PARAMETRIC ENTROPY BASED DENSITY ESTIMATION IN DISTANCE SAMPLING

{ KEVIN JOYCE } THE UNIVERSITY OF MONTANA

DENSITY ESTIMATION

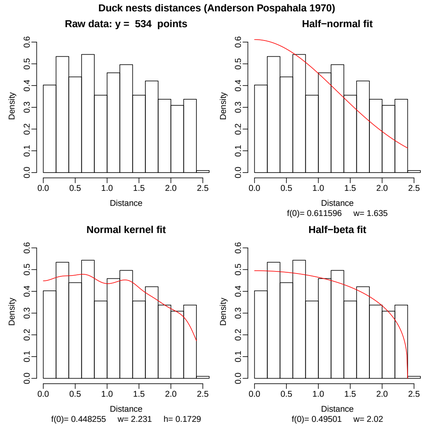


Figure 1: Distance data for duck nests in the Monte Vista National Wildlife Refuge in Colorado.

The problem of interest in this work is to estimate the density of objects in a known area from distance data collected along line-transects within the area. This problem has been extensively studied, and much work involves estimation based on the density estimator

$$\hat{D}_i = \frac{y_i \hat{f}(0)}{2L_i} \tag{1}$$

where y_i are the enumerated counts along the ith transect, L_i is the length of the ith transect[**?**] [**?**]. By far, the most involved part of the estimation is obtaining $\hat{f}(0)$, the peak of the probability distribution of the detected distances from the transect. This work is aimed at finding a robust parametric method for estimating $\hat{f}$ where detectability has relatively low kurtosis, or heavy central weighting within a fixed region.

BETA DISTRIBUTION

Of distributions supported on the interval $[0, 1]$, symmetric beta distributions form a flexible parametric family given by

$$f(x) = B(\alpha) x^{\alpha} (1 - x)^{\alpha} \tag{4}$$

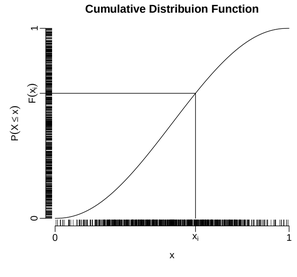


Figure 2: Finitely supported density data mapped through a beta CDF that maximizes entropy.

ESTIMATING ENTROPY

A well-known result from mathematical statistics states that if a random variable X is distributed with a cumulative distribution function (CDF) $F(x)$, then the random variable $U = F(X)$ is distributed uniformly over $[0, 1]$. Of all random variables supported on $[0, 1]$ with density function $f(x)$, it can be shown that the entropy statistic

$$H(X) = \int_0^1 f \log f \, dx \tag{2}$$

is maximized when the distribution is uniform [**?**]. There is much work on estimating the entropy statistic , and we employ a straight-forward histogram based estimator

$$\widehat{H} = \sum_{\text{bins}} f(n_i) \log \frac{f(n_i)}{w_i}. \tag{3}$$

This is known to be MLE [**?**] and has been implemented in `R` in the freely available `CRAN` library `entropy`.

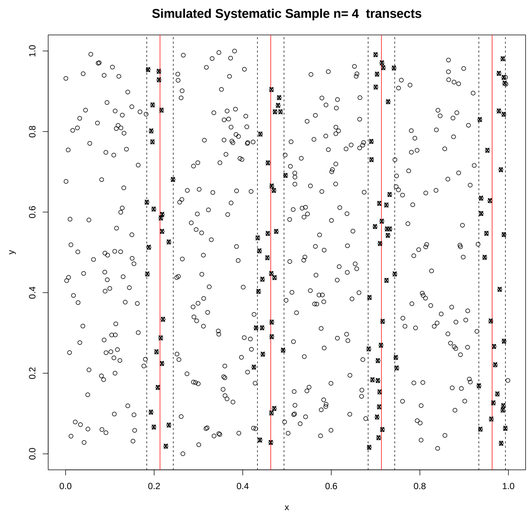
DENSITY ESTIMATION WITH ENTROPY MAXIMIZATION

The beta parametric family provides a flexible collection of functions that have sufficiently low kurtosis to fit the problematic distance data given in the introduction. If we reflect the distance data about zero and rescale to $[0, 1]$, we can fit the data to a half-beta distribution by optimizing our entropy estimate. This leads to the following optimization problem for estimating a symmetric density supported on a finite interval:

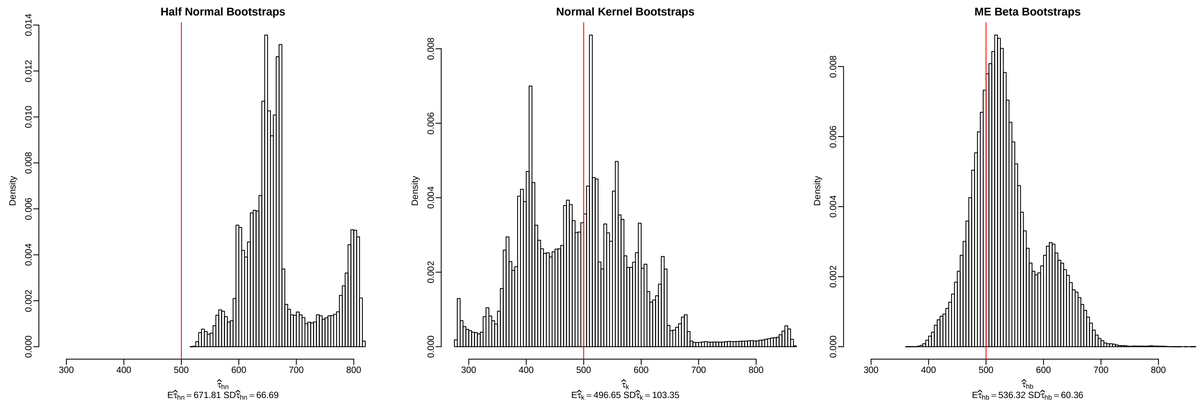
$$\hat{f} = \underset{f \in \text{beta}}{\arg\min}\, \widehat{H}(f)^{-1} \tag{5}$$

To mimic Anderson and Pospahala's data, we conducted a simulation where detected points were given by Bernouli trials with probabilities

$$p(d) = \begin{cases} 1 & |d| < e \\ (1 - d)^p & |d| < b\, e. \end{cases} \tag{6}$$



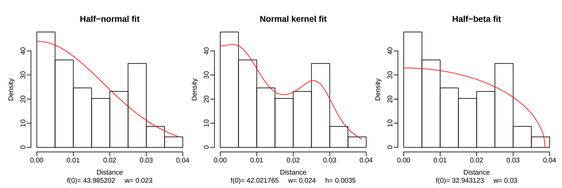
SIMULATION RESULTS



We estimated the total $\tau = 500$ with 500,000 ideal bootstrap systematic samples using half-normal estimation, normal kernel estimation with the recommended kernel width [**?**], and EM half-beta estimation with histogram bin sizes given by the Sturges method.

CONCLUSIONS

The half-normal estimator appears to be bimodal with a large positive bias. The normal kernel estimate is skewed right, but appears to be relatively unbiased in estimating the total. Our estimator has a significantly lower variance than the normal kernel estimator, but appears to have a positive bias and slight biomodality. In the figure to the right, it can be seen that the normal kernel estimator seems to overfit densities which may be the for the larger variance in the population total estimator. The EM half-beta estimator, on the other hand, appears to be more stable. Note that it is constrained to have zero probability at the maximum data value, and this could be the reason for the upward bias. A modification to the curves to account for this may remove the bias.



REFERENCES