Multivariate Local Region KDE: Initial Concept

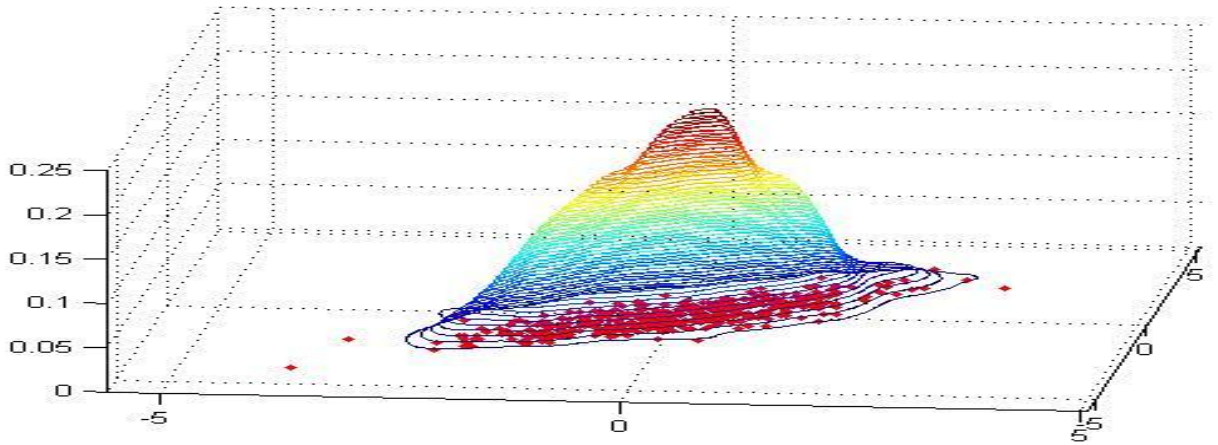
Arnold P. Boedihardjo

Outline

- Definition of multivariate KDE
- Mathematical formulation
- Issue with local region approach
- Possible solution…

What is the multivariate KDE?

- In the univariate setting, KDE is essentially a process of imposing a weighting function (i.e., kernel) to each available data sample in a single dimension R^1
- In the multivariate setting, the KDE is a process of imposing a kernel function to each data sample in the multiple dimensions R^d



Mathematical Formulation of the Multivariate KDE

$$f(x) = \frac{1}{N}\sum_{i=1}^{N}\prod_{j}^{d} h_j^{-1} K\left(\frac{x_j - X_{ij}}{h_j}\right)$$

- Employs product a kernel – multiplication of univariate Kernel functions
- Other multivariate kernel forms exists, such as radial kernels; however our focus will be on product kernels since it is the predominant approach in current applications
- The bandwidths are global within each dimension

Challenges of Local Region based Approach in Multivariate Setting

- Proposed LR-KDE employed a data clustering scheme that was not readily amenable to multivariate data sets
  - Numerical method is suggested to determine the minimum kernel merge loss
  - Potentially high computational cost
- LR-KDE also employed PAD criterion which can also be costly to compute in multidimension
- However, the recently developed General Local rEgion AlgorithM (GLEAM) KDE employs an SSE based criterion (fast algorithms exist for multidimension) and allows the use of any data maintenance strategy and hence not limited to costly clustering based methods
  - Employ a rapid subsampling or grid based data maintenance scheme

Extending GLEAM to Multivariate Setting

- Key idea: induce the local regions to one dimension at a time and apply product kernels!
- How? Simple. See below

$$f(x) = \frac{1}{N}\sum_{i=1}^{N}\prod_{j}^{d} H_j^{-1}(X_{ij}) \cdot K\left(\frac{x_j - X_{ij}}{H_j(X_{ij})}\right)$$

$$H_j^{-1}(x) = h(j,x) | x \cap \{l \in L_j\} \neq \emptyset$$

Remarks

- Avoids DIRECT construction of multidimensional LR
- It has been theoretically shown that Local Regions produce higher variance than standard KDE that is proportional to the ratio of the global spread and product of the spreads of each local regions
- Variance also increases as dimension increases if the number of data samples is held fixed
- Limit LR to low dimensions and provide a more suitable regularization function for multidimension
- Future: perform simple experiments.. Would like to use LR for spatiotemporal outlier detection within sensor networks