Waiting Time Estimation in Nonstationary Queueing System

221870264 Shaozhe Ke
221870001 Zihao Ma

April 27, 2025

Background

- Arrival rate: $\lambda(t)$ changes over time
- Service times: Exponentially distributed with rate μ
- Goal: Estimate expected waiting time $E[W(t)]$ over time

k-Nearest Neighbors

Given a query time τ_0, the kNN waiting time estimator is:

$$\widehat{W}_{\mathrm{kNN}}(\tau_0) = \frac{1}{k} \sum_{i \in \mathcal{N}_k(\tau_0)} w_i$$

where $\mathcal{N}_k(\tau_0)$ denotes the set of k nearest arrival times to τ_0.

Algorithm implementation

Algorithm 1 kNN Estimation via LORO-CV

Input: Search range $k_L < k_U$, NN = "nearest neighbors."
for $j = 1, 2, \dots, n$ **do** ▷ Loop through all replications
  $S_{\text{test}} \leftarrow \{W(t_{ij}), t_{ij}; i = 1, 2, \dots, M_j\}$ ▷ Set aside one replication as test set
  $S_{\text{train}} \leftarrow$ all data except S_{test} ▷ Use remaining data as training set
  Find k_U nearest neighbors in S_{train} for each $t_{ij} \in S_{\text{test}}$
  Store indices of k_U nearest neighbors into $\mathbf{M}_{\text{ind}} \in \mathbb{R}^{|S_{\text{test}}| \times k_U}$
end for
for $k = k_L$ to k_U **do** ▷ Search for optimal k
  Extract the first k columns from $\mathbf{M}_{\text{ind}}$
  Find k-nearest neighbors for each $t_{ij} \in S_{\text{test}}$ and compute $\hat{W}(t_{ij}, k)$
end for
for $k = k_L$ to k_U **do** ▷ Compute the EMSE for each k
  Compute the empirical mean squared error (EMSE):

$$\text{EMSE}(k) = \frac{\sum_{j=1}^{n} \sum_{i=1}^{M_j} \left[W_{ij} - \hat{W}(t_{ij}, k)\right]^2}{\sum_{j=1}^{n} M_j}$$

end for
Output: k^* that minimizes $\text{EMSE}(k)$

KDE-Mode Clustering Estimator

First, estimate the arrival time density using KDE:

$$\widehat{p}_h(t) = \frac{1}{nh} \sum_{i=1}^{n} K\left(\frac{t - t_i}{h}\right)$$

Then, identify the cluster $C(\tau_0)$ containing τ_0, and compute:

$$\widehat{W}_{\text{KDE-mode}}(\tau_0) = \frac{1}{|C(\tau_0)|} \sum_{i \in C(\tau_0)} w_i$$

KDE-Mode Clustering

- Step 1: Estimate arrival time density $\widehat{p}_h(t)$ using KDE [Silverman, 1986].
- Step 2: Detect all modes (local maxima) $\{m_1, m_2, \ldots, m_K\}$.
- Step 3: Assign each arrival time t_i to the nearest mode:

$$m_k = \arg\min_k |t_i - m_k|$$

- Step 4: For a query τ_0, find its nearest mode $m^*(\tau_0)$.
- Step 5: Estimate $\widehat{W}_{\text{KDE-mode}}(\tau_0)$ by averaging waiting times within cluster $C(m^*(\tau_0))$.

Mean-Shift Clustering (Iterative Mode Seeking)

- Step 1: Estimate arrival time density $\widehat{p}_h(t)$ using KDE [Silverman, 1986].
- Step 2: Initialize each sample $x^{(0)} = t_i$.
- Step 3: Iteratively update according to Mean-Shift rule [Fukunaga and Hostetler, 1975]:

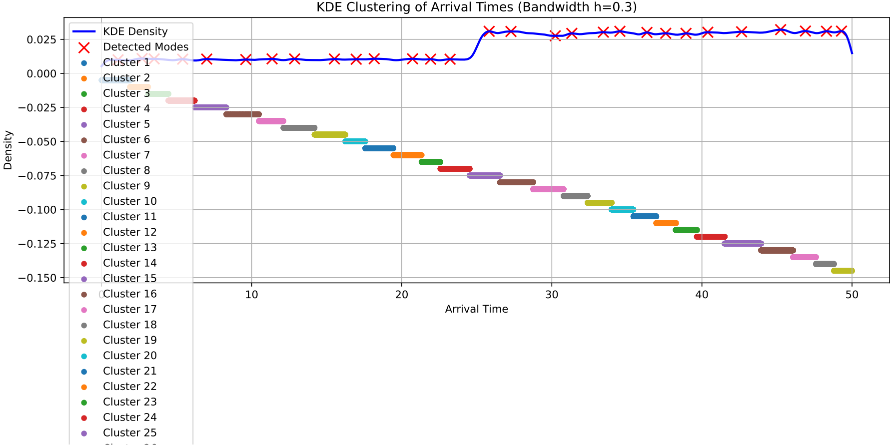
$$x^{(k+1)} = \frac{\sum_{i=1}^{n} K\left(\frac{x^{(k)}-t_i}{h}\right) t_i}{\sum_{i=1}^{n} K\left(\frac{x^{(k)}-t_i}{h}\right)}$$

  until convergence.
- Step 4: Each sample t_i converges to a local mode m_k.
- Step 5: Samples converging to the same mode form a cluster.

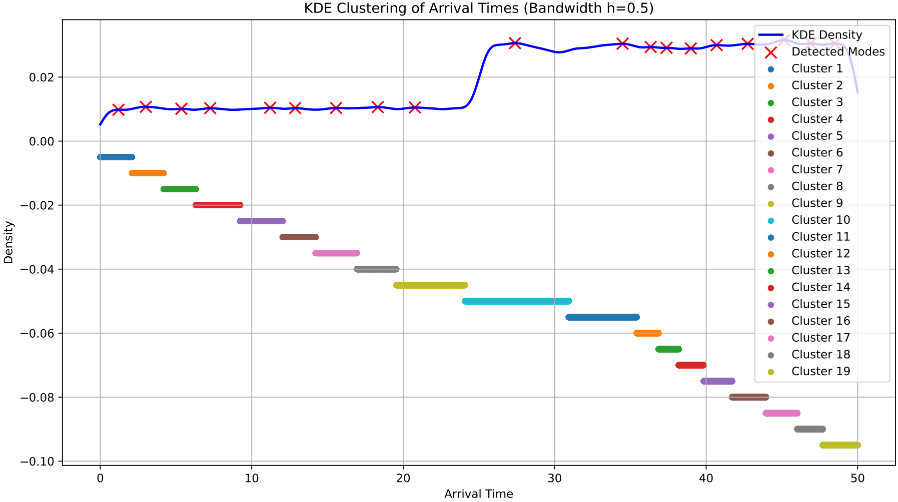
Remarks

KDE-Mode Clustering can be viewed as a non-iterative simplification of Mean-Shift clustering, where samples are directly assigned to pre-detected modes.

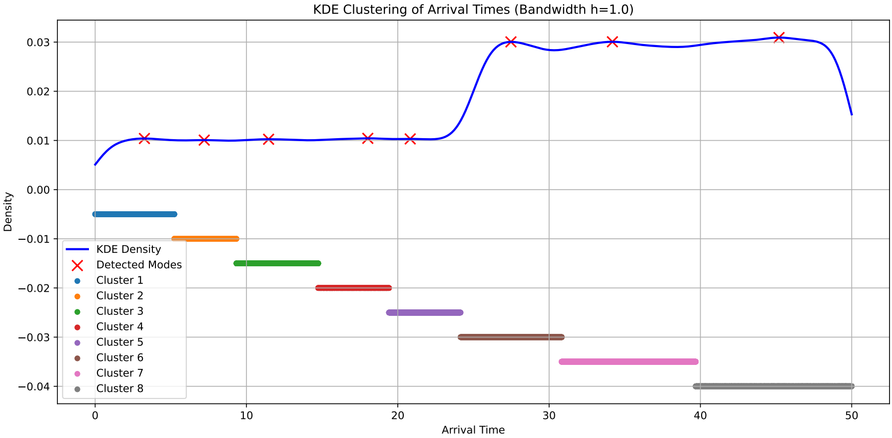
Bandwidth $h = 0.3$



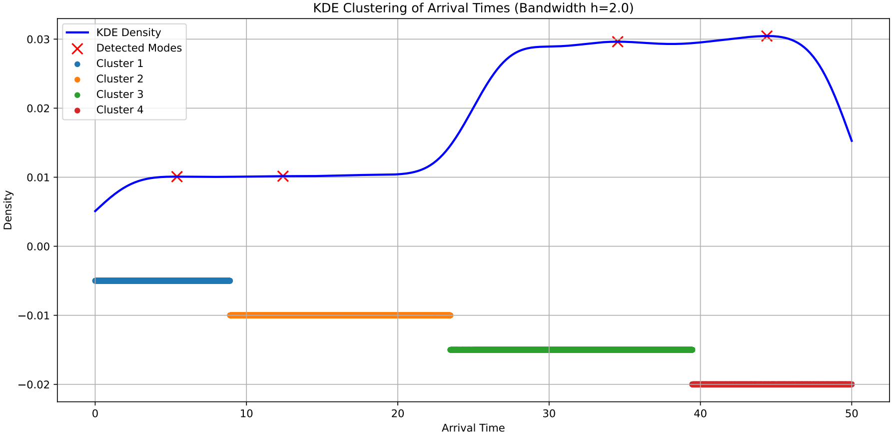
Bandwidth $h = 0.5$



Bandwidth $h = 1$



Bandwidth $h = 2$



Kolmogorov Forward Equations

For the nonstationary M(t)/M/1 queue:

For $n = 0$ (empty system):

$$\frac{d}{dt}p_0(t) = \mu p_1(t) - \lambda(t)p_0(t)$$

For $1 \leq n \leq N_{\text{max}} - 1$:

$$\frac{d}{dt}p_n(t) = \lambda(t)p_{n-1}(t) + \mu p_{n+1}(t) - (\lambda(t) + \mu)p_n(t)$$

For $n = N_{\text{max}}$ (maximum capacity):

$$\frac{d}{dt}p_{N_{\text{max}}}(t) = \lambda(t)p_{N_{\text{max}}-1}(t) - \mu p_{N_{\text{max}}}(t)$$

Computation of Expected Waiting Time

Once the state probabilities $p_n(t)$ are obtained by solving the Kolmogorov Forward Equations:

Step 1: Expected number of customers (queue length)

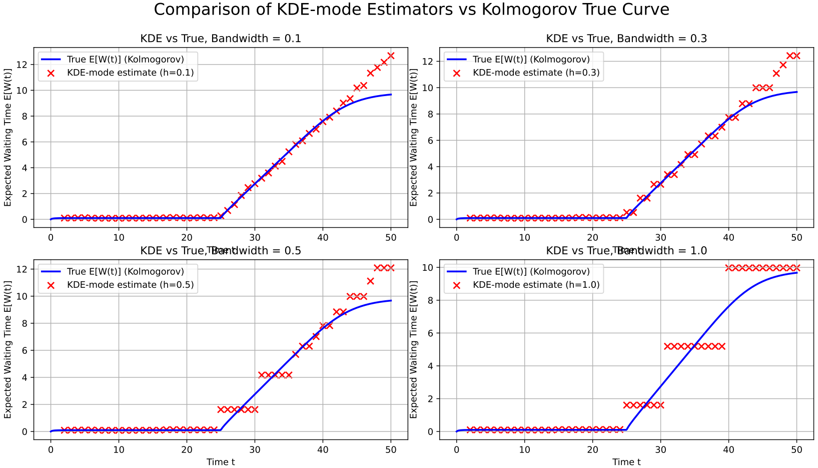
$$E[L(t)] = \sum_{n=0}^{N_{\max}} n \cdot p_n(t)$$

Step 2: Apply Little's Law

$$E[W(t)] = \frac{E[L(t)]}{\mu}$$

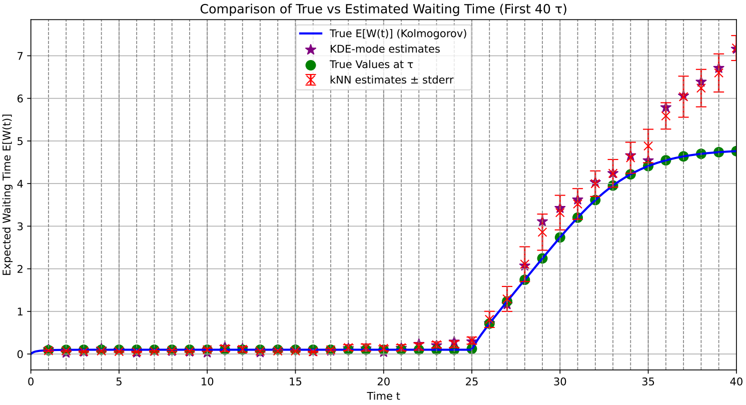
where μ is the constant service rate.

Results: E[W(t)] Comparison

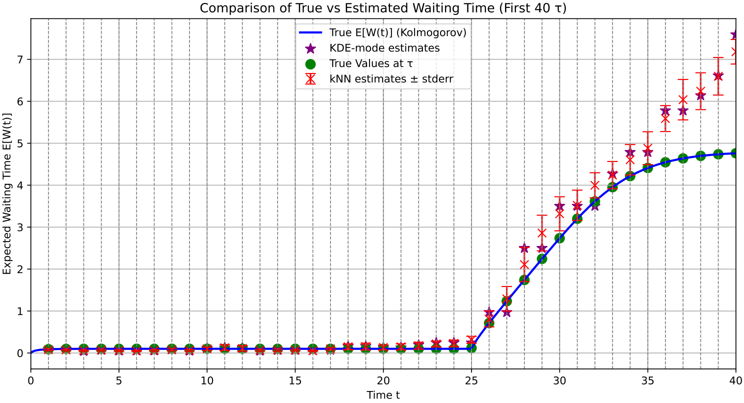


True $E[W(t)]$ vs KDE-mode estimates

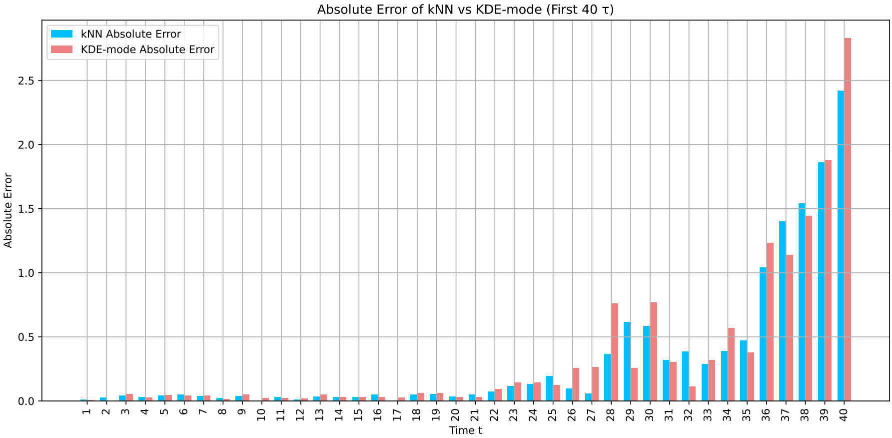
Results: kNN-KDE Comparison (bandwidth=0.1)



Results: kNN-KDE Comparison (bandwidth=0.3)



Results: kNN-KDE Comparison (bandwidth=0.3)



Erlang Arrival Process

- Erlang(k, λ) interarrival times: sum of k i.i.d. exponential(λ) random variables.
- PDF of interarrival time:

$$f_T(t) = \frac{\lambda^k t^{k-1} e^{-\lambda t}}{(k-1)!}, \quad t \geq 0$$

- Special case:
  - $k = 1$: reduces to standard Poisson process.
- Variability (coefficient of variation) decreases as k increases.

E_k(t)/M/1 Queue Model

- Customers arrive according to Erlang(k, λ) process.
- Service times are exponentially distributed with rate μ.
- State description:
  - n: number of customers in the system.
  - $s \in \{0, 1, \ldots, k-1\}$: current arrival phase.
- System state: (n, s).

Kolmogorov Forward Equations

Define $p_{n,s}(t) = P$system has n customers and phase s at time t.
Then, the equations are:

$$\frac{d}{dt} p_{n,s}(t) = (\text{inflow}) - (\text{outflow})$$

where:

- Phase transitions due to Erlang arrival stages.
- Customer arrivals when phase completes.
- Service completions.

Mean Queue Length

- Define:

$$L(t) = \text{number of customers in the system at time } t$$

- Then:

$$\mathbb{E}[L(t)] = \sum_{n=0}^{\infty} n \sum_{s=0}^{k-1} p_{n,s}(t)$$

- Differentiating $\mathbb{E}[L(t)]$ and using the forward equations leads to an ODE for $E[L(t)]$.

References I

Fukunaga, K. and Hostetler, L. (1975).
The estimation of the gradient of a density function, with applications in pattern recognition.
IEEE Transactions on Information Theory, 21(1):32–40.

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