Note on *"Kernel density estimation for undirected dyadic data"* by Graham B., Niu, F. and Powell J.

Arkadiusz Szydłowski

University of Leicester

June 26, 2023

Abstract

In this note I show that the results in Graham et al. (2022) are linked to the well known result in the U-statistics literature by Frees (1994). In fact, the asymptotic $\sqrt{N}$ normality of the kernel density estimator in Graham et al. (2022) follows from arguments in Frees (1994).

Graham et al. (2022) analyse nonparametric estimation of marginal density of:

$$W_{ij} = W(A_i, A_j, V_{ij})$$

where $\{A_i\}_{i=1}^N$ and $\{V_{ij}\}_{i,j=1}^N$ are i.i.d. and mutually independent and the function W is symmetric in the first two arguments. Note that this implies that $W_{ij} \perp W_{kl}$ unless at least one of the indices in (i,j) and (k,l) coincide. They show that the kernel density estimator:

$$\hat{f}_W(w) = \frac{2}{N(N-1)} \sum_{i<j} \frac{1}{h_N} K\left(\frac{w - W_{ij}}{h_N}\right)$$

converges to a normal distribution at rate $\sqrt{N}$.

Frees (1994) analyses nonparametric estimation of marginal density of $g(A_1, A_2, \ldots, A_m)$, where $\{A_i\}_{i=1}^N$ is an i.i.d. sequence and g is symmetric in all arguments[1], and shows that the kernel density

[1] Giné & Mason (2007) extend his results to a uniform-in-bandwidth result.

estimator:

$$\hat{f}_g(t) = \binom{N}{m}^{-1} \sum_{1 \le i_1 < i_2 < \ldots < i_m \le N} \frac{1}{h_N} K\left(\frac{t - g(A_{i_1}, A_{i_2}, \ldots, A_{i_m})}{h_N}\right)$$

converges to a normal distribution at rate $\sqrt{N}$.

To see the relationship between the two results, first assume that V_{ij} is drawn from the same distribution as A_i's. As V_{ij}'s are independent of A_i's, without loss of generality we can write $W_{ij} \equiv W_{ijk} = W(A_i, A_j, A_k)$. Define the symmetrised version of W_{ijk} as:

$$g(A_i, A_j, A_k) = W(A_i, A_j, A_k) + W(A_k, A_i, A_j) + W(A_i, A_k, A_j)$$

(note that W is symmetric in the first two arguments). Now asymptotic $\sqrt{N}$ normality of the kernel density estimate of the density of g follows from the main theorem in Frees (1994). Note that, beyond standard conditions on the kernel function, Frees (1994) requires the density of $g(a, A_j, A_k)$, $w_1(t; a)$, to exist and satisfy $\sup_t E_A |w_1(t; A)|^{2+\delta} < \infty$, which is implied by smoothness conditions for W and density of V_{ij} imposed by Graham et al. (2022).

The previous discussion imposed some additional assumptions on the model in Graham et al. (2022). Here I show that even without restricting the distribution of V_{ij} (beyond assumptions in Graham et al. (2022)) and without symmetrising the function W in the third argument, the asymptotic $\sqrt{N}$ normality of the kernel density estimator follows from arguments in Frees (1994) as the shock V gets integrated out in this argument anyway. Below I show formally that the main result in Frees (1994), Theorem A, holds for the setup in Graham et al. (2022). The remaining part of the proof of normality deals with

$$W_{1N}(a, t) = h_N^{-1} E[K((t - W(a, A_2, V_{12}))/h_n)] - h_N^{-1} E\left[K\left((t - W(A_1, A_2, V_{12}))/h_N\right)\right],$$

thus follows exactly from Frees (1994).[2]

[2]Note that the bandwidth condition $Nh_N^4 \to 0$ in Graham et al. (2022) implies that the bias goes to 0, an assumption in the main theorem of Frees (1994).

Proposition 1. *Define* $R_N(t) = \frac{2}{N(N-1)} \sum_{1 \le i_1 < i_2 \le N} \tilde{g}(A_{i_1}, A_{i_2}, V_{i_1 i_2}; t)$ *where:*

$$\tilde{g}(a_1, a_2, v_{12}; t) = \frac{1}{h_N} K\left(\frac{t - W(a_1, a_2, v_{12})}{h_N}\right) - \frac{1}{h_N} E\left[K\left(\frac{t - W(A_1, A_2, V_{12})}{h_N}\right)\right] - W_{1N}(a_1, t) - W_{1N}(a_2, t)$$

Assume that K is a symmetric, bounded function that integrates to one and satisfies A5 in Graham et al. (2022), $w_1(t; a)$ exists and satisfies $\sup_t E_A |w_1(t; A)|^{2+\delta} < \infty$ for $\delta > 0$. Then:

$$R_N(t) = O_p(h_N^{-1/2} N^{-1}).$$

Proof. Note that $E[\tilde{g}(A_{i_1}, A_{i_2}, V_{i_1 i_2}; t) | A_{i_1}] = 0$. We have:

$$Var(R_N(t)) = \frac{4}{N^2(N-1)^2} \sum_{1 \le i_1 < i_2 \le N} \sum_{1 \le j_1 < j_2 \le N} E[\tilde{g}(A_{i_1}, A_{i_2}, V_{i_1 i_2}; t) \tilde{g}(A_{j_1}, A_{j_2}, V_{j_1 j_2}; t)]. \quad (1)$$

When $\{i_1, i_2\}$ and $\{j_1, j_2\}$ have 0 or 1 element in common the expectation under the sum is zero. Otherwise, $E[\tilde{g}^2(A_{i_1}, A_{i_2}, V_{i_1 i_2}; t)] \le h_N^{-1} E\left[K\left(\frac{t - W(A_1, A_2, V_{12})}{h_N}\right)^2\right] + h_N^{-1} E[W_{1N}(A_1, t)]^2 = O(h_N^{-1})$ by a standard argument. Finally, since the number of non-zero elements in the sum in (1) is of order $O(N^{-2})$ we have:

$$Var(R_N(t)) = O_p(h_N^{-1} N^{-2})$$

which is sufficient for the result. □

References

Frees, E. W. (1994), 'Estimating densities of functions of observations', *Journal of the American Statistical Association* **89**(426), 517–525.

Giné, E. & Mason, D. M. (2007), 'On local U-statistic processes and the estimation of densities of functions of several sample variables', *The Annals of Statistics* **35**(3), 1105–1145.

Graham, B. S., Niu, F. & Powell, J. L. (2022), 'Kernel density estimation for undirected dyadic data', *Journal of Econometrics* .