Adventures in Kernel Density Estimation

Clay Scott

joint with

JooSeuk Kim, Robert Vandermeulen, and
Efrén Cruz Cortés

ELECTRICAL &
COMPUTER ENGINEERING

UNIVERSITY OF MICHIGAN

Kernel Density Estimation

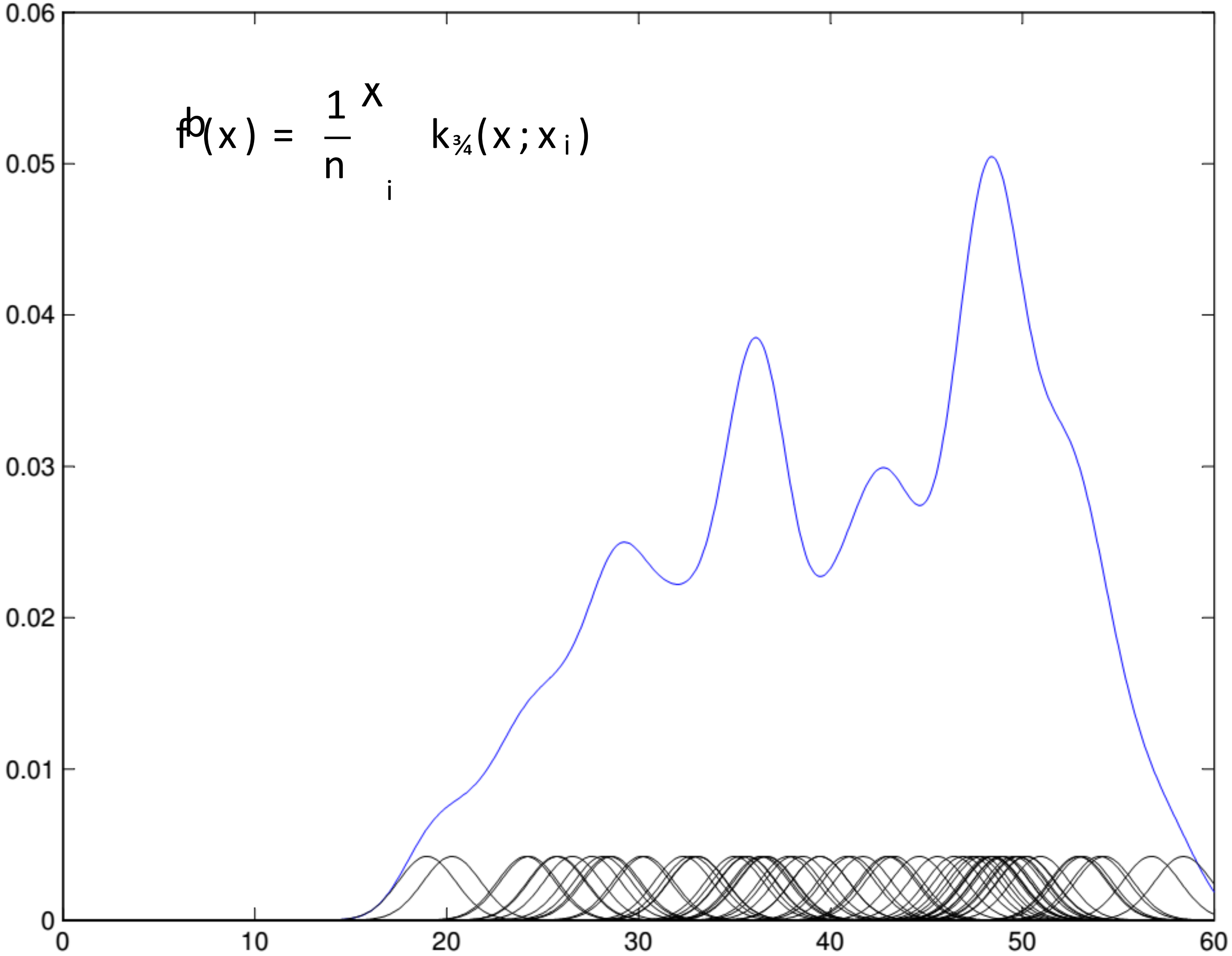
- $\mathbf{x}_1, \ldots, \mathbf{x}_n \overset{iid}{\sim} f$ (unknown density), $\mathbf{x}_i \in \mathbb{R}^d$
- Estimate f via

$$\widehat{f}(\mathbf{x}) = \frac{1}{n} \sum_i k_\sigma(\mathbf{x}, \mathbf{x}_i)$$

- Example: Gaussian kernel

$$k_\sigma(\mathbf{x}, \mathbf{x}') = (2\pi\sigma^2)^{-d/2} \exp\left(-\frac{\|\mathbf{x} - \mathbf{x}'\|^2}{2\sigma^2}\right)$$

Kernel Density Estimation



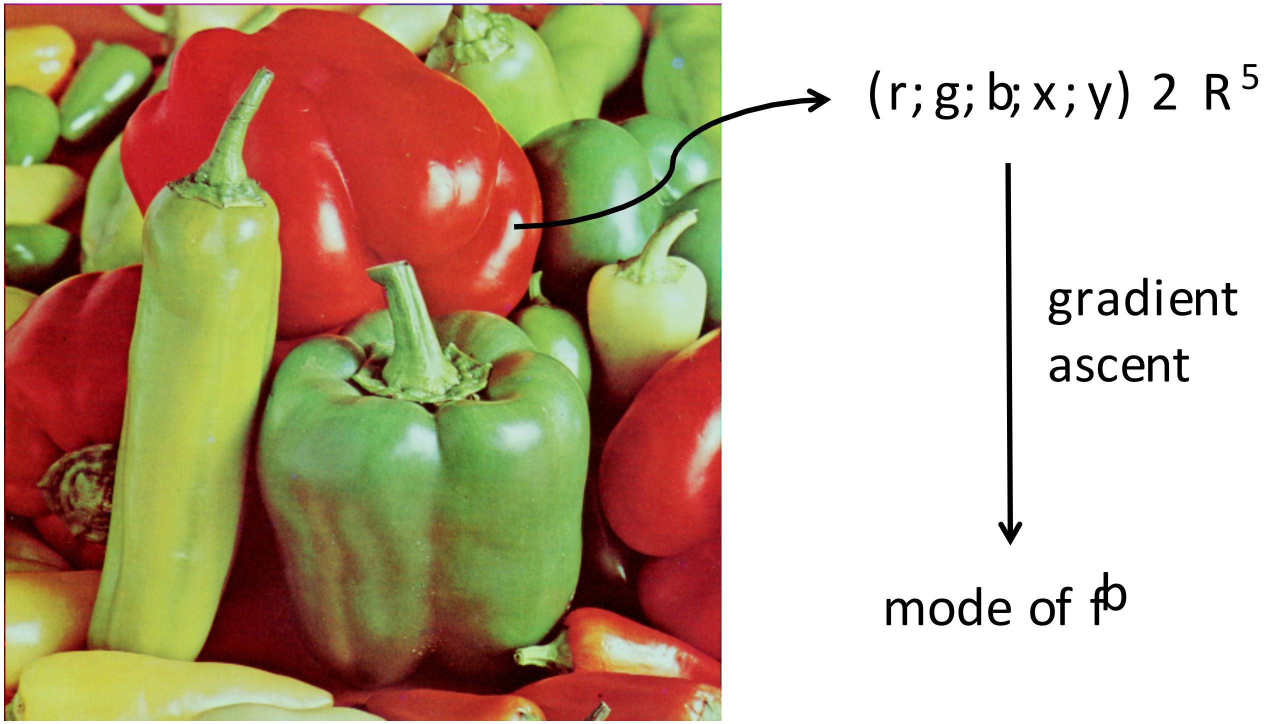
$$\hat{f}(x) = \frac{1}{n}\sum_{i} k_{\sigma}(x;x_i)$$

Applications of KDEs

Predictors based on KDEs: Classification, regression, anomaly detection

Estimates of information theoretic measures (entropy, KL divergence)

Clustering (e.g., image segmentation) via mean-shift algorithm



$(r; g; b; x; y) \in \mathbb{R}^5$

gradient ascent

mode of $\widehat{f}$

Conventional Analysis

² Estimation and approximation errors

kfb¡ f k · kfb¡ f ¤ k¾k + kf ¤ k¾ ¡ f k

² Note

fb = 1/n X_i k(¢ x_i)

= k convolved with empirical distribution

² Need ¾ ! 0 for approximation error to vanish

² Need n¾^d ! 1 for estimation error to vanish (by properties of convolutions and concentration inequalities)

Outline

Topics

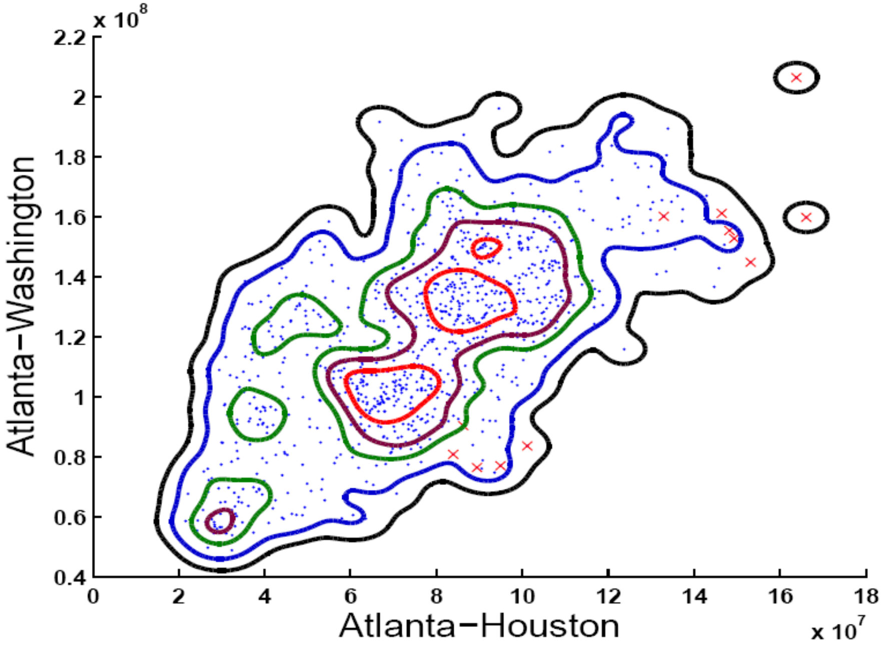
- Robust KDEs
- Sparse KDEs
- Consistency with fixed bandwidth

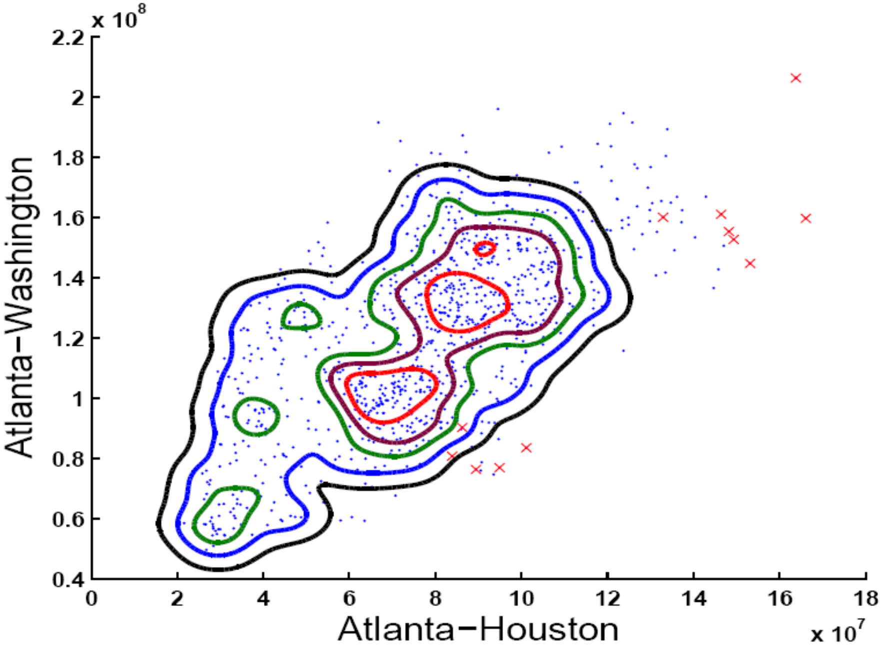
Themes

- KDE = mean in a function space
- Weighted KDEs to achieve above goals

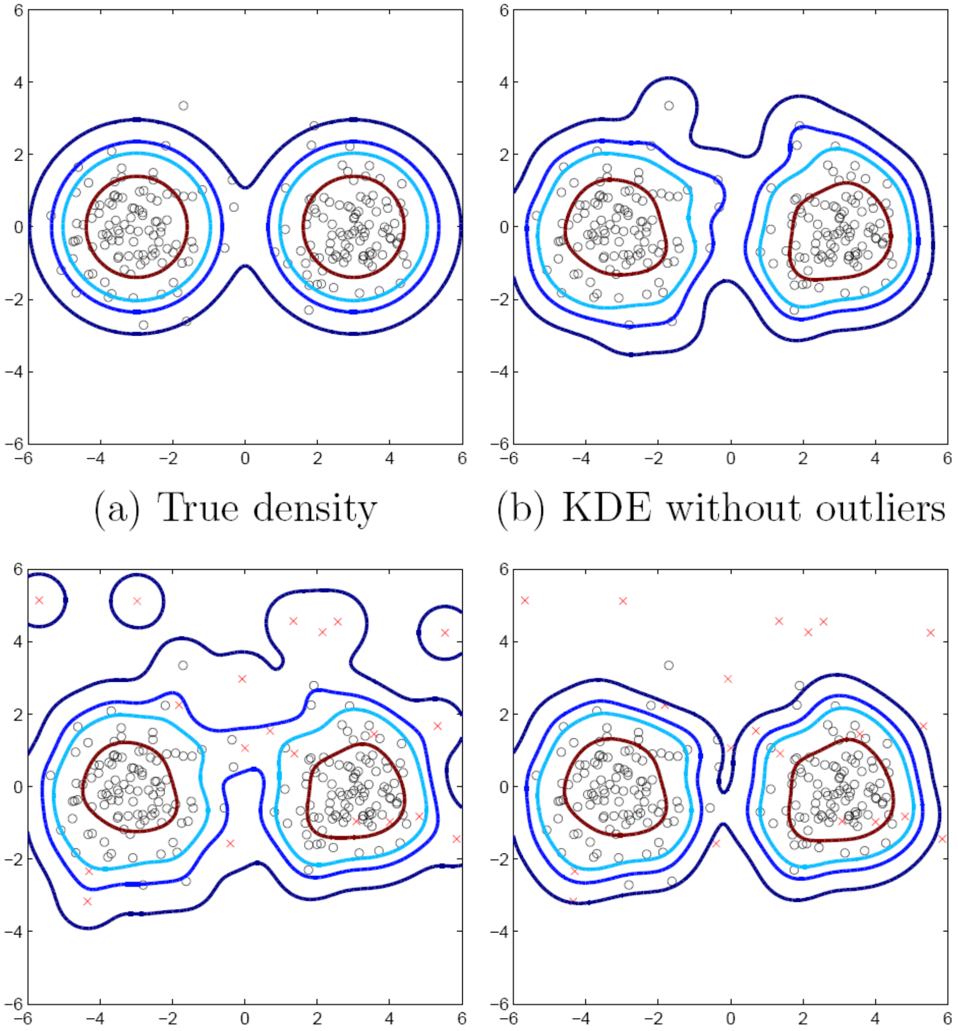
Robust KDE for Anomaly Detection

Abilene network traffic volumes





Synthetic Example



(a) True density

(b) KDE without outliers

(c) KDE with outliers

(d) RKDE with outliers

Problem Statement

- $\mathbf{X}_1, \ldots, \mathbf{X}_n \sim f(\mathbf{x}) = (1-\epsilon) f_0(\mathbf{x}) + \epsilon f_1(\mathbf{x})$
- Tasks
  - ○ Estimate $f_0(\mathbf{x})$
  - ○ Estimate $\{\mathbf{x} : f_0(\mathbf{x}) > \lambda\}$

Kernel Density Estimate

- $$\widehat{f}_{KDE}\left(\mathbf{x}\right) = \frac{1}{n}\sum_{i=1}^{n} k_{\sigma}\left(\mathbf{x}, \mathbf{X}_i\right)$$

- Gaussian kernel

$$k_{\sigma}\left(\mathbf{x}, \mathbf{x}'\right) = \left(\frac{1}{\sqrt{2\pi}\sigma}\right)^{d} \exp\left(-\frac{\left\|\mathbf{x} - \mathbf{x}'\right\|^2}{2\sigma^2}\right)$$

Gaussian RKHS

- There exists a **Hilbert space** $\mathcal{H}_\sigma$ and a **feature map** $\Phi_\sigma : \mathbb{R}^d \to \mathcal{H}_\sigma$ such that

$$k_\sigma\left(\mathbf{x}, \mathbf{x}'\right) = \left\langle \Phi_\sigma(\mathbf{x}), \Phi_\sigma(\mathbf{x}') \right\rangle_{\mathcal{H}_\sigma}$$

- Canonical feature map

$$\Phi_\sigma(\mathbf{x}) = k_\sigma(\cdot, \mathbf{x})$$

- Reproducing property

$$\forall\ g \in \mathcal{H}_\sigma, \quad g(\mathbf{x}) = \langle \Phi_\sigma(\mathbf{x}), g \rangle_{\mathcal{H}_\sigma}$$

- $\|\Phi_\sigma(\mathbf{x})\|^2 = k_\sigma(\mathbf{x}, \mathbf{x}) = (\sqrt{2\pi}\sigma)^{-d}$

KDE = mean in RKHS

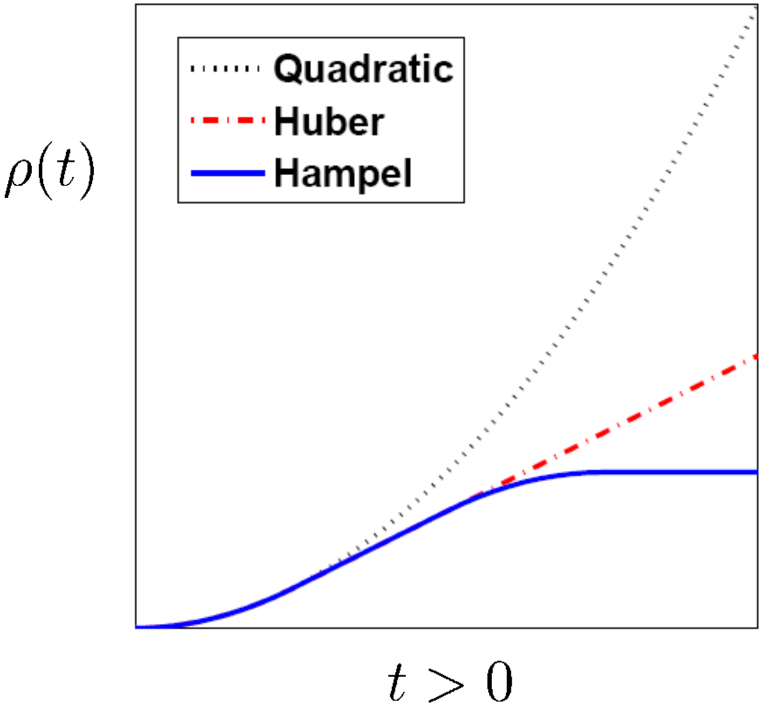
- $$\begin{aligned}\widehat{f}_{KDE} &= \frac{1}{n}\sum_{i=1}^{n} k_\sigma\left(\cdot, \mathbf{X}_i\right) \\ &= \frac{1}{n}\sum_{i=1}^{n} \Phi_\sigma(\mathbf{X}_i)\end{aligned}$$

- Idea: Estimate this mean **robustly**

Robust Kernel Density Estimate

$$\widehat{f}_{KDE} = \operatorname*{arg\,min}_{g \in \mathcal{H}_\sigma} \sum_{i=1}^{n} \|\Phi_\sigma(\mathbf{X}_i) - g\|^2_{\mathcal{H}_\sigma}$$

$$\widehat{f}_{RKDE} = \operatorname*{arg\,min}_{g \in \mathcal{H}_\sigma} \sum_{i=1}^{n} \rho\big(\|\Phi_\sigma(\mathbf{X}_i) - g\|_{\mathcal{H}_\sigma}\big)$$



Representer Theorem

- Recall

$$\widehat{f}_{RKDE} = \arg\min_{g \in \mathcal{H}_\sigma} \sum_{i=1}^{n} \rho\big(\|\Phi_\sigma(\mathbf{X}_i) - g\|_{\mathcal{H}_\sigma}\big)$$

- **Theorem**: If ρ satisfies certain common assumptions, then

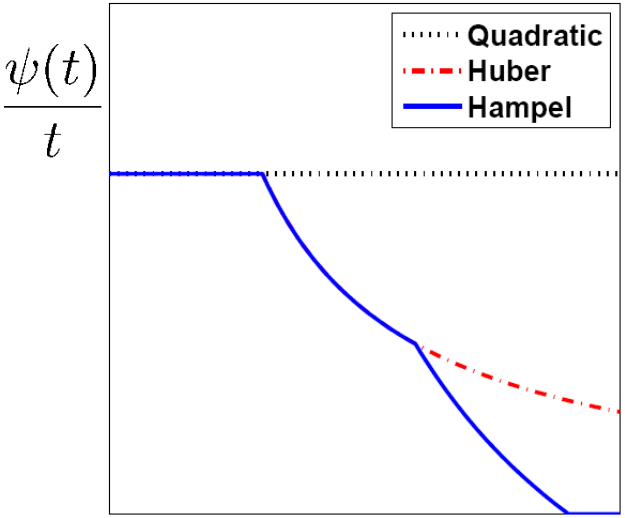
$$\widehat{f}_{RKDE}(\mathbf{x}) = \sum_{i=1}^{n} w_i k_\sigma(\mathbf{x}, \mathbf{X}_i)$$

for some $w_i \geq 0$, $\sum_{i=1}^{n} w_i = 1$.

- Furthermore

$$w_i \propto \frac{\psi(\|\Phi_\sigma(\mathbf{X}_i) - \widehat{f}_{RKDE}\|)}{\|\Phi_\sigma(\mathbf{X}_i) - \widehat{f}_{RKDE}\|}$$

where $\psi = \rho'$.



Robustness Interpretation

- Notice that

$$\begin{aligned}\|\Phi_\sigma(\mathbf{x}) - \widehat{f}\|^2_{\mathcal{H}_\sigma} &= \langle \Phi_\sigma(\mathbf{x}) - \widehat{f}, \Phi_\sigma(\mathbf{x}) - \widehat{f}\rangle_{\mathcal{H}_\sigma} \\ &= \|\Phi_\sigma(\mathbf{x})\|^2_{\mathcal{H}_\sigma} - 2\langle \Phi_\sigma(\mathbf{x}), \widehat{f}\rangle_{\mathcal{H}_\sigma} + \|\widehat{f}\|^2_{\mathcal{H}_\sigma} \\ &= (\sqrt{2\pi}\sigma)^{-d} - 2\widehat{f}(\mathbf{x}) + \|\widehat{f}\|^2_{\mathcal{H}_\sigma}\end{aligned}$$

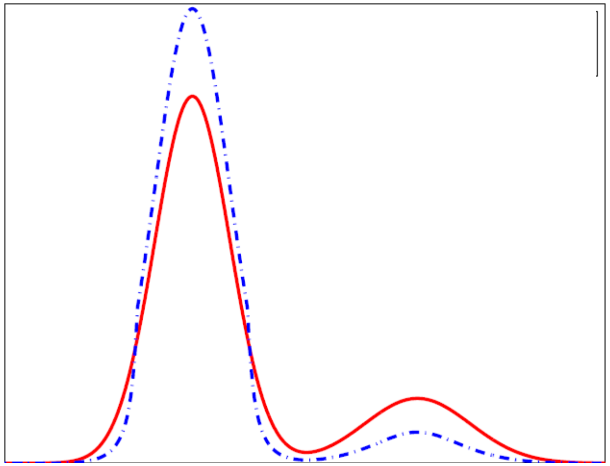
- **Conclusion:**

$$\begin{aligned}w_i \text{ is small} &\iff \|\Phi_\sigma(\mathbf{X}_i) - \widehat{f}_{RKDE}\| \text{ is large} \\ &\iff \widehat{f}_{RKDE}(\mathbf{X}_i) \text{ is small}\end{aligned}$$

- RKDE **down-weights** outlying points

Other Results

- Efficient algorithm: iterative reweighted least squares (converges to global or local optimum depending on whether ρ is convex)
- Influence function: also reveals robustness to outliers relative to standard KDE
- Consistency
  - Convex ρ: converges to density of data (same as KDE)
  - Nonconvex ρ: converges to transformed version of density of data; this equals the uncontaminated density under certain assumptions on the contamination
- Experimental validation



Sparse Approximation of Kernel Means

² Sparse kernel mean:

$$\underset{i}{X} \; ®_i Á(¢; x_i) \; ¼ \; \frac{1}{n} \underset{i}{X} \; Á(¢; x_i)$$

where $jf i : ®_i 6= 0gj \cdot k$

² Major complexity gains

± Evaluation of kernel density estimate: $O(n)$ ¡ ! $O(k)$

± Mean shift clustering: $O(n^2)$ ¡ ! $O(nk)$

² Set z_i := $Á(¢; x_i)$ 2 V

± V = $L^2(R^d)$ for density estimation, or

± V = RKHS of Á, if Á is a reproducing kernel

Sparse Approximation of a Sample Mean

Setting

² inner product space (V; h¢; ¢i)

² $z_1; : : : ; z_n$ 2 V, ½ = $\frac{1}{n}$ P $_i$ z_i

² desired sparsity k ¿ n

² ® = (®$_1$; : : : ; ®$_n$)

Goal: (approximately) solve

$$\min_{\text{® 2 R}^n} \text{ ° ° ½ ¡ P}_i \text{ ®}_i z_i \text{ ° °}_V$$

s.t. k®k$_0$ · k

using an algorithm with O(nk) time and space complexity

Existing SA Methods Too Slow

Matching pursuit:

- Requires computing $\langle \bar{z}, z_i \rangle$ for each i
- Each $\langle \bar{z}, z_i \rangle = \frac{1}{n} \sum_j \langle z_j, z_i \rangle$ requires $O(n)$ operations
- $\Rightarrow$ $O(n^2)$ complexity

Problem Simplification

- Denote $I \subseteq [n] := \{1, 2, \ldots, n\}$
- Equivalent formulations:

$$\min_{\alpha \in \mathbb{R}^n} \left\| \rho - \sum_i \alpha_i z_i \right\|^2$$

$$\text{s.t. } \|\alpha\|_0 \le k$$

$$\Updownarrow$$

$$\min_{\substack{I \subseteq [n] \\ |I| = k}} \; \min_{(\alpha_i)_{i \in I}} \left\| \rho - \sum_{i \in I} \alpha_i z_i \right\|^2$$

$$\Updownarrow$$

$$\min_{\substack{I \subseteq [n] \\ |I| = k}} \left\| \rho - z_I \right\|^2 .$$

- where z_I is the projection of ρ onto $\mathrm{span}\{z_i | i \in I\}$

Incoherence-Based Bound

Theorem: Assume $\langle z_i, z_i \rangle = C$ for all i. For any $I \subseteq [n]$,

$$\|\bar{z} - z_I\| \leq \left(1 - \frac{|I|}{n}\right) \sqrt{\frac{1}{C}(C^2 - \nu_I^2)}.$$

where

$$\nu_I := \min_{j \notin I} \max_{i \in I} \langle z_i, z_j \rangle$$

measures the incoherence of $\{z_i : i \in I\}$

Bound Minimization

² For most kernels of interest,

$$\begin{aligned} hz_i;z_j i &= h\text{Á}(\text{¢};x_i);\text{Á}(\text{¢};x_j)i \\ &= g(kx_i \text{ i } x_j k) \end{aligned}$$

for some strictly decreasing g

² Example: Gaussian kernel, V = RKHS

$$\begin{aligned} h\text{Á}(\text{¢};x_i);\text{Á}(\text{¢};x_j)i &= \text{Á}(x_i;x_j) \\ &= (2\text{¼¾}^2)^{\text{i } d=2} \exp \text{µ} \text{ i } \frac{kx_i \text{ i } x_j k^2}{2\text{¾}^2} \text{¶} \end{aligned}$$

Bound Minimization = k-center problem

- For most kernels of interest,

$$\begin{aligned} \langle z_i, z_j \rangle &= \langle \phi(\cdot, x_i), \phi(\cdot, x_j) \rangle \\ &= g(\|x_i - x_j\|) \end{aligned}$$

for some strictly decreasing g

- Bound minimization

$$\begin{aligned} I_k^* &= \arg\max_{\substack{I \subseteq [n] \\ |I| = k}} \nu_I \\ &= \arg\max_{\substack{I \subseteq [n] \\ |I| = k}} \min_{j \notin I} \max_{i \in I} g(\|x_i - x_j\|) \\ &= \arg\min_{\substack{I \subseteq [n] \\ |I| = k}} \max_{j \notin I} \min_{i \in I} \|x_i - x_j\| \end{aligned}$$

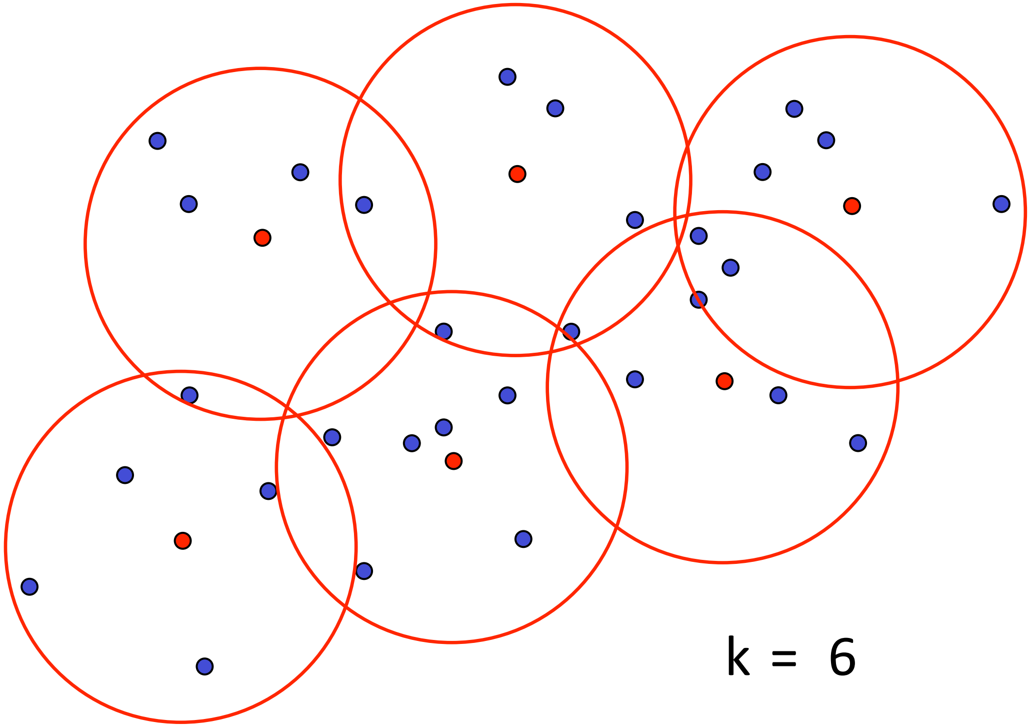
$\Rightarrow$ k-center problem

k-center problem and algorithm

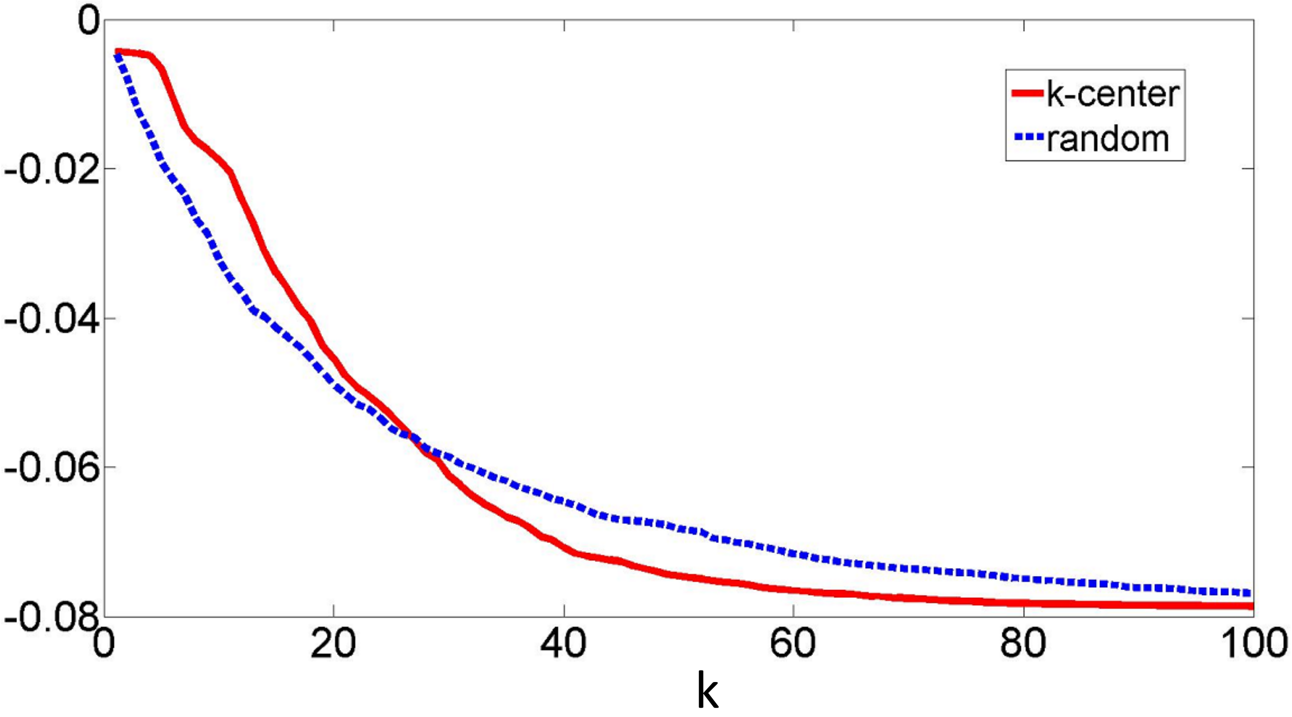
² Given n cities, place warehouses in k cities to minimize the maximum distance of a city to the nearest warehouse

² NP-Complete

² Greedy O(nk) 2-approximation algorithm



Example



Weighted KDEs for Density Estimation

- Estimate f via

$$\hat{f}_{\alpha}(x) = \sum_{i} \alpha_i k_{\sigma}(x; x_i)$$

- Theoretical question: Can we establish consistency (and rates) with σ fixed as $n \to 0$?

- Empirical question: can we get better estimates in finite sample settings?

Estimation Method

$$\|\hat{f}_\alpha - f\|_{L_2}^2 = \int (\hat{f}_\alpha(x) - f(x))^2 dx$$

$$= \int \hat{f}_\alpha(x)^2 dx - 2\int \hat{f}_\alpha(x) f(x) dx + \int f(x)^2 dx$$

First term:

$$\int \hat{f}_\alpha(x)^2 dx = \sum_i \sum_j \alpha_i \alpha_j \int k_\sigma(x; x_i) k_\sigma(x; x_j) dx$$

$$= \alpha^T Q \alpha$$

Second term:

$$\int \hat{f}_\alpha(x) f(x) dx = \sum_i \alpha_i \left(\int k_\sigma(x; x_i) f(x) dx \right)$$

$$\approx \sum_i \alpha_i \left[\frac{1}{n-1} \sum_{j \neq i} k_\sigma(x_j; x_i) dx \right]$$

$$= r^T \alpha$$

Fixed-Bandwidth KDE

Objective function:

$$J(\alpha) = \alpha^T Q \alpha - 2\mathbf{r}^T \alpha$$

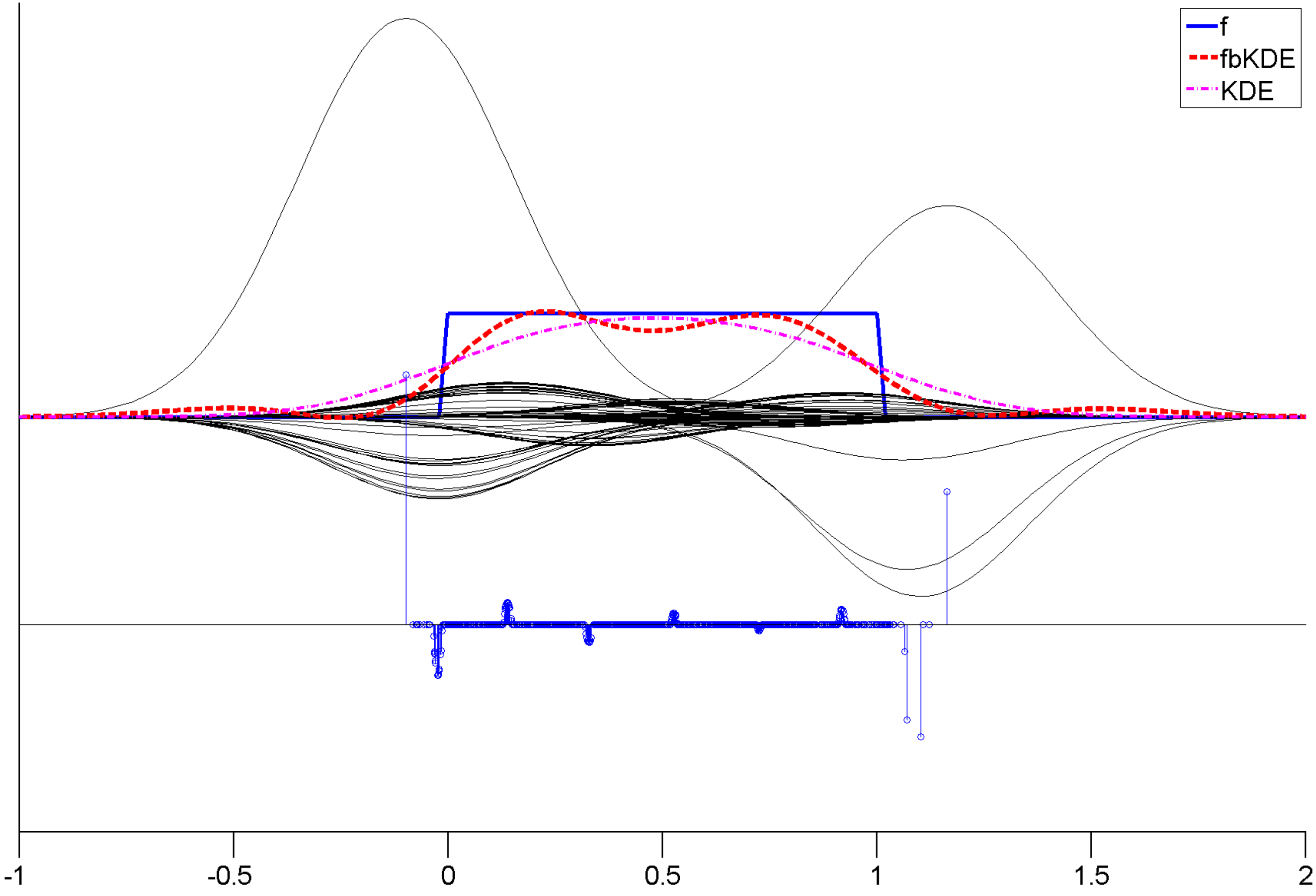
Optimization problem:

$$\begin{aligned} \widehat{\alpha} \longleftarrow & \mathop{\arg\min}_{\alpha \in \mathbb{R}^n} \;\; J(\alpha) \\ & \text{s.t.} \;\; \|\alpha\|_1 \leq R \end{aligned}$$

Density estimator:

$$\widehat{f}(\mathbf{x}) = \sum_i \widehat{\alpha}_i k_\sigma(\mathbf{x}, \mathbf{x}_i)$$

Example – Uniform Density



Oracle Inequality

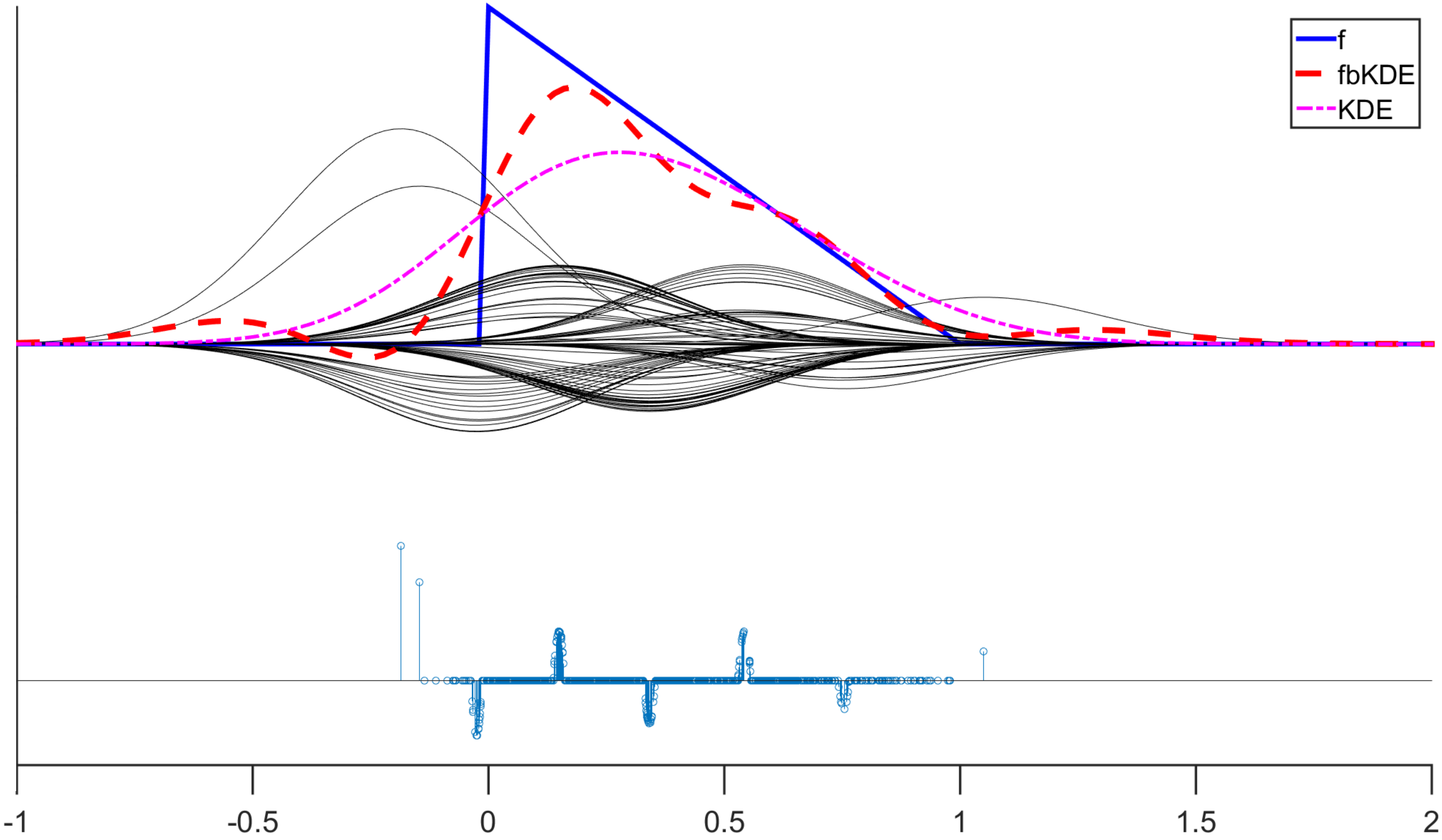
- There is a constant C such that with probability at least $1-\delta$

$$\|\hat{f}-f\|_{L_2}^2 \le \inf_{\alpha:\|\alpha\|_1\le R}\|f_\alpha - f\|_{L_2}^2 + CR\sqrt{\frac{\log(1/\delta)}{n}}$$

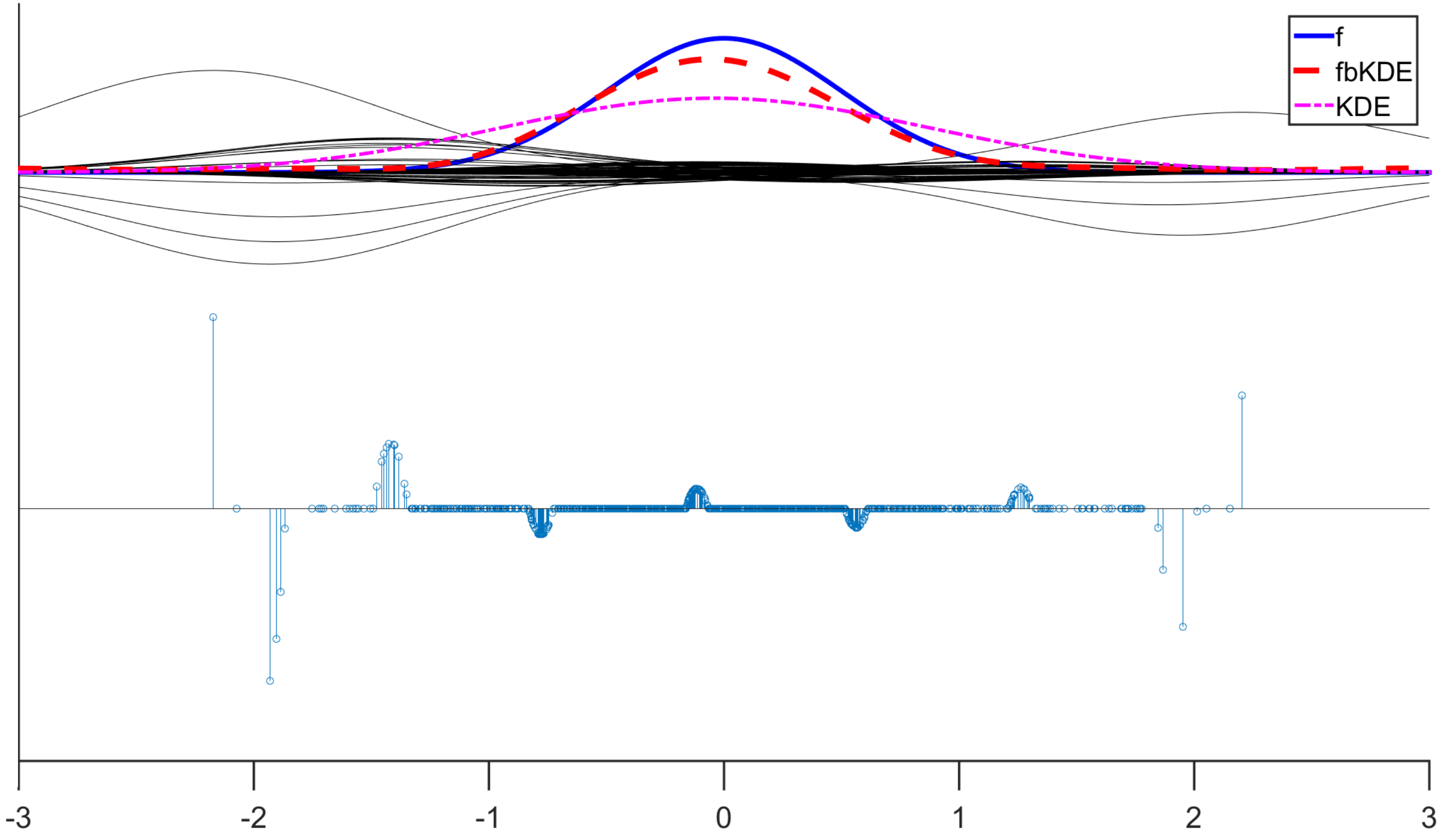
  (ignoring $\log n$ terms).

- Can be used to deduce consistency, rates of convergence with fixed σ.

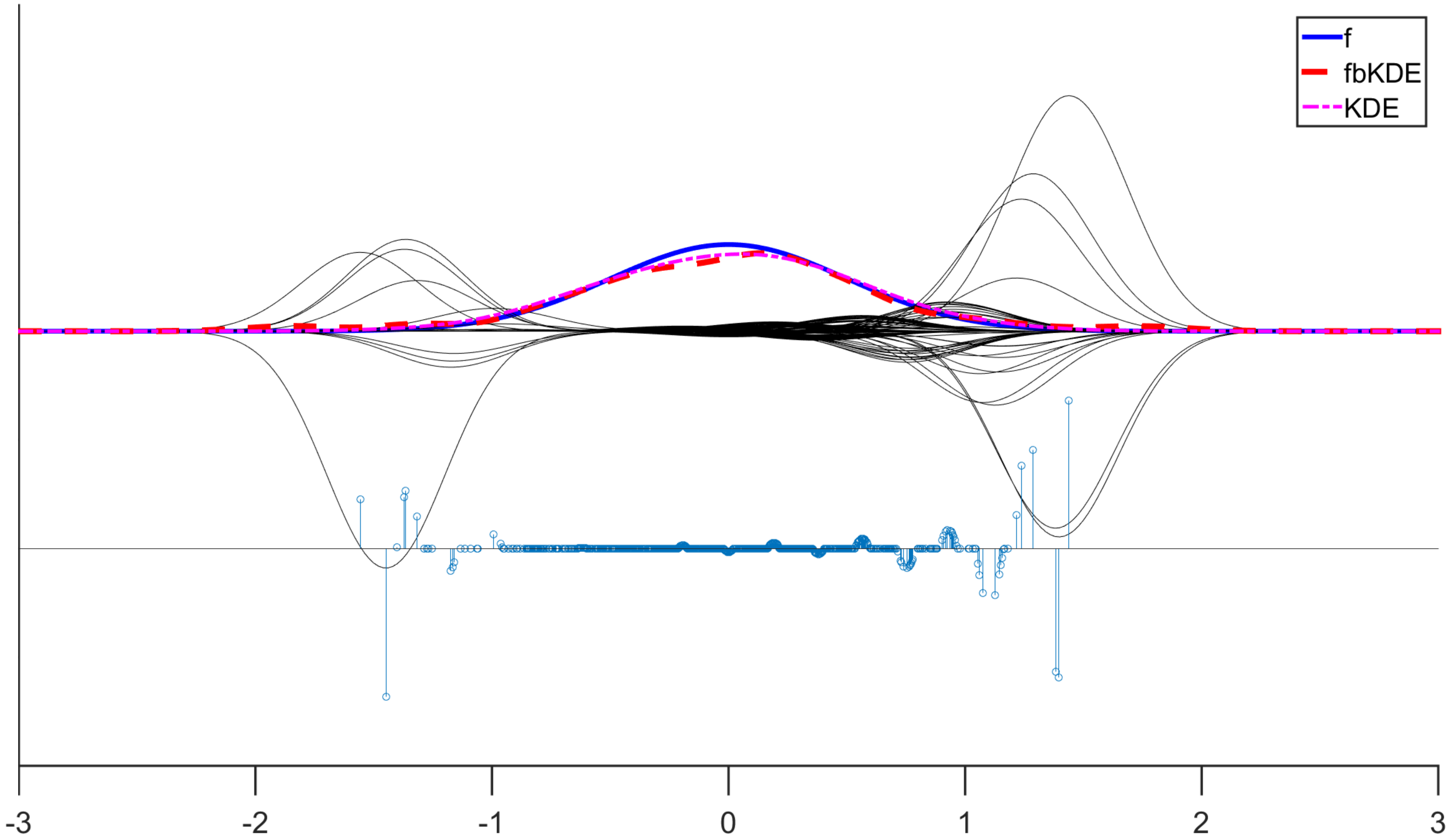
Triangular Density



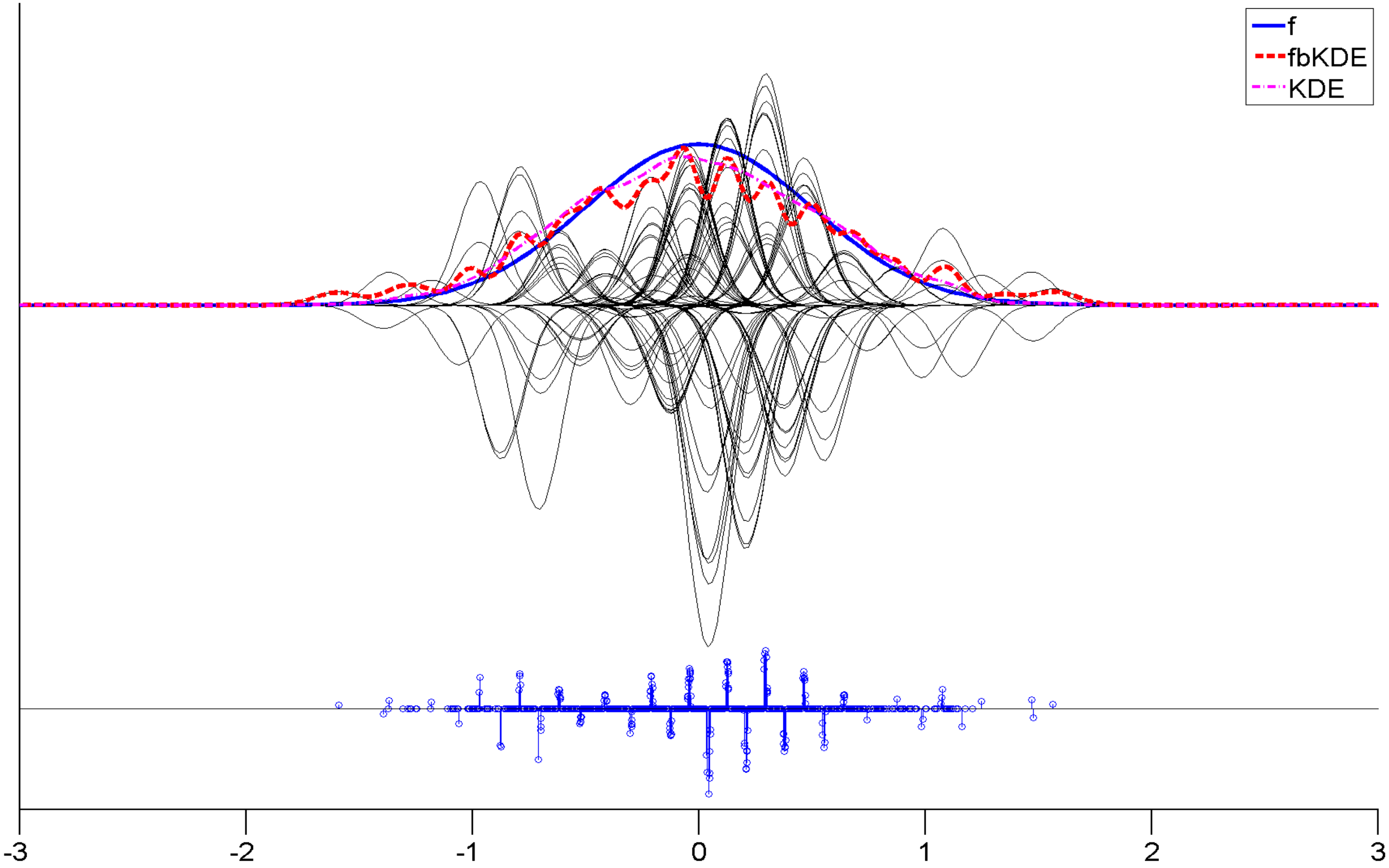
Gaussian Density – Bandwidth Too Large



Gaussian Density – Bandwidth Just Right



Gaussian Density – Not Enough Regularization



Summary

Topics

- Robust KDEs
- Sparse KDEs
- Consistency with fixed bandwidth

Themes

- KDE = mean in a function space
- Weighted KDEs to achieve above goals