Ensemble Kernel Density Estimation

USU Student Research Symposium – Ethan Ancell

Department of Mathematics and Statistics

Research Mentor: Kevin Moon

Kernel Density Estimation

Kernel density estimation (KDE) is the problem of estimating a probability density function from a finite sample of data.

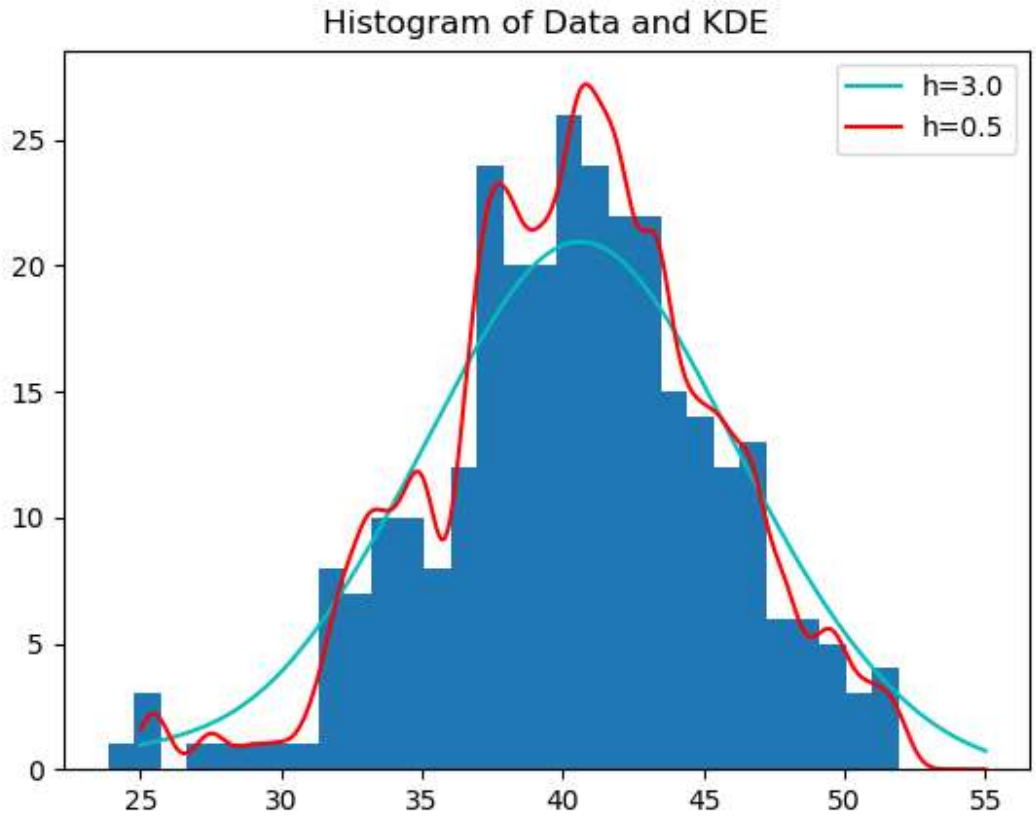
KDEs are of the form

$$\hat{p}(x) = \frac{1}{nh}\sum_{i=1}^{n} k\left(\frac{x - x_i}{h}\right)$$

n is the sample size, x_i is the indexed known data, $k(\cdot)$ is the kernel function, and h is the bandwidth size that controls the degree of smoothing.

Kernel density functions are a *nonparametric* technique to estimate densities and thus do not require any assumptions on the data, such as the class of distribution.

Example of KDE in action: I simulated 300 data points from a $\mathcal{N}(\mu = 40, \sigma^2 = 25)$ distribution with both $h = 3.0$ and $h = 0.5$ bandwidths.



Ensemble Estimation

Given an individual estimator $\hat{E}_l$, we define a weighted ensemble

$$\hat{E}_w \coloneqq \sum_{l \in \bar{l}} w_l \, \hat{E}_l$$

where $\bar{l}$ is a set of parameter values (with cardinality L) and $w \coloneqq \{w_{l_1}, \ldots, w_{l_L}\}$ is a set of weights such that $\sum_{l \in \bar{l}} w_l = 1$.

(These estimators are very general)

Previous work has shown that if:

1. The bias can be expressed $\mathbb{B}\left(\hat{E}_l\right) = \sum_{i \in \mathcal{J}} c_i \psi_i(l) n^{-i/2d} + O(1/\sqrt{n})$

   Where c_i are constants that depends on the underlying density, $\mathcal{J} \coloneqq \{i_1, \dots, i_I\}$ is an index set with cardinality $I < L$, and $\psi_i(l)$ are basis functions that only depend on l.

2. The variance can be expressed $Var\left(\hat{E}_l\right) = c_v \left(\frac{1}{n}\right) + o\left(\frac{1}{n}\right)$

Then, $\exists \boldsymbol{w}_0 \in \mathbb{R}^L$ such that the MSE: $\mathbb{E}\left[\left(\hat{E}_{\boldsymbol{w}_0} - E\right)^2\right] = O\left(\frac{1}{n}\right)$. (This is on the same rate of convergence as parametric estimators)

The optimal weight vector $\boldsymbol{w}_0$ in the previous slide can be found by solving the (fortunately) convex optimization problem:

$$\min_{w} \quad \left|\left|\boldsymbol{w}\right|\right|_2$$

$$s.t. \sum_{l \in \bar{l}} w_l = 1$$

$$\gamma_w(i) = \sum_{l \in \bar{l}} w_l \psi_i(l) = 0, i \in \mathcal{I}$$

The conditions that guarantee such a weight vector $\boldsymbol{w}_0$ are satisfied in density functionals (such as information divergences) and entropy estimation. KDEs do NOT meet those conditions.

The research task is to define KDEs $\hat{p}_l(\mathrm{x})$ with certain assumptions and prove that a weight vector $\boldsymbol{w}_0$ exists (obtained by solving an optimization problem) such that we can improve the rate of convergence of the MSE. The original aim was to see if the parametric rate can be achieved.

The proof involves controlling both the bias and variance terms, because $MSE(E) = \mathbb{B}(E)^2 + Var(E)$.

In the case that $Var(\hat{E}_l) = c_v\left(\frac{1}{n}\right) + o\left(\frac{1}{n}\right)$ the proof shows we can control the variance by relying on the weight vector:

$$Var(\hat{E}_w) = Var(\sum_{l\in\bar{l}} w_l \hat{E}_l) \leq \frac{L\left|\left|w\right|\right|^2}{n}$$ (some steps omitted)

For a KDE $\hat{f}_l$,

$$Var(\hat{f}_l) = \frac{f(x)R(k)}{nh} + O\left(\frac{1}{n}\right)$$

The first term in the above is the problem, because $h \to 0$ and thus $h^{-1} \to \infty$. Therefore, we can not use the same technique above. Because the first term converges slower than $O\left(\frac{1}{n}\right)$.

Can we reach into the other entries of the covariance matrix of $\hat{f}_w$ to control the variance?

For two KDEs f_1 and f_2 with a uniform kernel assumption on both with bandwidths $h_1 > h_2$

$$Cov(f_1, f_2) = \frac{n-1}{n}\left(f(x)^2 + c_1 h_1^{\nu_1} + c_2 h_2^{\nu_2}\right) + \frac{f(x)}{nh_2} + o\left(h_1^{\nu_1} + h_2^{\nu_2}\right) + o\left(\frac{1}{nh_2} + \frac{1}{n}\right)$$

We are fairly certain that it is impossible to achieve the parametric rate of convergence with the above. As a sketch of the idea,

$$Var\left(\sum_{l \in \bar{l}} \hat{f}_l\right) = \sum_{l,l' \in \bar{l}} w_l w_{l'} \Sigma_L(l, l') = \sum_{l,l' \in \bar{l}} w_l w_{l'} (h_l^{\nu} + h_{l'}^{\nu} + \frac{1}{nh_l}) = \sum_{l,l' \in \bar{l}} \frac{w_l}{nh_l}$$

The last equality in combination with the bias of KDE can get extremely close to the parametric rate of convergence but can not reach it.

Future Work

- Achieving a parametric rate of convergence with a KDE with this method is probably impossible. However, can we at least improve the convergence rate?

- It is likely that we can get the convergence of bias-cancelling KDEs (these are KDEs that require the assumption of the existence of higher order derivatives and further smoothness). This would be a very big deal even if we can't achieve the original goal of the parametric rate of convergence.

- It is also likely that we can asymptotically approach the parametric rate of convergence (i.e., get arbitrarily close to the parametric rate).