Robust Kernel Density Estimation

Clayton Scott

EECS and Statistics
University of Michigan

Problem Statement

- $\mathbf{X}_1, \ldots, \mathbf{X}_n \sim f(\mathbf{x}) = (1-\epsilon) f_{\text{norm}}(\mathbf{x}) + \epsilon f_{\text{anom}}(\mathbf{x})$
- Tasks
  - Estimate $f_{\text{norm}}(\mathbf{x})$
  - Estimate $\{\mathbf{x} : f_{\text{norm}}(\mathbf{x}) > \lambda\}$

Kernel Density Estimate

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

- Gaussian kernel

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

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

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

- 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

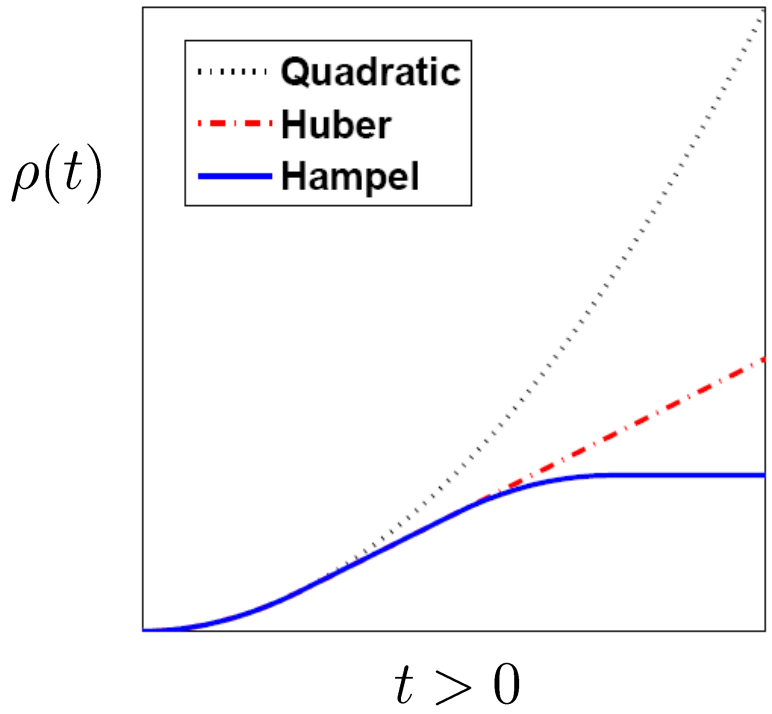
- $$\widehat{f}_{KDE} = \frac{1}{n}\sum_{i=1}^{n} k_\sigma\left(\cdot, \mathbf{X}_i\right)$$
$$= \frac{1}{n}\sum_{i=1}^{n} \Phi_\sigma(\mathbf{X}_i)$$

- Idea: Estimate this mean **robustly**

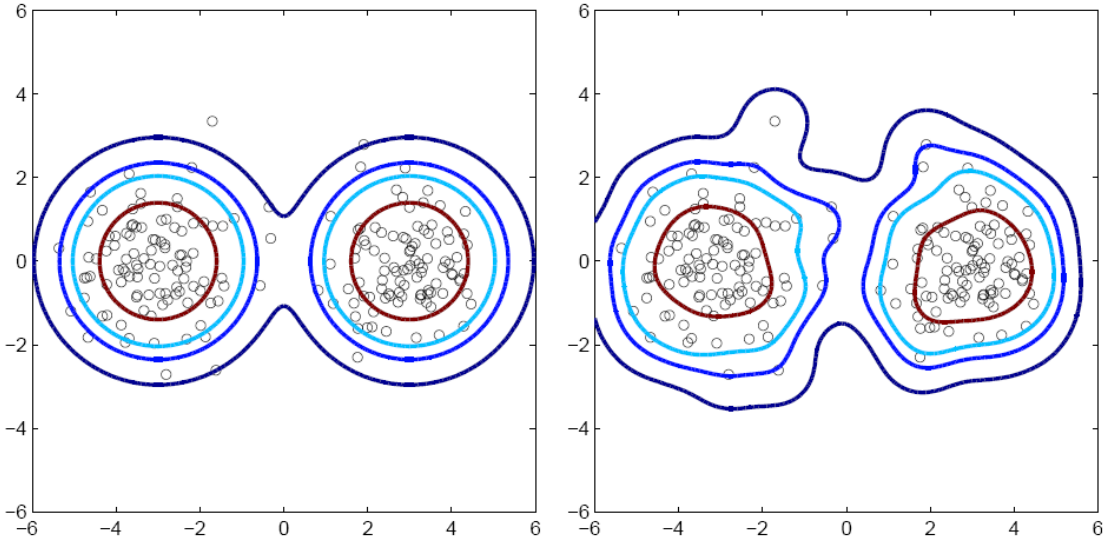
Robust Kernel Density Estimate

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

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

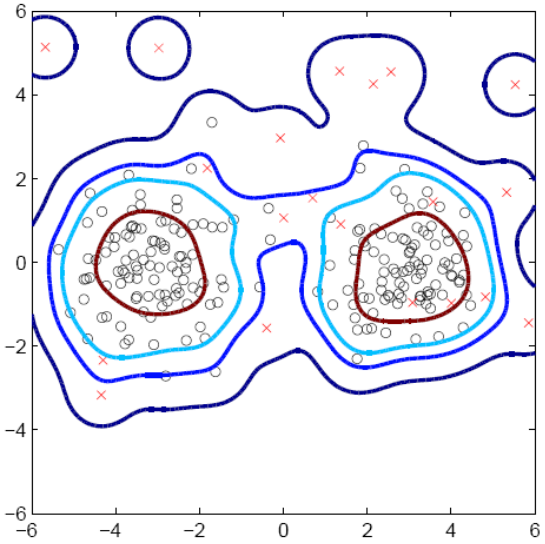


Example

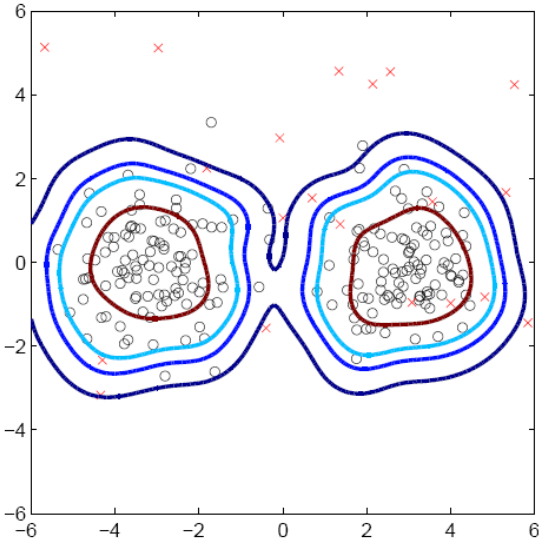


(a) True density

(b) KDE without outliers



(c) KDE with outliers



(d) RKDE with outliers

Outline

- Algorithm
- Influence Function
- Asymptotics
- Experiments: Anomaly Detection

Robust Multivariate Mean

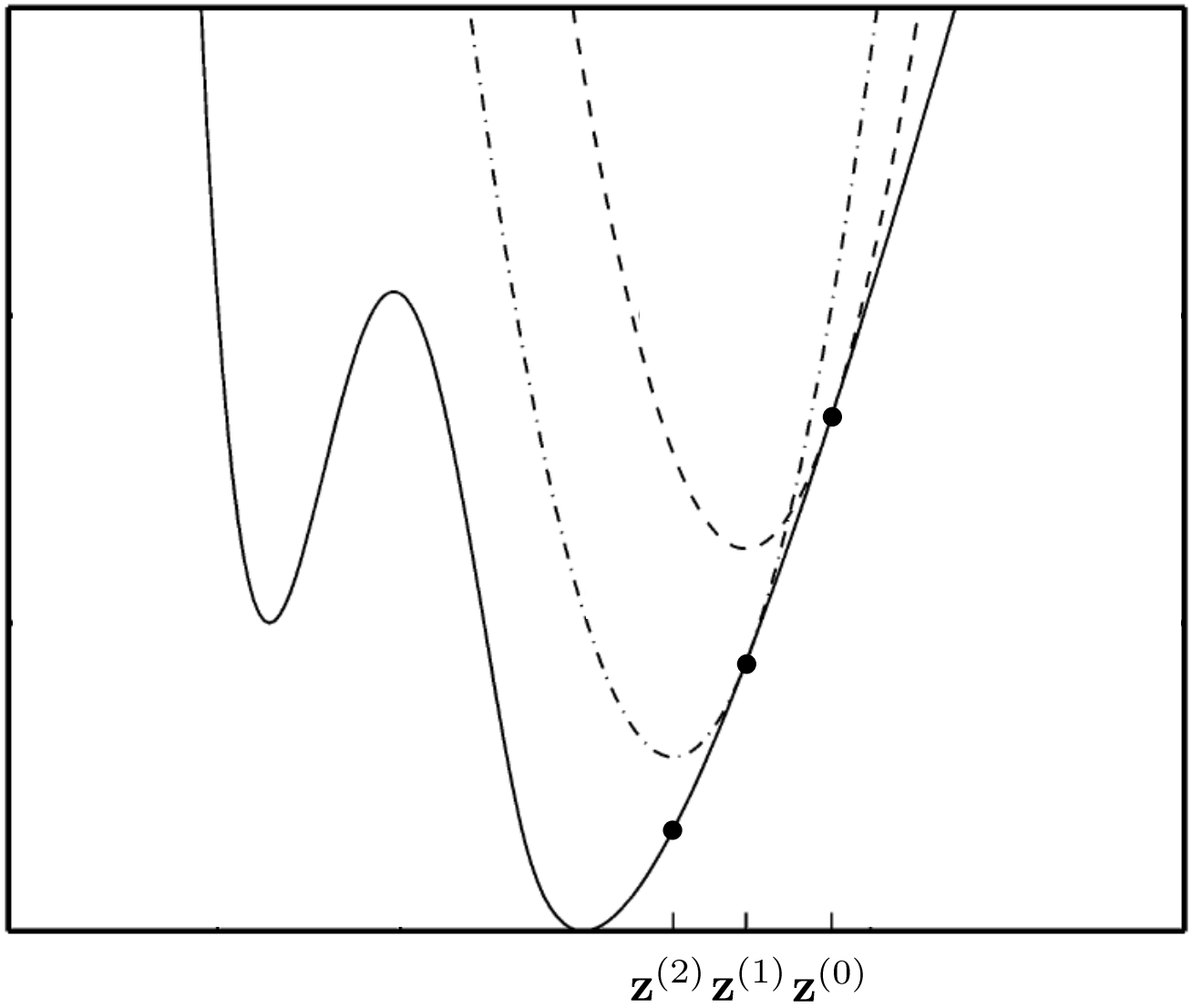
- $\mathbf{z}_1, \ldots, \mathbf{z}_n \in \mathbb{R}^d$
- Sample mean

$$\bar{\mathbf{z}} = \arg\min_{\mathbf{z} \in \mathbb{R}^d} \sum_{i=1}^{n} \|\mathbf{z}_i - \mathbf{z}\|^2$$

- Robust sample mean

$$\bar{\mathbf{z}}_{\text{rob}} = \arg\min_{\mathbf{z} \in \mathbb{R}^d} \sum_{i=1}^{n} \rho\big(\|\mathbf{z}_i - \mathbf{z}\|\big)$$

Majorization / Minimization



Iterative Re-Weighted Least Squares

- Initialize

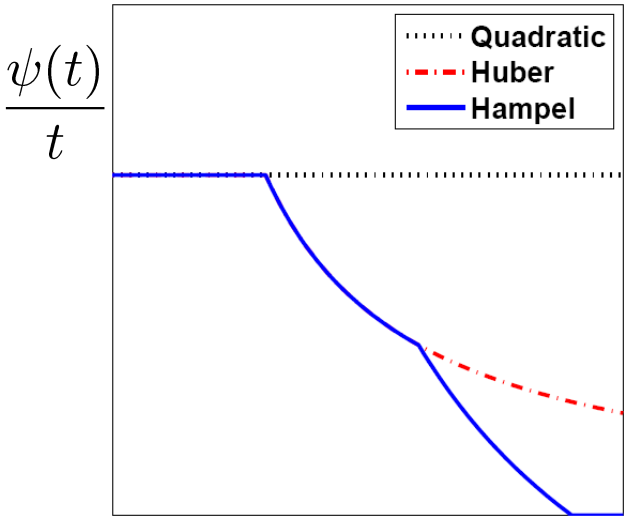
$$w_i^{(0)} = \frac{1}{n}$$

- Iterate

  - $\bar{\mathbf{z}}_{\text{rob}}^{(k)} = \sum_{i=1}^{n} w_i^{(k-1)} \mathbf{z}_i,$
  - $w_i^{(k)} \propto \frac{\psi(\|\mathbf{z}_i - \bar{\mathbf{z}}_{\text{rob}}^{(k)}\|)}{\|\mathbf{z}_i - \bar{\mathbf{z}}_{\text{rob}}^{(k)}\|}$

    s.t. $\sum_{i=1}^{n} w_i^{(k)} = 1$

where $\psi = \rho'$



IRWLS Computation

- $$\|\mathbf{z}_i - \bar{\mathbf{z}}_{\text{rob}}^{(k)}\|^2 = \|\mathbf{z}_i - \textstyle\sum_{j=1}^{n} w_j^{(k-1)} \mathbf{z}_j\|^2$$

- Involves data only through inner products

$$\langle \mathbf{z}_i, \mathbf{z}_j \rangle$$

- Can **kernelize**! Substitute

$$\begin{aligned} \mathbf{z}_i &\rightarrow \Phi_\sigma(\mathbf{X}_i) \\ \langle \mathbf{z}_i, \mathbf{z}_j \rangle &\rightarrow \langle \Phi_\sigma(\mathbf{X}_i), \Phi_\sigma(\mathbf{X}_j) \rangle \\ &\qquad\qquad || \\ &\qquad\quad k_\sigma(\mathbf{X}_i, \mathbf{X}_j) \end{aligned}$$

Kernel IRWLS

- Initialize

$$w_i^{(0)} = \frac{1}{n}$$

- Iterate

  - $$f^{(k)} = \sum_{i=1}^{n} w_i^{(k-1)} \Phi_\sigma(\mathbf{X}_i)$$
    $$= \sum_{i=1}^{n} w_i^{(k-1)} k_\sigma(\cdot, \mathbf{X}_i)$$
  - $$w_i^{(k)} \propto \frac{\psi(\|\Phi_\sigma(\mathbf{X}_i) - f^{(k)}\|)}{\|\Phi_\sigma(\mathbf{X}_i) - f^{(k)}\|}$$
    $$\text{s.t. } \textstyle\sum_{i=1}^{n} w_i^{(k)} = 1$$

Kernel IRWLS Convergence

- **Theorem**: If ρ satisfies certain common assumptions, then $f^{(k)}$ converges to a stationary point.

  If in addition ρ is convex and strictly increasing, then

  $$f^{(k)} \rightarrow \widehat{f}_{RKDE}$$

  in $\mathcal{H}_\sigma$.

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

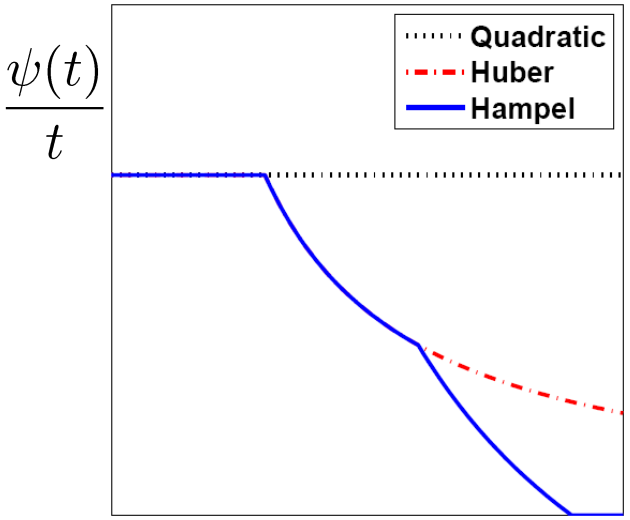
$$\widehat{f}_{RKDE}(\mathbf{x}) = \sum_{i=1}^{n} w_i k_\sigma(\mathbf{x}, \mathbf{X}_i)$$

  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 # 1

- 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

Connection to Data Depth

- Just argued that

$$\|\Phi_\sigma(\mathbf{x}) - \widehat{f}\|^2_{\mathcal{H}_\sigma} = (\sqrt{2\pi}\sigma)^{-d} - 2\widehat{f}(\mathbf{x}) + \|\widehat{f}\|^2_{\mathcal{H}_\sigma}$$

- Therefore

$$\|\Phi_\sigma(\mathbf{x}) - \widehat{f}\|_{\mathcal{H}_\sigma} < a \Longleftrightarrow \widehat{f}(\mathbf{x}) > \lambda$$

where $\lambda = \frac{1}{2}((\sqrt{2\pi}\sigma)^{-d} + \|\widehat{f}\|^2_{\mathcal{H}_\sigma} - a^2)$

- Likelihood depth w.r.t. $\{\mathbf{X}_1, \ldots, \mathbf{X}_n\}$

$$\Updownarrow$$

"centroid" depth w.r.t. $\{\Phi_\sigma(\mathbf{X}_1), \ldots, \Phi_\sigma(\mathbf{X}_n)\}$

Influence Function

- Scalar parameter estimator

$$\widehat{\theta}(F)$$

- e.g., $\theta =$ mean of F

$$\widehat{\theta}(F) = \int x \, dF(x)$$

$$\widehat{\theta}(F_n) = \frac{1}{n} \sum_{i=1}^{n} X_i$$

- Influence function

$$IF(x'; \widehat{\theta}, F) = \lim_{s \to 0} \frac{\widehat{\theta}((1-s)F + s\delta_{x'}) - \widehat{\theta}(F)}{s}$$

Influence Function

- Density estimator

$$\widehat{f}(\mathbf{x}; F)$$

- e.g., KDE

$$\widehat{f}_{KDE}(\mathbf{x}; F) = \int k_\sigma(\mathbf{x}, \mathbf{y})\, dF(\mathbf{y})$$

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

- Influence function

$$IF(\mathbf{x}, \mathbf{x}'; \widehat{f}, F) = \lim_{s \to 0} \frac{\widehat{f}(\mathbf{x}; (1-s)F + s\delta_{\mathbf{x}'}) - \widehat{f}(\mathbf{x}; F)}{s}$$

Influence Function

- KDE

$$IF(\mathbf{x}, \mathbf{x}'; \widehat{f}_{KDE}, F_n) = -\frac{1}{n}\sum_{i=1}^{n} k_\sigma(\mathbf{x}, \mathbf{X}_i) + k_\sigma(\mathbf{x}, \mathbf{x}')$$

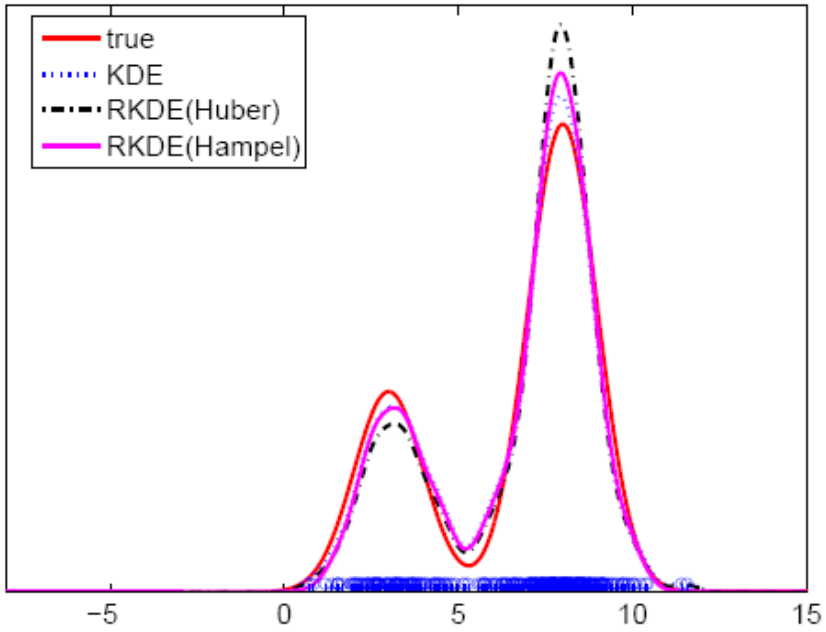
- RKDE: **Theorem**:

$$IF(\mathbf{x}, \mathbf{x}'; \widehat{f}_{RKDE}, F_n) = \sum_{i=1}^{n} \alpha_i k_\sigma(\mathbf{x}, \mathbf{X}_i) + \alpha' k_\sigma(\mathbf{x}, \mathbf{x}')$$

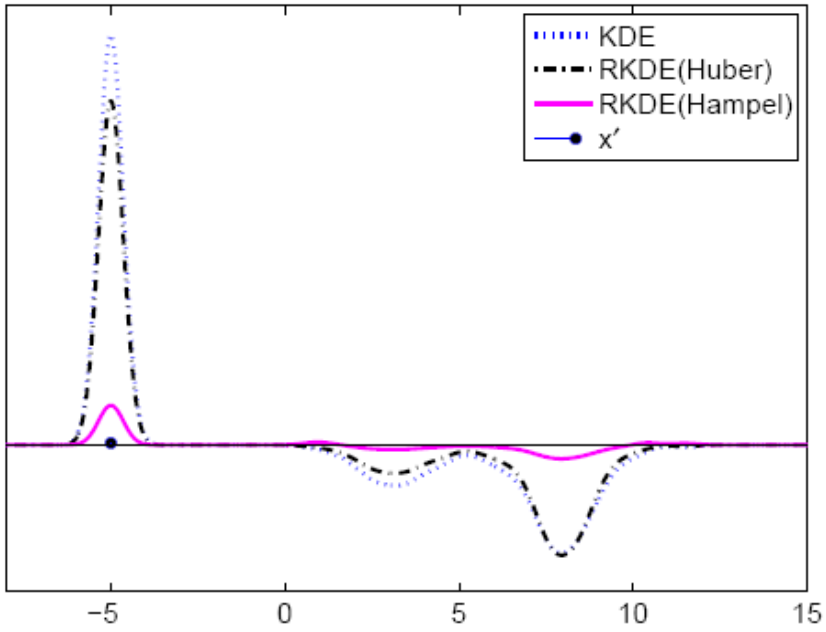
where $(\alpha_1, \ldots, \alpha_n, \alpha')$ is the solution of a system of linear equations.

Example

$\widehat{f}(\mathbf{x}; F_n)$

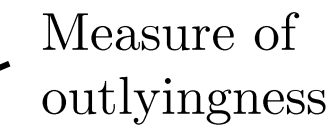


$IF(\mathbf{x}, \mathbf{x}'; \widehat{f}, F_n)$



Robustness Interpretation # 2

- Exact formula



$$\alpha' = \frac{\frac{\psi(\|\Phi_\sigma(\mathbf{x}') - \widehat{f}_{RKDE}\|)}{\|\Phi_\sigma(\mathbf{x}') - \widehat{f}_{RKDE}\|}}{\frac{1}{n}\sum_{i=1}^{n} \frac{\psi(\|\Phi_\sigma(\mathbf{X}_i) - \widehat{f}_{RKDE}\|)}{\|\Phi_\sigma(\mathbf{X}_i) - \widehat{f}_{RKDE}\|}}$$

- KDE

$$\alpha' = 1$$

- RKDE

$$\alpha' < 1$$

if $\mathbf{x}'$ is an outlier

Asymptotics: σ fixed

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

$$f_\sigma = \arg\min_{g \in \mathcal{H}_\sigma} \underbrace{\int \rho\big(\|\Phi_\sigma(\mathbf{x}) - g\|_{\mathcal{H}_\sigma}\big)\, f(x)\, dx}_{J(g)}$$

Theorem: If f has compact support, then

$$J(f_\sigma^n) \xrightarrow{i.p.} J(f_\sigma) \quad \text{as } n \to \infty$$

Asymptotics: σ fixed

- **Corollary**: If f has compact support, and ρ is convex and strictly increasing, then

$$\|f_\sigma^n - f_\sigma\|_{\mathcal{H}_\sigma} \xrightarrow{i.p.} 0 \qquad \text{as } n \to \infty$$

- **Corollary**: Under the same assumptions

$$\begin{aligned}\|f_\sigma^n - f_\sigma\|_\infty &= \sup_{\mathbf{x}} |\langle \Phi_\sigma(\mathbf{x}), f_\sigma^n - f_\sigma \rangle_{\mathcal{H}_\sigma}| \\ &\leq \underbrace{\|\Phi_\sigma(\mathbf{x})\|_{\mathcal{H}_\sigma}}_{\| \\ (\sqrt{2\pi}\sigma)^{-d/2}} \|f_\sigma^n - f_\sigma\|_{\mathcal{H}_\sigma} \xrightarrow{i.p.} 0 \qquad \text{as } n \to \infty\end{aligned}$$

Robustness Interpretation # 3

- Representer theorem (infinite sample):

$$f_\sigma = k_\sigma * p_\sigma$$

where p_σ is a density, $\text{supp}(p_\sigma) \subseteq \text{supp}(f)$

- Quadratic loss (KDE)

$$p_\sigma = f$$

- Robust loss (RKDE)

Tails are down-weighted

$$p_\sigma(\mathbf{x}) = \frac{w_\sigma(\mathbf{x}) f(\mathbf{x})}{\int w_\sigma(\mathbf{y}) f(\mathbf{y}) dy}$$

where

$$w_\sigma(\mathbf{x}) = \frac{\psi(\|\Phi_\sigma(\mathbf{x}) - f_\sigma\|_{\mathcal{H}_\sigma})}{\|\Phi_\sigma(\mathbf{x}) - f_\sigma\|_{\mathcal{H}_\sigma}}$$

Asymptotics: $\sigma \to 0$

- Does

$$\lim_{\sigma \to 0} f_\sigma$$

  exist?

- If $f_\sigma \to f_0$, and $\sigma = \sigma_n \to 0$, under what conditions does

$$f^n_{\sigma_n} \to f_0 \ ?$$

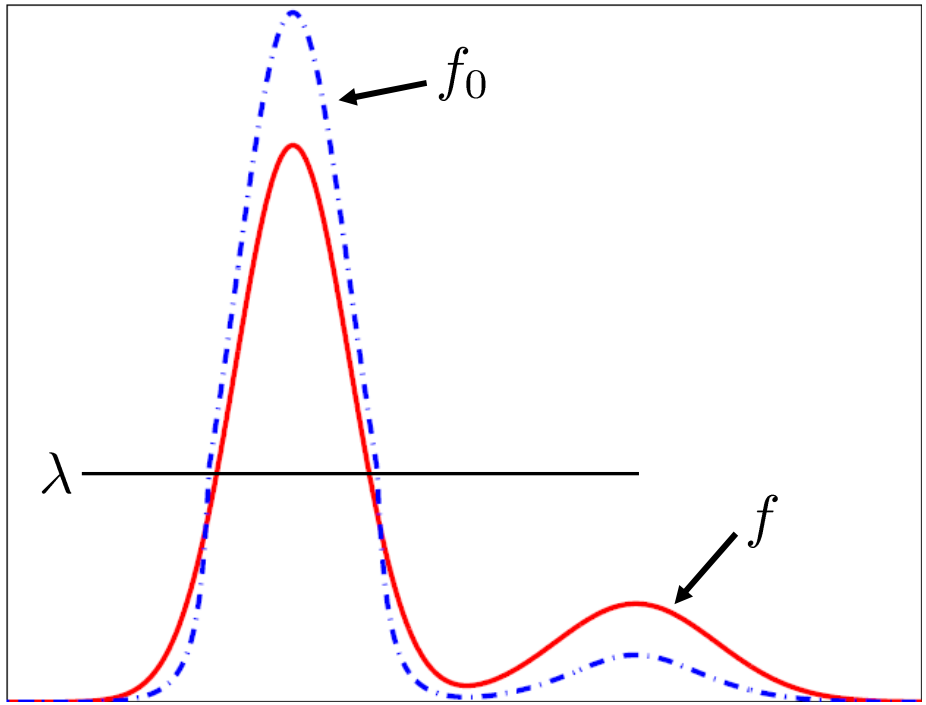
- How does f_0 depend on ρ?

Hampel Loss

- **Conjecture:**

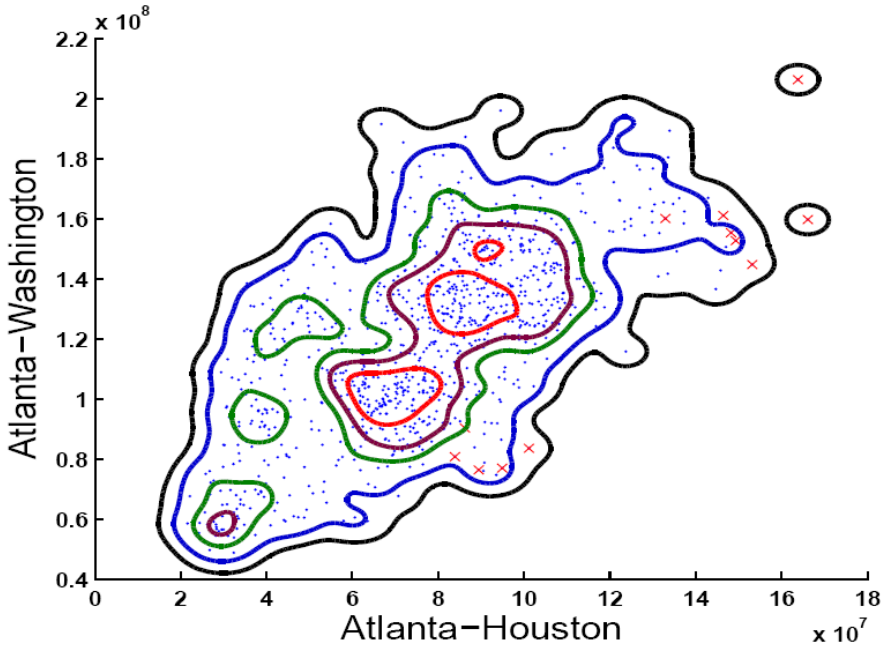
$$f_0(\mathbf{x}) = \begin{cases} af(\mathbf{x}), & f(\mathbf{x}) \geq \lambda \\ \frac{f(\mathbf{x})}{b - cf(\mathbf{x})}, & f(\mathbf{x}) < \lambda \end{cases}$$

for some $a > 1,\ b, c, \lambda > 0$

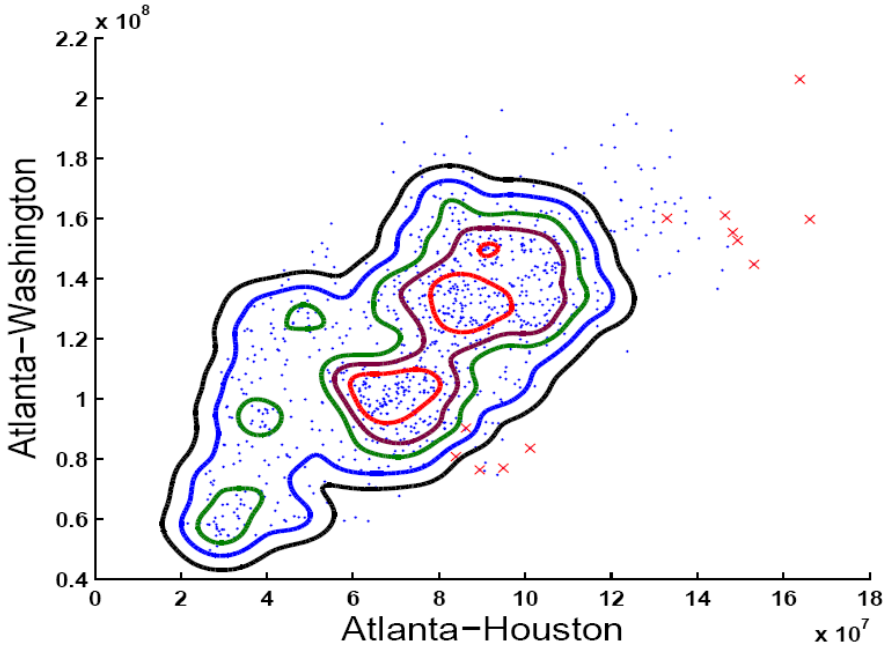


Example: Hampel Loss

KDE



RKDE



Anomaly Detection

- Training data

$$\mathbf{X}_1, \ldots, \mathbf{X}_n \sim f(\mathbf{x}) = (1-\epsilon) f_{\text{norm}}(\mathbf{x}) + \epsilon f_{\text{anom}}(\mathbf{x})$$

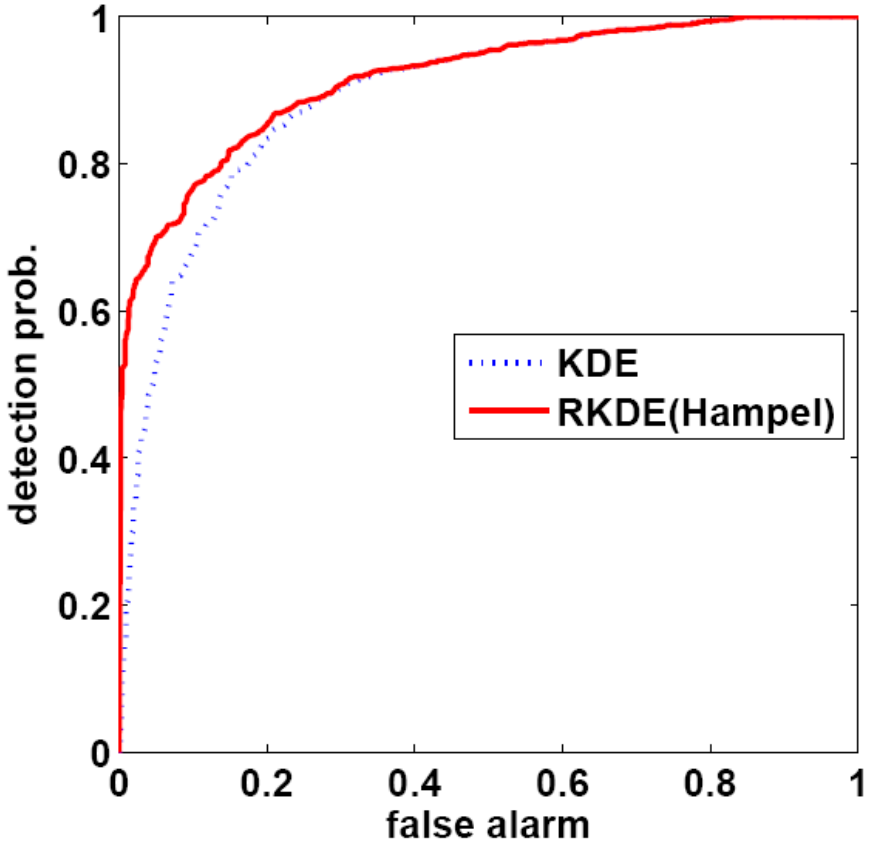
- Task: Classify normal examples versus anomalies
- Anomaly detector:

$$\{\mathbf{x} : \widehat{f}(\mathbf{x}) > \lambda\}$$

  for all $\lambda \geq 0$

- Performance:
  Area under ROC (AUC)
- Testing data

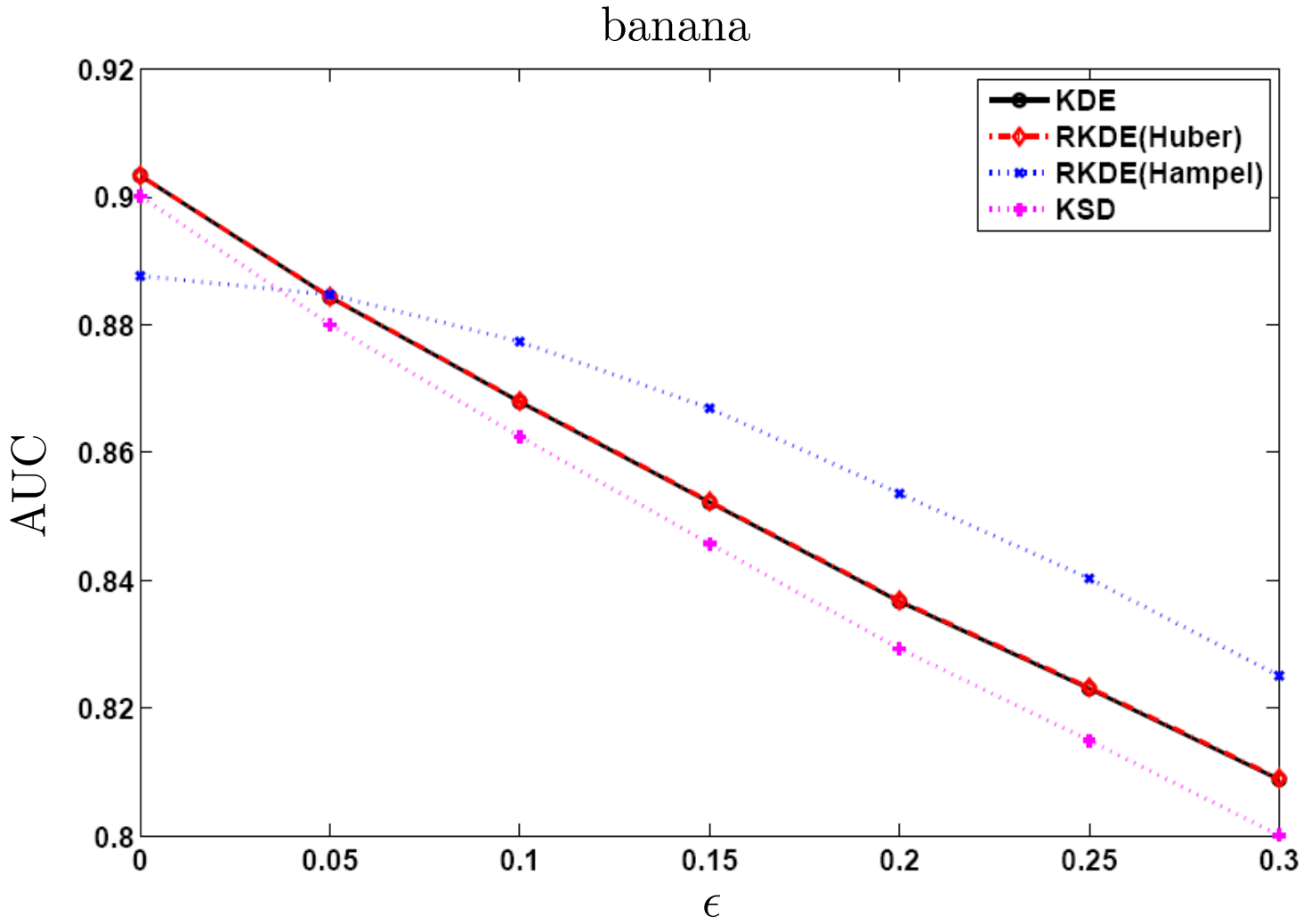
$$\mathbf{X}'_1, \ldots, \mathbf{X}'_{n'} \sim f_{\text{norm}}(\mathbf{x})$$
$$\mathbf{X}''_1, \ldots, \mathbf{X}''_{n''} \sim f_{\text{anom}}(\mathbf{x})$$



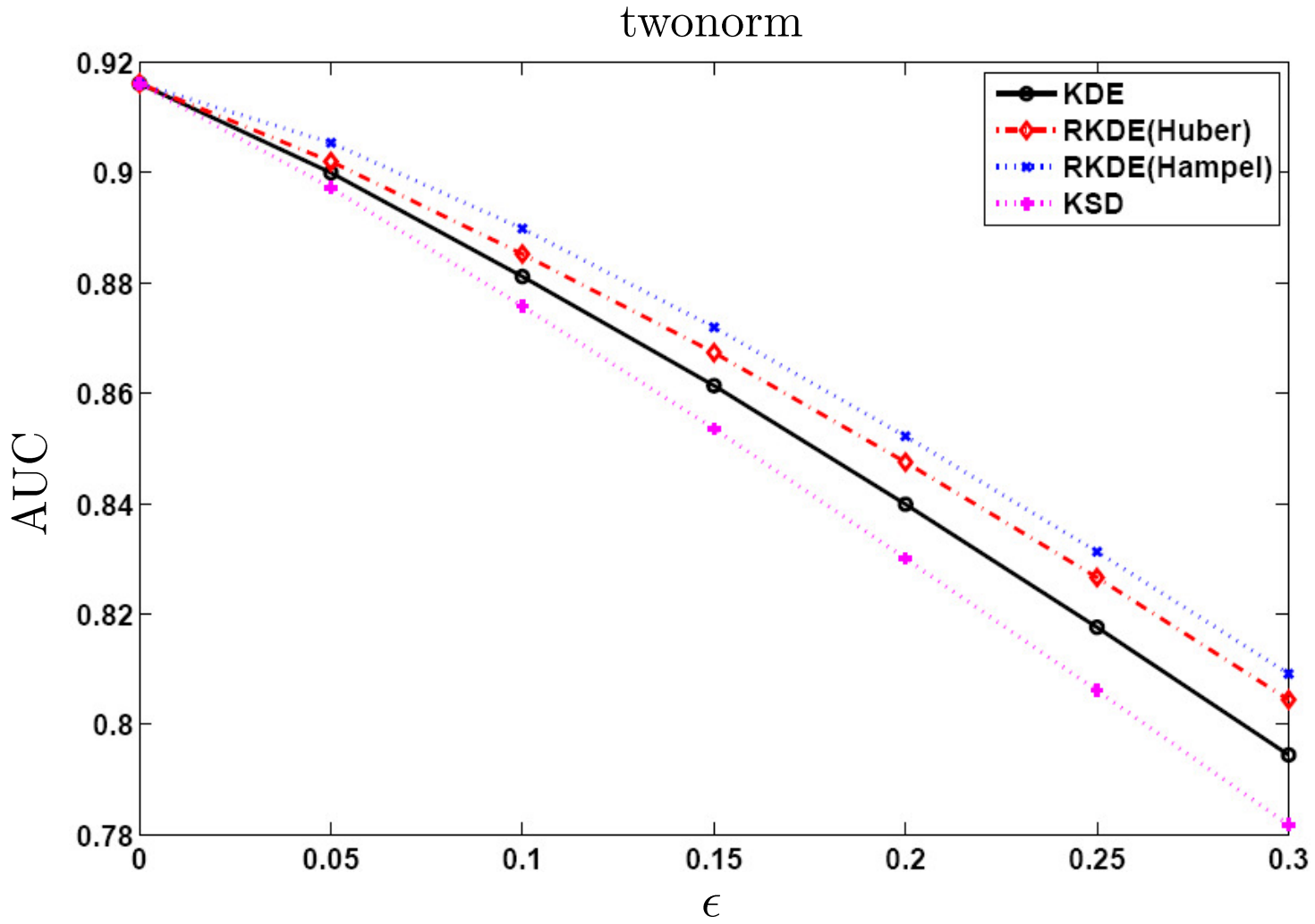
Anomaly Detection

- 15 benchmark datasets for binary classification
- $\epsilon = 0, 0.05, 0.1, 0.15, 0.2, 0.25, 0.3$
- Algorithms compared
  - KDE
  - RKDE (Huber)
  - RKDE (Hampel)
  - Kernelized spatial depth (Chen et al., 2009)

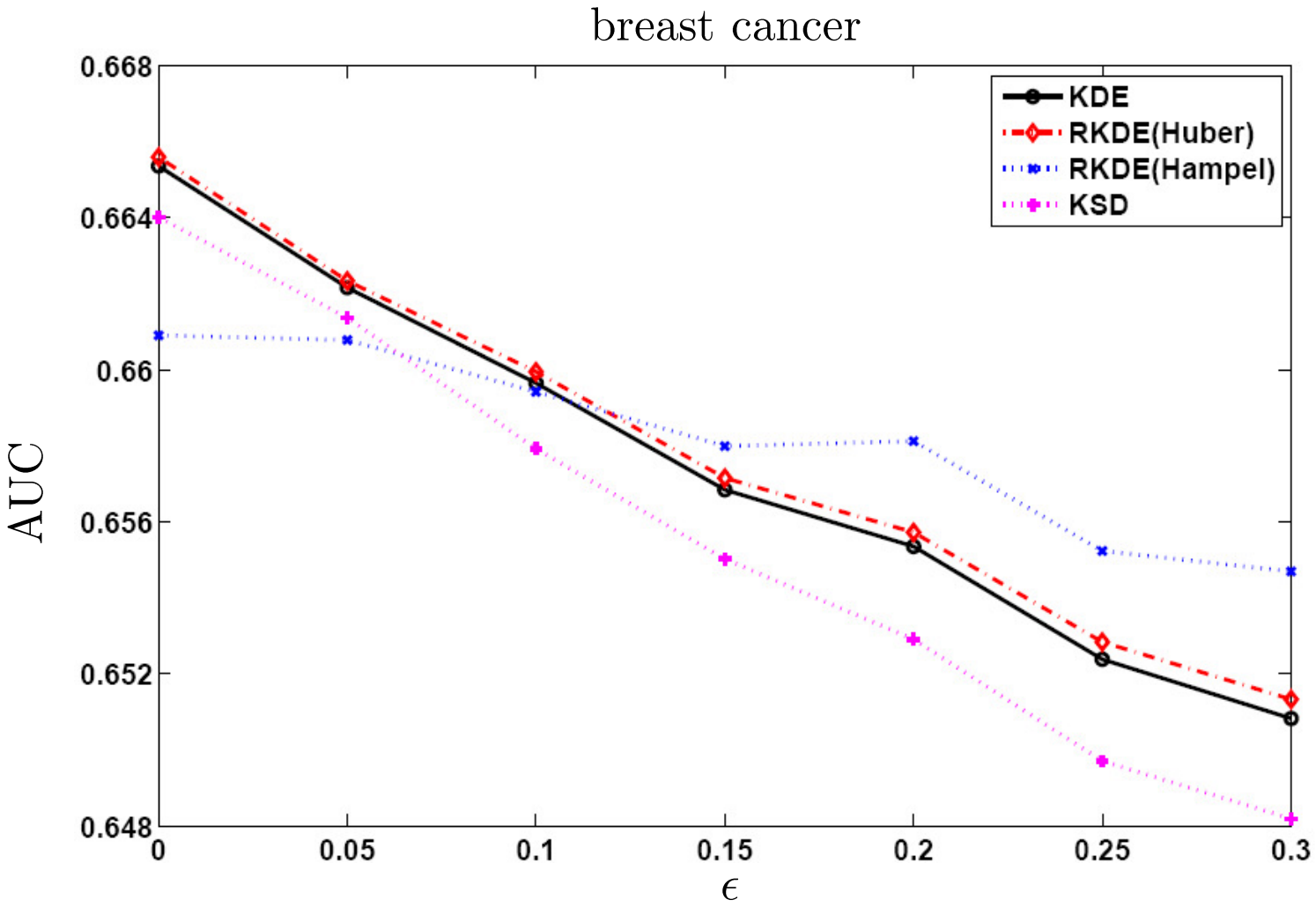
Anomaly Detection: AUC versus ε



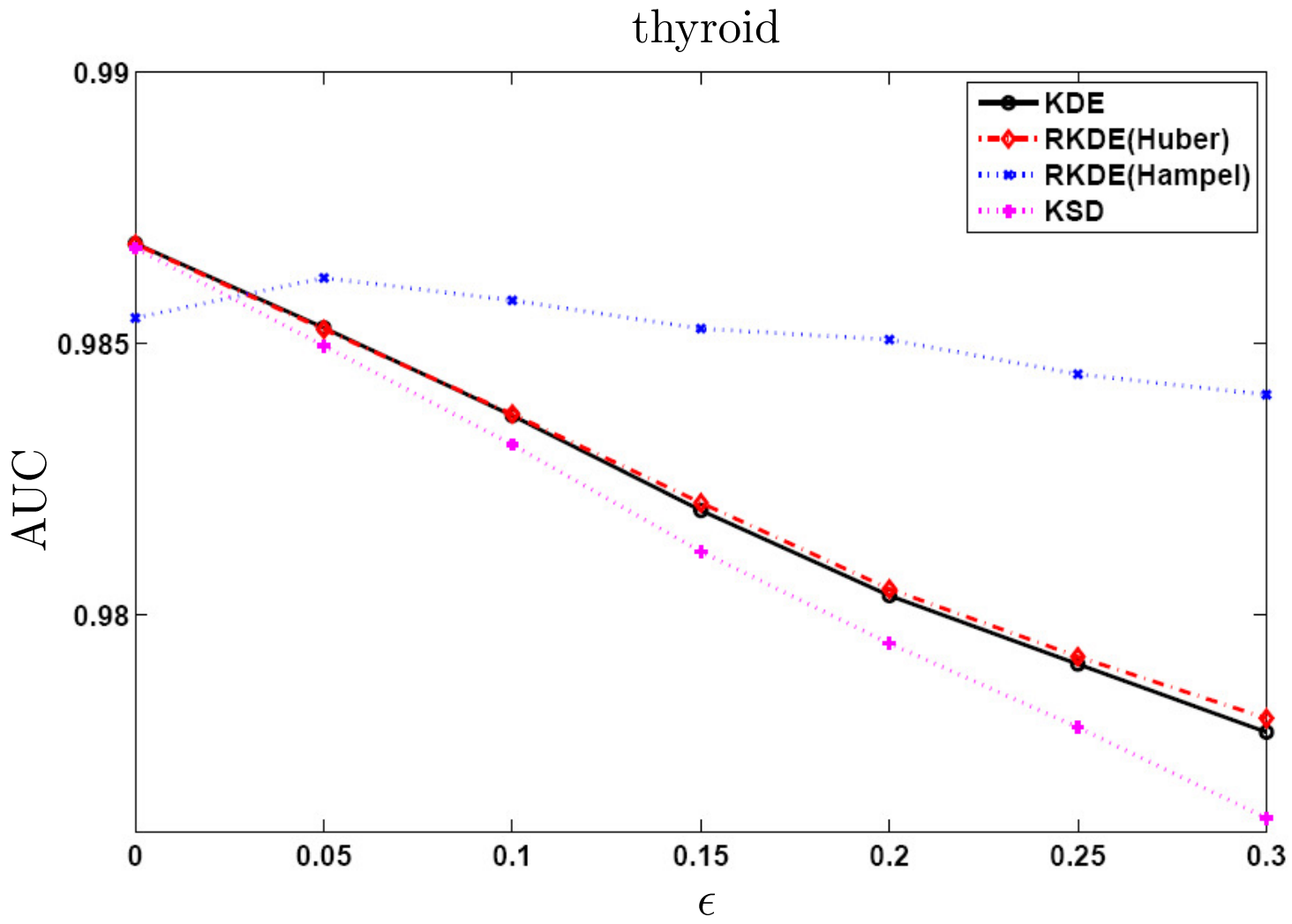
Anomaly Detection: AUC versus ε



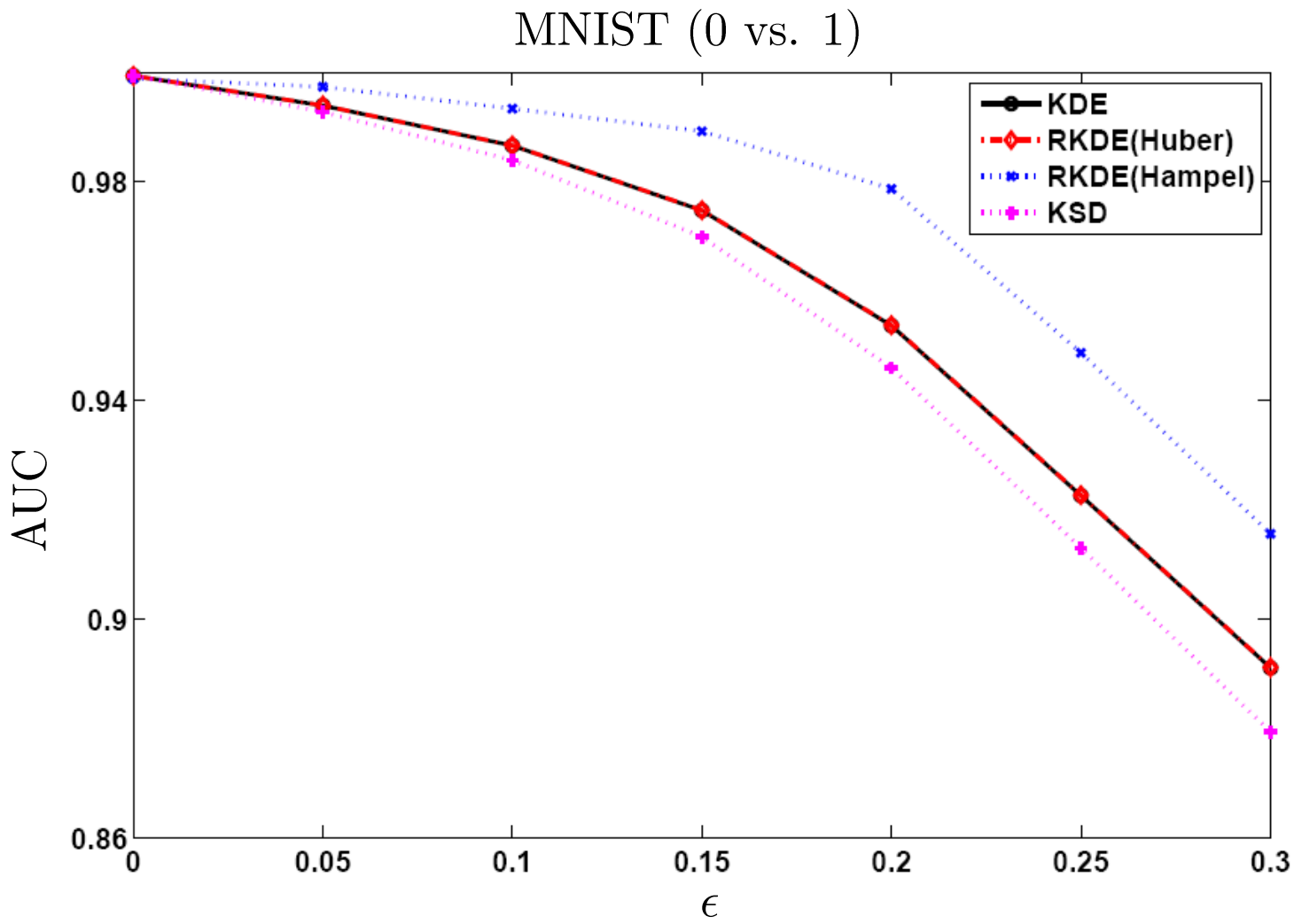
Anomaly Detection: AUC versus ε



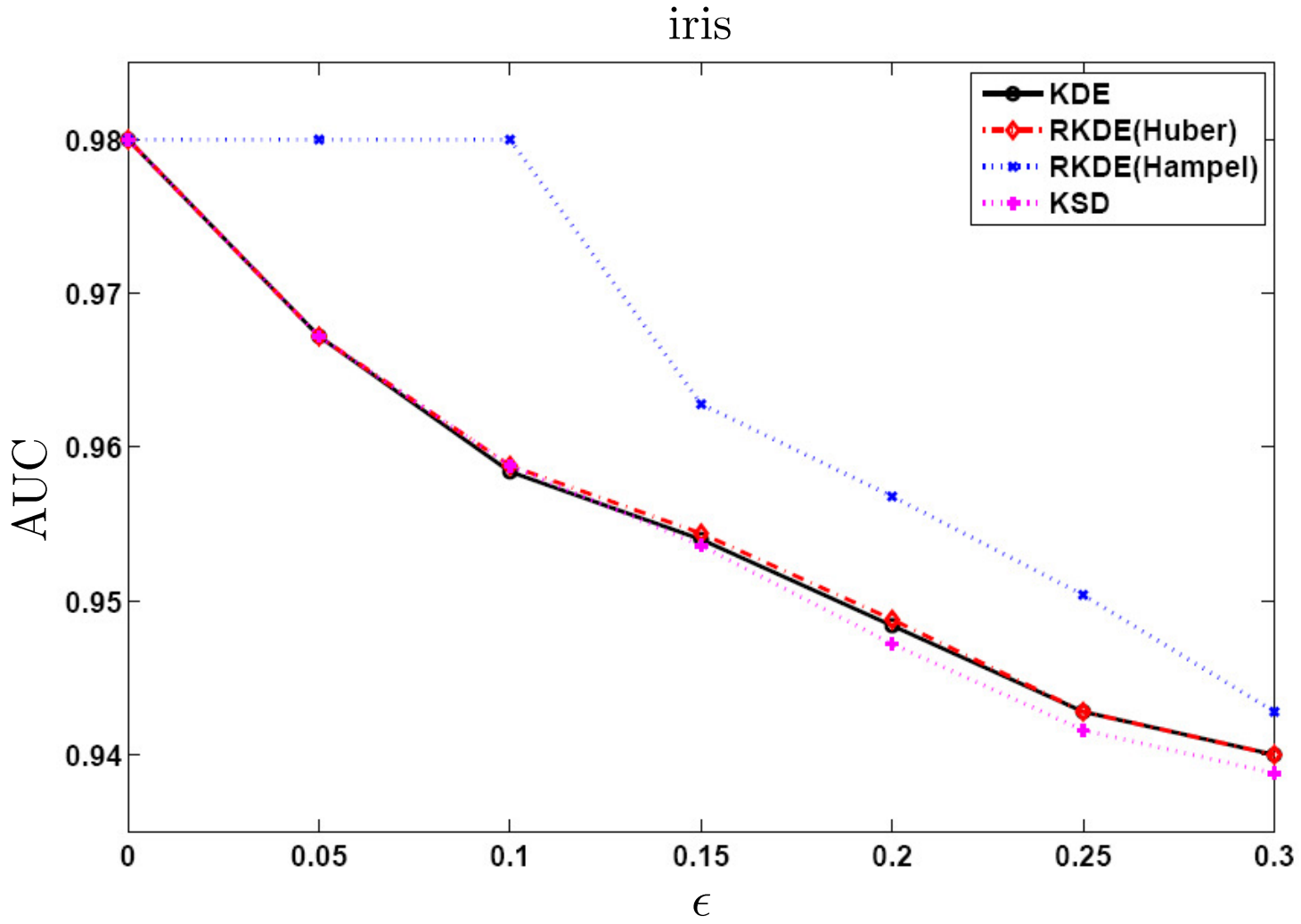
Anomaly Detection: AUC versus ε



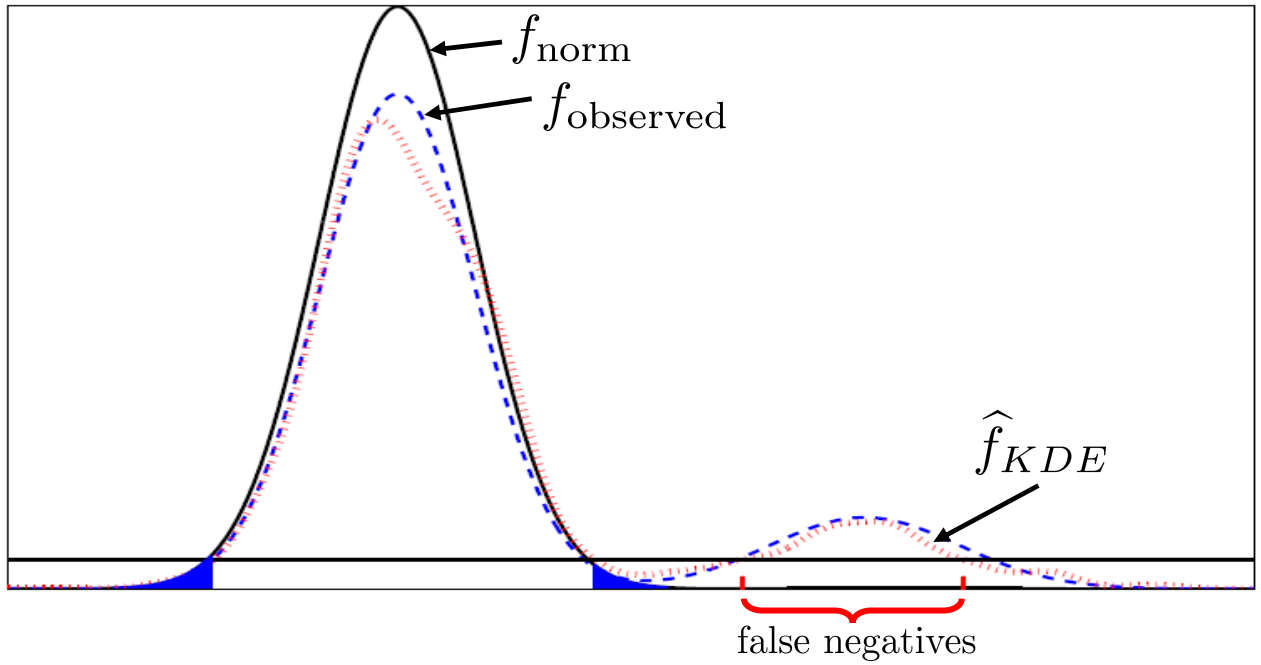
Anomaly Detection: AUC versus ε



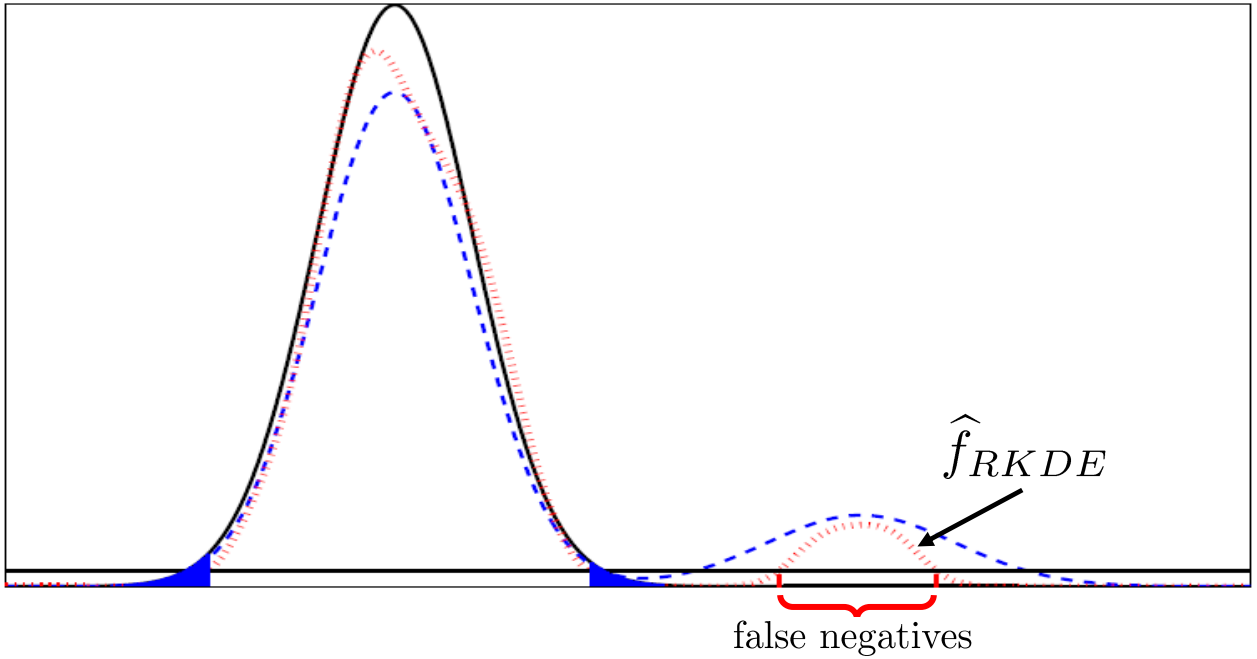
Anomaly Detection: AUC versus ε



Explanation



levels chosen to ensure same false alarm rate (shaded area)



Robustness Interpretation # 4:

- Define

  $\alpha(\widehat{f}) =$ smallest α such that some level set of $\widehat{f}$
  is a consistent estimate of the
  $(1-\alpha)$-**generalized quantile set** of f_{norm}

- **Conjecture**: If
  - ϵ is sufficiently small
  - f_{anom} and f_{norm} are sufficiently distinct

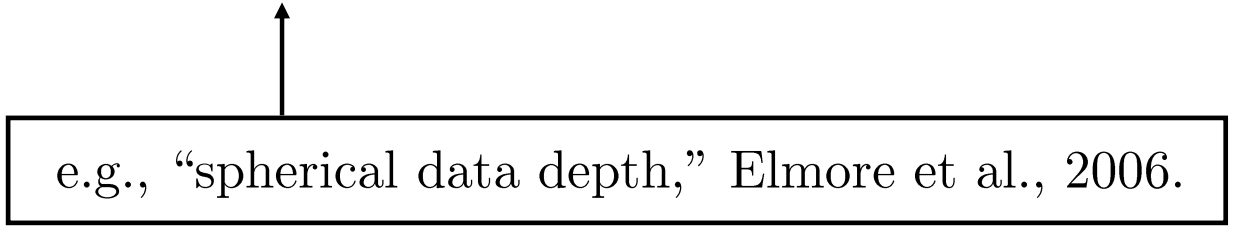
  then
  $$\alpha(\widehat{f}_{RKDE}) < \alpha(\widehat{f}_{KDE})$$
  for the RKDE based on Hampel's loss

Conclusions

- Robust nonparametric density estimation is possible, if
  - ϵ is small
  - f_{anom} is outlying with respect to f_{norm}
- Easy implementation:
  - Training: Kernel IRWLS: $O(n^2)$ steps per iteration
  - Testing: $O(n)$ steps
- Interesting theory and works well in practice
- Data depth can be simple if it is kernelized
- Many open questions

Extensions / Open Questions

- Other kernels?
  - for anomaly detection, kernel need not integrate to 1
- Consistency, rates, limiting distribution?
- Re-analysis of KDE?
- Breakdown point?
- Other kernelizable depth functions / multivariate medians?



Acknowledgments

- Joint work with JooSeuk Kim, Ph.D. candidate at the University of Michigan
- Supported by NSF Awards 0830490 and 0953135.

Parameter Tuning

- **Kernel bandwidth**: σ = median distance to nearest neighbor (among training data)
- **Hampel loss parameters**:
  - Train RKDE with $\rho(t) = t$ to get $\widehat{f}_1$
  - Set $d_i = \|\Phi_\sigma(\mathbf{X}_i) - \widehat{f}_1\|$
  - Define a, b, and c to be the 50th, 95th, and 100th percentiles of $\{d_i\}$

Proof Idea

- Representer theorem (infinite sample):

$$f_\sigma = k_\sigma * p_\sigma$$

where p_σ is a density, $\mathrm{supp}(p_\sigma) \subseteq \mathrm{supp}(f)$

- Define

$$D_\sigma = \left\{ k_\sigma * p \;\middle|\; \begin{array}{l} p \text{ is a density with} \\ \mathrm{supp}(p) \subseteq \mathrm{supp}(f) \end{array} \right\}$$

- Uniform error bound

$$P\left\{ \sup_{g \in D_\sigma} \left| J_n(g) - J(g) \right| > \epsilon \right\} \leq \underbrace{N(D_\sigma, \|\cdot\|_{\mathcal{H}_\sigma}, \epsilon)}_{\text{covering number}} \exp\left\{ \frac{-cn\epsilon^2}{\rho(\sigma^{-d/2})^2} \right\}$$

Anomaly Detection: Average Ranks

	epsilon						
	0	0.05	0.1	0.15	0.2	0.25	0.3
KDE	2.43	2.63	2.67	2.73	2.80	2.73	2.70
RKDE_Huber	2.10	2.33	2.13	2.20	2.13	2.33	2.23
RKDE_Hampel	3.10	2.13	1.87	1.67	1.67	1.47	1.60
KSD	2.37	2.90	3.33	3.40	3.40	3.47	3.47

CD at 0.05	1.21
CD at 0.10	1.08

Spatial Depth

- Spatial depth:

$$\text{outlyingness}(\mathbf{x}) \propto \underbrace{\left\| \sum_{i=1}^{n} \frac{\psi(\|\mathbf{x}_i - \mathbf{x}\|)}{\|\mathbf{x}_i - \mathbf{x}\|} (\mathbf{x}_i - \mathbf{x}) \right\|}_{= 0 \text{ at multivariate median}}$$

  with $\psi \equiv 1$

- If ψ is decreasing (ρ nonconvex), then multiple stationary points exist (besides multivariate median), and "outlyingness" interpretation may be lost