Ensemble Kernel Density Estimation

2022 Student Research Symposium

Ethan Ancell

Department of Mathematics and Statistics, Utah State University

Research Mentor: Dr. Kevin Moon

Estimators (general)

In statistics, we often try to guess some unknown parameter θ using an estimator $\hat{\theta}$. Typically, $\hat{\theta}$ is a function of our observed data.

<u>Examples:</u>

- Mean of a normal distribution μ with estimator $\hat{\mu} = \frac{1}{n}\sum x_i$
- Rate parameter of exponential distribution λ with estimator $\hat{\lambda} = \frac{n}{\sum x_j}$
- Probability density function $f(x)$ with estimator $\hat{f}(x)$ (there are a few ways to do this, kernel density estimation being one) **(we will return to this in a few slides)**

What makes an estimator “good”?

The MSE (mean squared error) measures the error of an estimator.

$$MSE(\hat{\theta}) = \mathbb{E}((\hat{\theta} - \theta)^2)$$

It is well-known that the MSE can be decomposed into:

$$MSE(\hat{\theta}) = Bias(\hat{\theta})^2 + Var(\hat{\theta})$$

Approaches to minimize the MSE of an estimator often involve simultaneously minimizing the bias and the variance.

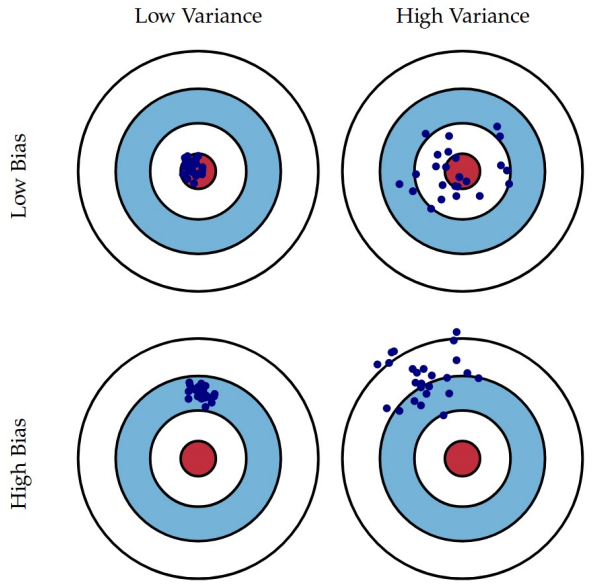
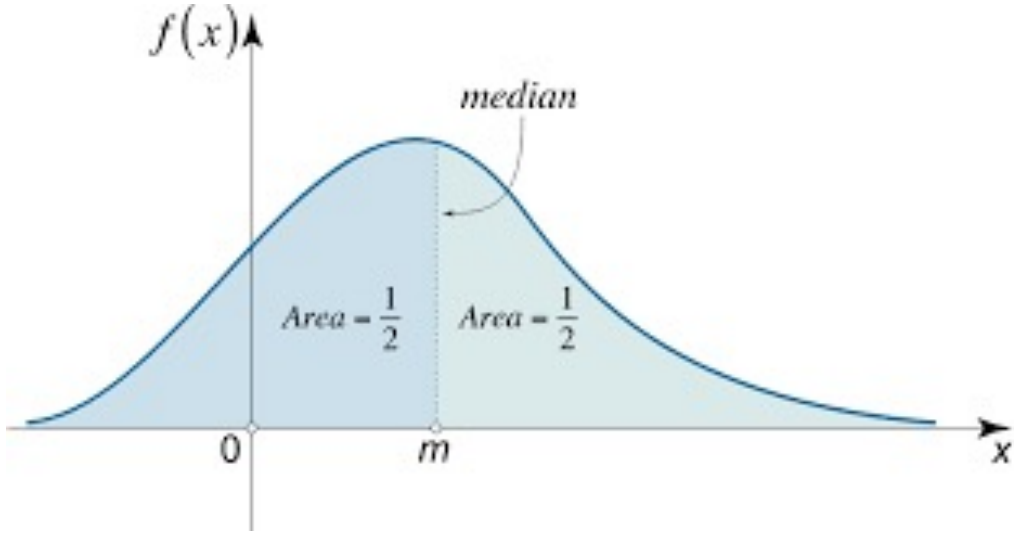


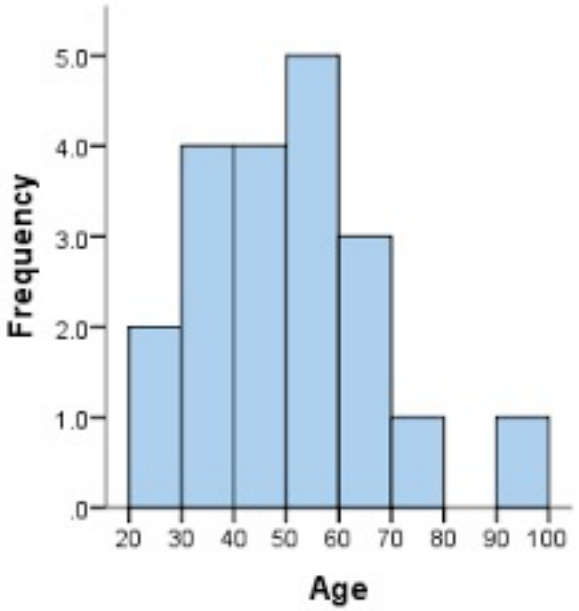
Fig. 1 Graphical illustration of bias and variance.

Probability density functions

Probability density functions contain all the important information regarding the distribution of a continuous random variable.



We might want to guess what this is using a finite sample of data. A naïve approach to guess this from data is with a histogram.



Kernel density estimation

Another way to estimate a PDF is with kernel density estimation (KDEs). This is nonparametric (doesn't assume distributional form).

KDEs are of the form:

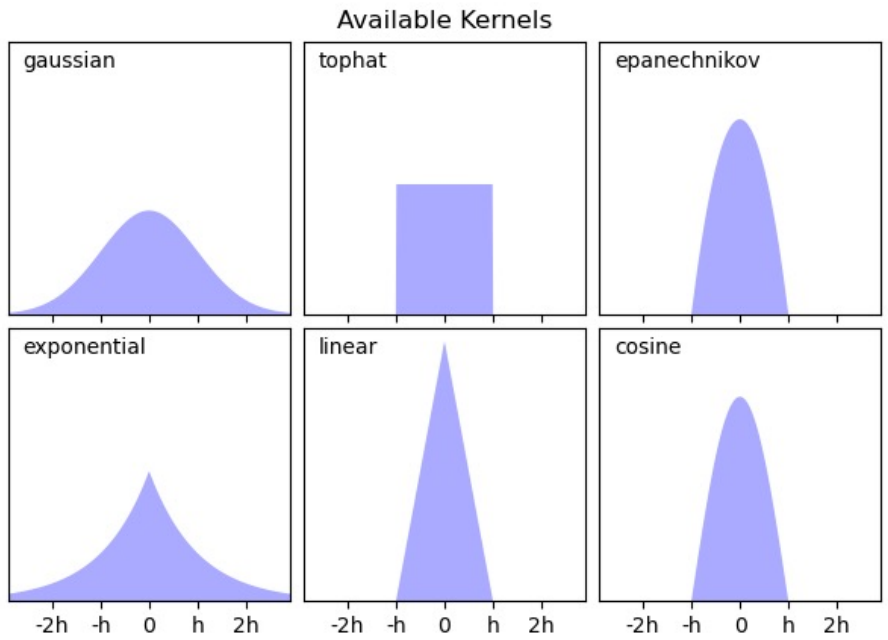
$$\hat{p}(x) = \frac{1}{nh}\sum_{i=1}^{n} k\left(\frac{x - x_i}{h}\right)$$

n: the sample size

x_i: known data

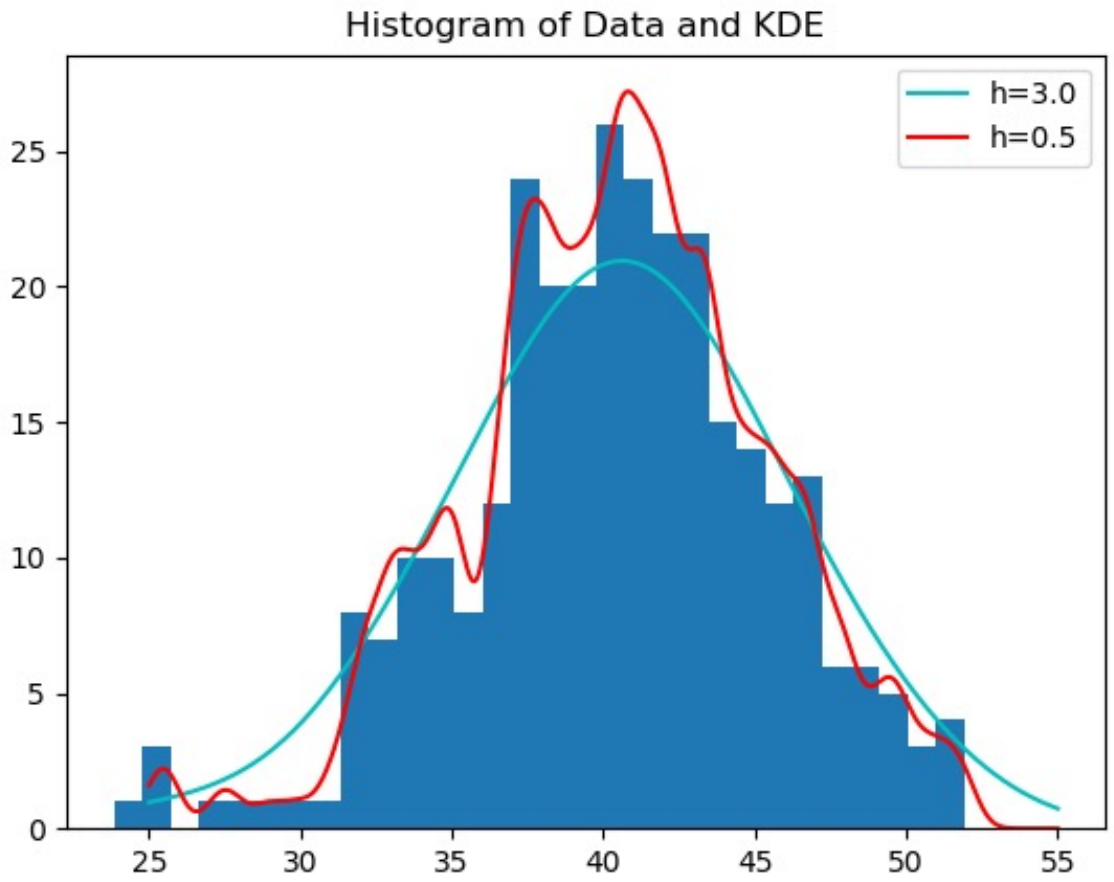
$k(\cdot)$: kernel function

h: bandwidth (degree of smoothing)



Example of a KDE

Simulation of 300 data points from a $\mathcal{N}(\mu = 40, \sigma^2 = 25)$ distribution. The red line is a KDE with $h = 0.5$ and the blue line is a KDE with $h = 3.0$.



Parametric vs nonparametric

As mentioned before, KDEs are nonparametric. If we know the true distribution of our data (e.g., normal, exponential, Rayleigh, gamma), then it is better in terms of MSE it is better to use parametric estimation.

For the mathematically inclined,

- MSE for parametric density estimation is on the order of $O\left(\frac{1}{n}\right)$
- MSE for nonparametric density estimation is on the order of $O(\frac{1}{\sqrt{n}})$

Can you make multiple estimators work in a team?

Yes. These are called ‘ensemble’ techniques. Examples from ML:

- Random forests
- AdaBoost
- Gradient boosting

Goal: use **ensemble** to improve kernel density estimation

Assumed form for ensemble

Let $\hat{\theta}_l$ be an estimator for a parameter θ. Define a weighted ensemble estimator as

$$\hat{\theta}_w \coloneqq \sum_{l \in \mathcal{L}} w_l \, \hat{\theta}_l$$

Where $\mathcal{L}$ is an index set with $|\mathcal{L}| = L$ and $w = \{w_{l_1}, \ldots, w_{l_L}\}$ is a set of weights such that $\sum_{l \in \mathcal{L}} w_l = 1$. The $\hat{\theta}_w$ is the ensemble estimator, and each $\hat{\theta}_l$ is a ”weak estimator.”

We will try to apply general ensemble estimation theory to kernel density estimation to see if we can get a better KDE.

MSE theorem for weighted ensemble estimator

Theorem 1 (Sricharan et al. 2013):

Suppose the following conditions hold for each weak estimator $\hat{\theta}_l$:

(C1) The bias can be expressed as

$$\mathbb{B}(\hat{\theta}_l) = \sum_{i \in \mathcal{J}} c_i \psi_i(\ell) \phi_i(n, d) + O(1/\sqrt{n})$$

(C2) The variance can be expressed

$$Var(\hat{\theta}_l) = c_v \left(\frac{1}{n}\right) + o\left(\frac{1}{n}\right)$$

Then, $\exists w \in \mathbb{R}^L$ such that the MSE: $\mathbb{E}\left[\left(\hat{\theta}_w - \theta\right)^2\right] = O\left(\frac{1}{n}\right)$

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

But how do we find $\boldsymbol{w}_0$?

The theorem states that the weight vector $\boldsymbol{w}_0$ giving the MSE rate of $O\left(\frac{1}{n}\right)$ can be found by solving the convex optimization problem:

$$
\begin{aligned}
\min_{\boldsymbol{w},\epsilon} \quad & \epsilon \\
\text{s.t.} \quad & \sum_{\ell \in \mathcal{L}} w(\ell) = 1, \\
& \left|\gamma_{\boldsymbol{w}}(i) n^{\frac{1}{2}} \phi_{i,d}(n)\right| \leq \epsilon, \quad \forall i \in \mathcal{J} \\
& ||\boldsymbol{w}||_2^2 \leq \eta\epsilon
\end{aligned}
$$

Does the KDE meet the hypotheses of Theorem 1?

Short answer: no.

Long answer: The bias meets the condition, but not the variance. However, we can extend the theorem and use weaker conditions to obtain an ensemble kernel density estimator that performs better than using a single estimator.

Controlling the Bias

The bias of a KDE fits condition (C1) in Theorem 1, but some work is needed to derive it. With some assumptions on smoothness (s times differentiable density), the bias of the KDE can be derived to be

$$\mathbb{B}(\hat{f}_j) = \sum_{i=1}^{s/2} c_i \ell_j^{2i} n^{\frac{-2i}{(d+\epsilon)}} + o(h_j^s)$$

c_i is a constant depending on the underlying density, and ℓ_j is a function of the bandwidth which is used in the optimization problem. Matching the above to the form of Theorem 1, $\psi_i(\ell) = \ell_j^{2i}$ and $\phi_{i,d}(n) = n^{-\frac{2i}{d+\epsilon}}$. Furthermore, if $h_j^S \to 0$ quickly enough in relation to n, then the asymptotic term can be replaced with $O(\frac{1}{\sqrt{n}})$.

Controlling variance (part 1)

As said before, variance does not meet hypothesis of Theorem 1. As a trick, we'll force the problematic part of the variance to be very small in our constrained optimization problem. For our ensemble KDE $\hat{f}_w$, the variance is

$$Var(\hat{f}_w) = \sum_{i,j \in \mathcal{J}} w_i w_j [\Sigma_L]_{i,j}$$

Where $[\Sigma_L]_{i,j}$ is the covariance between the ith and jth weak estimators, parameterized with bandwidths h_i and h_j respectively.

Controlling the Variance (part 2)

For two different KDEs designated $\hat{f}_\ell$ and $\hat{f}_{\ell'}$, the covariance can be derived to be

$$[\Sigma_L]_{i,j} = Cov(\hat{f}_\ell, \hat{f}_{\ell'}) = \frac{f(x)}{n \max(h_\ell, h_{\ell'})^d} + O\left(\frac{1}{n}\right) + o(h_i^s + h_j^s)$$

The above is nontrivial to derive and requires a few strong assumptions on our kernel function (compact domain).
If $h_i \to 0$ no slower than $\frac{1}{n} \to 0$ for all i, then, we substitute the above into $Var(\hat{f}_w)$ and we have

$$Var(\hat{f}_w) = \frac{f(x)}{n^{\frac{\epsilon}{d+\epsilon}}} \left(\sum_{i \leq j} w_i w_j \ell_j^{-d} + \sum_{j<i} w_i w_j l_i^{-d} \right) + O\left(\frac{|w|^2}{n}\right)$$

Trick is to take the first term of the above and "force it small" using optimization problem. If we keep $||w||^2$ small using optimization problem, this guarantees the variance of the ensemble will be on the order of $O\left(\frac{1}{n}\right)$ which is what we want.

Optimization problem

We use a similar optimization problem as in Theorem 1, except we add an extra constraint to accommodate the variance of the ensemble (third line of constraints below).

$$
\begin{aligned}
\min_{\boldsymbol{w},\epsilon} \quad & \epsilon \\
\text{s.t.} \quad & \sum_{l \in \bar{l}} w(l) = 1, \\
& \left| \gamma_w(i) n^{1/2} \phi_{i,d}(n) \right| \leq \epsilon \ \forall i \in \mathcal{J}, \\
& \frac{1}{n^{\epsilon_0/(d+\epsilon_0)-1}} \left[\sum_{i \leq j} w(\ell_i) w(\ell_j) \ell_j^{-d} + \sum_{j<i} w(\ell_i) w(\ell_j) \ell_i^{-d} \right] \leq \lambda \epsilon \\
& ||\boldsymbol{w}||_2^2 \leq \eta \epsilon
\end{aligned}
$$

Ensemble KDE in Practice

As input, you select:

- L: number of estimators
- ϵ: small positive constant for stability
- $\{h_\ell\}_{\ell \in \mathcal{L}}$: a set of bandwidths

After you select the above,

- Use optimization software to solve the constrained problem. This gives you a vector of weights.
- Estimate density by using the estimator $\hat{f}_w(x) = \sum_\ell w_\ell \hat{f}_\ell(x)$.

Experiment Design

To check this is better in practice, let us compare the following types of estimators with a given set of bandwidths $\{h_l\}_{l\in\mathcal{L}}$.

1. Ensemble estimator
2. Single estimator (one bandwidth chosen from $\{h_l\}_{l\in\mathcal{L}}$ using cross-validation)
3. Single estimator (rule of thumb chosen bandwidth)

Example Experiment Results for $d = 8, n = 500$

Average testing mean-squared error from 50 simulations using a randomized multivariate Gaussian density with randomized mean and covariance matrix.

Estimator	Average MSE on fixed 200 test points (50 randomizations)
Scott's rule of thumb	89.97×10^{-12}
Silverman's rule of thumb	121.5×10^{-12}
Single KDE (limited range)	1.873×10^{-12}
Single KDE (CV – large range)	1.867×10^{-12}
Ensemble	1.359×10^{-12}

Future Work

- Convergence toward specific higher order kernel?
- Algorithms for selecting bandwidth
- Compare against other modern KDE approaches
- Multivariate density estimation proof

References

- Sricharan, Kumar, Dennis Wei, and Alfred O. Hero. "Ensemble estimators for multivariate entropy estimation." *IEEE transactions on information theory* 59.7 (2013): 4374-4388.