR. Symmetric kernels are generally used at the core of these LLR, but this results in a local Wald estimator whose asymptotic distribution depends on the specific choice of kernel. The key to this result is the boundary bias problem affecting nonparametric regression. LLR circumvents this problem asymptotically, but not in small samples. Following previous results in Chen (2001), we showed that using the gamma kernel results in a more efficient local Wald estimator. Although our results focus on gamma kernels, other choices of kernel are available. Chen (2001) explores beta kernels, while Scaillet (2004) has used Inverse Gaussian and Reciprocal Inverse Gaussian kernels for density estimation. Other choices might be also feasible. However an study of alternative asymmetric kernels and their performance is beyond the scope of the present article, and it is left for future consideration.

## References


Alesina, A. (1988). Credibity and policy convergence in a two-party system with rational voters. *American Economic Review. 78*, 796–805.

Almond, D. and J. Doyle (2011). After midnight: A regression discontinuity design in length of postpartum hosptial stays. *American Economic Journal: Economic Policy. 3*, 1–34.

Almond, D., J. Doyle, A. Kowalski, and H. Williams (2010). Estimating marginal returns to medical care: Evidence from at-risk newborns. *Quarterly Journal of Economics. 125*, 591–634.

Angrist, J. and V. Lavy (1999). Using Maimonides' rule to estimate the effect of class size on scholastic achievement. *Quarterly Journal of Economics 114*, 533–575.

Battistin, E., A. Brugiavini, E. Rettore, and G. Weber (2009). The retirement consumption puzzle: Evidence from a Regression Discontinuity approach. *American Economic Review 99*, 2209–2226.

Besley, T. and S. Coate (1997). An economic model of representative democracy. *Quarterly Journal of Economics 112*, 85–114.

Bound, J., D. Jaeger, and R. Baker (1995). Problems with instrumental variables estimation when the correlation between the instruments and the endogenous explanatory variable is weak. *Journal of the American Statistical Association 90*, 443–450.

Brown, B. and S. X. Chen (1999). Beta-Bernstein smoothing for regression curves with compact supports. *Scandinavian Journal of Statistics. 26*, 47–59.

Chen, S. X. (2000). Beta kernel smoothers for regression curves. *Statistica Sinica 10*, 73–91.

Chen, S. X. (2001). Local linear smoothers using asymmetric kernels. *Annals of the Institute of Statistical Mathematics. 54*, 312–323.

Downs, A. (1957). *An economic theory of democracy.* Harper Collins.

Fan, J. (1993). Local linear regression smoothers and their minimax efficiencies. *Annals of Statistics 21*, 196–216.

Feir, D., T. Lemieux, and V. Marmer (2011). Weak identification in Fuzzy Regression Discontinuity desings. Technical report, Micro Theory Working Paper, Microeconomics.ca.

Frandsen, B., M. Frolich, and B. Melly (2011). Quantile treatment effects in the regression discontinuity design. *MIMEO*.

Groseclose, T., S. Levitt, and M. Snyder (1999). Comparing interest group scores across time and chambers: Adjusted ada scores for the u.s. congress. *The American Political Science Review 93*, 33–50.

Hahn, J., P. Todd, and W. van der Klaauw (1999). Evaluating the effect of an antidiscrimination law using a regression-discontinuity design. *NBER Working Paper 7131.*

Hahn, J., P. Todd, and W. van der Klaauw (2001). Identification and estimation of treatment effects with a regression discontinuity design. *Econometrica 69*, 201–209.

Hart, J. (1997). *Nonparametric Smoothing and Lack-of-Fit Tests.* Springer-Verlag New York.

Imbens, G. and T. Lemieux (2008). Regression Discontinuity Designs: A guide to practice. *Journal of Econometrics 142*, 615–635.

Lee, D. (2008). Randomized experiments from non-random selection in U.S. House elections. *Journal of Econometrics 142*, 675–674.

Lee, D., E. Moretti, and M. Butler (2004). Do voters affect or elect policies? Evidence from the U.S. House. *Quarterly Journal of Economics 119*, 807 – 859.

Ludwig, J. and D. Miller (2007). Does a head start improve childrens' life chances? evidence from a Regression Discontinuity Design. *Quarterly Journal of Economics 122*, 159–208.

Osborne, M. and A. Slivinski (1996). A model of political competition with citizen candidates. *Quarterly Journal of Economics 111*, 65–96.

Porter, J. (2003). Estimation in the Regression Discontinuity model. Technical report, Harvard University.

Ruppert, D., S. J. Sheather, and M. P. Wand (1995). An effective bandwidth selector for local least squares regression. *Journal of the American Statistical Association 90*, 1257–1270.

Ruppert, D. and M. P. Wand (1994). Multivariate locally weighted least squares regression. *Annals of Statistics 22*, 1346–1370.

Scaillet, O. (2004). Density estimation using inverse and Reciprocal Inverse Gaussian kernels. *Journal of Nonparametric Statistics 16*, 217–226.

Seifert, B. and T. Gasser (1996). Finite sample variance of local polynomials: Analysis and solutions. *Journal of the American Statistical Association 91*, 267–275.

Silverman, B. (1986). *Density Estimation for Statistics and Data Analysis.* Chapman and Hall.

Staiger, D. and J. Stock (1997). Instrumental variables regression with weak instruments. *Econometrica 65*, 557–586.

Thistlethwaite, D. and D. Campbell (1960). Regression-Discontinuity analysis: An alternative to the ex post facto experiment. *Journal of Educational Psychology 51*, 309–317.

van der Klaauw, W. (1996). A Regression-Discontinuity evaluation of the effect of financial aid offers on college enrollment. Technical report, New York University, Economics Department.

Wald, A. (1940). The fitting of straight lines if both variables are subject to error". *Annals of Mathematical Statistics 11*, 284300.

Wand, M. and M. Jones (1994). *Kernel smoothing.* Chapman and Hall/CRC.

# A Proofs

This appendix contains the proofs of the main results. I first obtain the asymptotic distribution of $\hat{m}^+$ at the boundary. The asymptotic distribution of the estimators of $m^-$, $p^+$ and $p^-$ is obtained identically, and so the proof concerning these cases is omitted.

Define $\mathbf{Z}_i^T = (1, (Z_i - z))$, $\theta = (m^+(z), m'^+(z))^T$ and $Y_i = m(Z_i) + u_i$. Throughout we use the notation $z = z_o$ to denote the boundary on which estimation is taking place -and equals 0 or 1, depending on the kernel employed. Define $\mathbf{Z}$ as the $n \times 2$ matrix with rows $\mathbf{Z}_i^T$, $\mathbf{K}$ is the diagonal matrix with elements $K_{z,b}(z_i)\mathbb{I}(z_i > z)$ and let $\mathbf{M}$ and $\mathbf{u}$ be $n \times 1$ matrices with elements $m(z_i)$ and $u_i$ respectively. The design points, $z_i$ have been normalised to fall within $[0, \infty)$. Finally let $\mathbf{e}_1^T = (1, 0)$.

The estimator of $m^+$ that solves equation (2.4) can be written as:

$$\begin{aligned}
\hat{m}^+(z) - m^+(z) &= \mathbf{e}_1^T(\mathbf{Z}^T\mathbf{K}\mathbf{Z})^{-1}\mathbf{Z}^T\mathbf{K}(\mathbf{M} + \mathbf{u} - \mathbf{Z}\theta) \\
&= \mathbf{e}_1^T(\mathbf{Z}^T\mathbf{K}\mathbf{Z})^{-1}\mathbf{Z}^T\mathbf{K}(\mathbf{M} - \mathbf{Z}\theta) \\
&\quad + \mathbf{e}_1^T(\mathbf{Z}^T\mathbf{K}\mathbf{Z})^{-1}\mathbf{Z}^T\mathbf{K}\mathbf{u} \qquad \text{(A-1)}
\end{aligned}$$

**Lemma A.1** *Under assumption 2.1,*

$$\mathbf{e}_1^T(\mathbf{Z}^T\mathbf{K}\mathbf{Z})^{-1}\mathbf{Z}^T\mathbf{K}(\mathbf{M} - \mathbf{Z}\theta) = b^2 m''^+(z) + o_p(1) \qquad \text{(A-2)}$$

*at z.*

**Proof.** Consider the LLR with the gamma kernel. The matrix $\mathbf{Z}^T\mathbf{K}\mathbf{Z}/n$ has typical element

$$S_l(z) = n^{-1}\sum_{i=1}^{n} K_{z,b}(Z_i)\mathbb{I}(Z_i > z)(Z_i - z)^l$$

for $l = 0, 1, 2$. Chen (2001) shows that $S_l(z) = E(S_l(z)) + o_p(1)$, where

$$\begin{aligned}
E\left(S_l(z)\right) &= \int_0^\infty K_{z,b}(z_i)(z_i - z)^l f(z_i) dz_i \\
&= E(f(\xi_i)(\xi_i - z)^l) \\
&= E\left\{\left(f(z) + f'(z)(z_i - z) + \frac{f''(z)}{2}(\xi_i - z)\right)(\xi_i - z)^l\right\} + O(b^2) \\
&= f(z)p_l(z) + f'(z)p_{l+1}(z) + \frac{f''(z)}{2}p_{l+2}(z) + O(b^2) \qquad \text{(A-3)}
\end{aligned}$$

where $\xi_i$ is a random variable with density $K_{z,b}(.)$. The third equality follows since $p_l = O(b^2)$ for $l \geq 2$. From this expression Chen provides the asymptotic approximation to the elements in $(\mathbf{Z}^T\mathbf{K}\mathbf{Z}/n)^{-1}$ (see Chen (2001), pg. 322). Secondly, $n^{-1}\mathbf{Z}^T\mathbf{K}(\mathbf{M} - \mathbf{Z}\theta)$ has typical element

$$n^{-1} \sum_i K_{z,b}(Z_i)\mathbb{I}(Z_i > z)(Z_i - z)^l (m(Z_i) - m^+(z) - m'^+(z)(Z_i - z))$$

for $l = 0, 1$. From assumption 2.2.3 it follows that

$$\begin{aligned}
& E\left\{\frac{1}{n}\sum_i K_{z,b}(Z_i)\mathbb{I}(Z_i > z)(Z_i - z)^l (m(Z_i) - m^+(z) - m'^+(z)(Z_i - z))\right\} \\
=\ & E\left\{\frac{1}{n}\sum_i K_{z,b}(Z_i)\mathbb{I}(Z_i > z)(Z_i - z)^l \left(\frac{1}{2}m''^+(z)(Z_i - z)^2 + \zeta(Z_i)\right)\right\} \\
=\ & \frac{1}{2}m''^+(z)E(f(\xi_i)(\xi_i - z)^{2+l}) + o(b^2)E(f(\xi_i)(\xi_i - z)^l) \\
=\ & \frac{1}{2}m''^+(z)\left(f(z)p_{2+l}(z) + O(b^2)\right) + o(b^2)\left(f(z)p_l + f'(z)p_{l+1} + \frac{f''(z)}{2}p_{l+2} + O(b^2)\right) \\
=\ & \frac{1}{2}m''^+(z)f(z)p_{2+l}(z) + o(1) \text{ for } l = 0, 1. \qquad \text{(A-4)}
\end{aligned}$$

The variance of each of these terms is such that,

$$
\begin{aligned}
& var\left(\frac{1}{n}\sum_i K_{z,b}(Z_i)\mathbb{I}(Z_i > z)(Z_i - z)^l \left\{\frac{1}{2}m''^{+}(z)(Z_i - z)^2 + \zeta(Z_i)\right\}\right) \\
\leq \; & \frac{1}{n}E\left\{K^2_{z,b}(Z_i)\mathbb{I}(Z_i > z)(Z_i - z)^{4+2l}\frac{1}{4}\left(m''^{+}(z)\right)^2\right\} \\
+ \; & \left\{o(b^2)\right\}^2 \frac{1}{n}E\left\{K^2_{z,b}(Z_i)(Z_i - z)^{2l}\right\} \\
\propto \; & \frac{A_b(z)}{n}E\left\{(\xi_i - z)^{4+2l}f(\xi_i)\right\} + \left\{o(b^2)\right\}^2 \frac{A_b(z)}{n}E\left\{(\xi_i - z)^{2l}f(\xi_i)\right\} \\
= \; & \frac{A_b(z)}{n}O(b^2) + \left\{o(b^2)\right\}^2 \frac{A_b(z)}{n}E\left\{f(z)p_{2l}(z) + O(b^2)\right\} \qquad \text{(A-5)}
\end{aligned}
$$

At the boundary, $A_b(z) = O(1/b)$ and so the first term above is $o(1)$, while for $l = 0, 1$, the second term is $o(1/nb)$. Therefore the above variance is bounded by a quantity of order $o(1)$. The result then follows from (A-4), since $p_2(z) = b^2(2 + (z/b)) \to b^2(2 + \kappa)$ at the boundary (see property 2.1). However, since $z = 0$, the above simplifies to $p_2(0) = 2b^2$, and the result then follows.

■

**Proof of Lemma 2.1.**

I only need to establish a Central Limit Theorem for the second term in equation (A-1). Begin by noting that the term $n^{-1}\mathbf{Z}^T\mathbf{K}\mathbf{u}$ has elements

$$T_l(z) = \sum_{i=1}^{n} K_{z,b}(Z_i)\mathbb{I}(Z_i > z)(Z_i - z)^l u_i$$

for $l = 0, 1$ with zero mean and conditional variance,

$$
\begin{aligned}
var(T_l(z)|z_1,\ldots,z_n) &= \frac{1}{n}E\left(K_{z,b}^2(Z_i)\mathbb{I}(Z_i>z)(Z_i-z)^{2l}\sigma^2(Z_i)\right)\\
&= \frac{A_b(z)}{n}E\left(f(\xi_i)\sigma^2(\xi_i)(z_i-z)^{2l}\right)\\
&= \frac{A_b(z)}{n}\left[f(z)\sigma^{2+}(z)p_{2l}(z)+O(b^2)\right]\\
&= \frac{\Gamma(2\kappa+1)}{nb\Gamma^2(\kappa+1)2^{2\kappa+1}}f(z)\sigma^{2+}(z)p_{2l}(z)+o(1) \qquad \text{(A-6)}
\end{aligned}
$$

for $l = 0, 1$. Let $\lambda = (\lambda_1, \lambda_2)^T$ be such that $\lambda^T\lambda = 1$, and consider,

$$
\sqrt{nb}\lambda^T\frac{\mathbf{Z}^T\mathbf{K}\mathbf{u}}{n} = \sum_{i=1}^n \frac{t_{ni}}{\sqrt{n}}
$$

where

$$
t_{ni} = \sqrt{b}\lambda_1 K_{z,b}(Z_i)\mathbb{I}(Z_i>z)u_i + \sqrt{b}\lambda_2 K_{z,b}(Z_i)\mathbb{I}(Z_i>z)(Z_i-z)u_i
$$

A sufficient Lyapounov condition is

$$
\lim_{n\to\infty}\sum_{i=1}^n E\left|\frac{t_{ni}}{\sqrt{n}}\right|^{2+\delta} = \lim_{n\to\infty}\frac{1}{n^{2+\delta/2}}\sum_{i=1}^n E\left|t_{ni}\right|^{2+\delta} = \lim_{n\to\infty} n^{-1-\delta/2}E|t_{ni}|^{2+\delta} = 0,
$$

and from this it follows that

$$
\begin{aligned}
|t_{ni}|^{2+\delta} &= E\left|\sqrt{b}\lambda_1 K_{z,b}(Z_i)\mathbb{I}(Z_i>z)u_i + \sqrt{b}\lambda_2 K_{z,b}(Z_i)\mathbb{I}(Z_i>z)(Z_i-z)u_i\right|^{2+\delta}\\
&\le 2^{1+\delta}E\left|\sqrt{b}\lambda_1 K_{z,b}(Z_i)\mathbb{I}(Z_i>z)u_i\right|^{2+\delta} + 2^{1+\delta}E\left|\sqrt{b}\lambda_2 K_{z,b}(Z_i)\mathbb{I}(Z_i>z)(Z_i-z)u_i\right|^{2+\delta}
\end{aligned}
$$

Now, the first term is such that,

$$
E\left|K_{z,b}(Z_i)u_i\right|^{2+\delta} = |u_i|^{2+\delta}\int_0^\infty |K_{z,b}(z_i)|^{2+\delta}f(z_i)dz_i \qquad \text{(A-7)}
$$

Note that

$$
\begin{aligned}
K_{z,b}(z_i)^{2+\delta} &= \frac{z_i^{(2+\delta)z/b} e^{-(2+\delta)z_i/b}}{b^{z(2+\delta)/b+(2+\delta)}\Gamma^{(2+\delta)}(z/b+1)} \frac{\Gamma((2+\delta)z/b+1)(2+\delta)^{-(2+\delta)z/b-1}}{\Gamma((2+\delta)z/b+1)(2+\delta)^{-(2+\delta)z/b-1}} \\
&= A_b^*(z) K_{x,b}^*(z_i), \qquad \text{(A-8)}
\end{aligned}
$$

where now,

$$
\begin{aligned}
A_b^*(z) &= \frac{(2+\delta)^{\frac{-(2+\delta)z}{b}} \Gamma\left(\frac{(2+\delta)z}{b}+1\right)}{b^{1+\delta}\Gamma^{(2+\delta)}\left(\frac{x}{b}+1\right)} \\
K_{x,b}^*(z_i) &= \frac{z_i^{k-1} e^{-z_i/\theta}}{\theta^k \Gamma(k)} \text{ for } \theta = \frac{b}{(2+\delta)} \text{ and } k = \frac{(2+\delta)z}{b} + 1
\end{aligned}
$$

so that $p_1(z) = 1$ and $p_2(z) \propto zb(2+\delta) + b^2$. Given the boundary condition $z/b \to \kappa$ ,

$$
A_b^*(z) \sim \frac{(2+\delta)^{-(2+\delta)\kappa}\Gamma((2+\delta)\kappa+1)}{b^{1+\delta}\Gamma^{(2+\delta)}(\kappa+1)} = O(1/b^{1+\delta}). \qquad \text{(A-9)}
$$

Thus, at the boundary,

$$
\begin{aligned}
E\left|K_{z,b}(Z_i)u_i\right|^{(2+\delta)} &= |u|_i^{(2+\delta)} A_b^*(z) E\left\{f(\xi_i)\right\} \\
&= |u|_i^{(2+\delta)} A_b^*(z) f(z) + |u|_i^{(2+\delta)} A_b^*(z) f'(z) + O(1/b^{\delta}) \\
&= O(1/b^{1+\delta}) \qquad \text{(A-10)}
\end{aligned}
$$

and similarly,

$$
\begin{aligned}
E\left|K_{z,b}(Z_i)(Z_i - z)u_i\right|^{(2+\delta)} &= |u_i|^{(2+\delta)} A_b^*(z) E\left\{f(\xi_i)(\xi_i - z)^{(2+\delta)}\right\} \\
&= |u_i|^{(2+\delta)} A_b^*(z) f(z) p_{(2+\delta)}(z) + O(b^2) \\
&= O(1/b^{1+\delta}) O(b^2) = o(1) \qquad \text{(A-11)}
\end{aligned}
$$

Therefore,

$$\lim_{n\to\infty}\frac{1}{n^{1+\delta/2}}E|t_{ni}|^{2+\delta}\leq\lim_{n\to\infty}\frac{b^{(2+\delta)/2}}{n^{1+\delta/2}}\left(O\left(\frac{1}{b^{1+\delta}}\right)+o(1)\right)=o(1)$$

provided that $nb\to\infty$ from which it follows that

$$\sqrt{nb}\left\{\hat{m}(x)^+-m(x)^+\right\}-\frac{\sqrt{nb}}{2}(2+\kappa)m''^+(x)b^2\sim N\left(0,\frac{\sigma^{2+}(x)\Gamma(2\kappa+1)}{2^{2\kappa+1}\Gamma(\kappa+1)f(x)}\right) \quad \text{(A-12)}$$

Noting that, by definition, $z=0$ (so that $\kappa=0$), the lemma follows.

## Proof of Theorem 2.1.

To prove the theorem, I first establish the covariance between $\hat{m}^+(z)$ and $\hat{p}^+(z)$. Consider firstly the terms $n^{-1}\mathbf{Z}^T\mathbf{K}\mathbf{u}$ and $n^{-1}\mathbf{Z}^T\mathbf{K}\mathbf{v}$ and note that,

$$\begin{aligned}
& E\left(\frac{1}{n}\sum_{i=1}^{n}K_{z,b}(Z_i)(Z_i-z)^l u_i\frac{1}{n}\sum_{i=1}^{n}K_{z,b}(Z_i)(Z_i-z)^l v_i|z_1,\ldots,z_n\right)\\
= & \frac{1}{n}E\left\{K_{z,b}^2(Z_i)(Z_i-z)^{2l}\eta(Z_i)\right\}\\
= & \frac{A_b(z)}{n}\left(f(z)\eta^+(z)p_{2l}(z)+O(b^{l+1})\right) \quad \text{(A-13)}
\end{aligned}$$

The final expression for the covariance between $\hat{m}^+$ and $\hat{p}^+$ follows from A-1, A-3 and results in page 322 of Chen (2001).

With the expression for the bias and the central limit theorem one deduces that

$$\sqrt{nb}\begin{pmatrix}(\hat{m}^+-\hat{m}^-)-(m^+-m^-)\\(\hat{p}^+-\hat{p}^-)-(p^+-p^-)\end{pmatrix}\to\mathcal{N}\left(\begin{pmatrix}\tau_m\\\tau_p\end{pmatrix},\begin{pmatrix}\lambda_m & \lambda_{mp}\\\lambda_{mp} & \lambda_p\end{pmatrix}\right) \quad \text{(A-14)}$$

where $\tau_m=\frac{b^2\sqrt{nb}}{2}(2+\kappa)(m''_+(z)-m''_-(z))$, $\lambda_m=f^{-1}(z)C(\sigma_+^2(z)+\sigma_-^2(z))$, $\lambda_{mp}=Cf^-1(z)(\eta^+(z)+\eta^-(z))$, and similarly for the remaining terms. Finally, a Taylor expansion (see Hahn et al. (1999) or proposition 1 in Porter (2003)), yields,

$$\sqrt{nb}\left(\frac{\hat{m}^{+}-\hat{m}^{-}}{\hat{p}^{+}-\hat{p}^{-}}-\frac{m^{+}-m^{-}}{p^{+}-p^{-}}\right)\rightarrow\mathcal{N}\left(\lambda,\tau\right) \tag{A-15}$$

where,

$$\lambda = \frac{b^2\sqrt{nb}}{2}(2+\kappa)\left(\frac{1}{p^{+}-p^{-}}(m''_{+}-m''_{-})-\frac{m^{+}-m^{-}}{(p^{+}-p^{-})^2}(p''_{+}-p''_{-})\right) \tag{A-16}$$

and

$$\begin{aligned}\tau &= \frac{C}{f(z)}\left(\frac{1}{(p^{+}-p^{-})^2}(\sigma^2_{+}+\sigma^2_{-})\right.\\ &- 2\frac{m^{+}-m^{-}}{(p^{+}-p^{-})^3}(\eta^{+}+\eta^{-})\\ &+ \left.\frac{(m^{+}-m^{-})^2}{(p^{+}-p^{-})^4}(p^{+}(1-p^{+})+p^{-}(1-p^{-}))\right)\end{aligned} \tag{A-17}$$

for $C=\frac{\Gamma(2\kappa+1)}{\Gamma^2(\kappa+1)2^{2\kappa+1}}$. The result then follows by noting that $\kappa$ is the limit of $z/b$, but since $z=0$, $\kappa=0$.