Robust Artificial Intelligence

Reading Seminar; Tsinghua University

Thomas G. Dietterich, Oregon State University

tgd@cs.orst.edu

The Class So Far

- Lecture 1: Calibrated Probabilities (Closed World)
- Lecture 2: Thresholding Confidence Indicators (Closed World)
- Lecture 3: Open Category Detection
- Lecture 4: Anomaly Detection

Lecture 4: Anomaly Detection

- Definition: An “anomaly” is a data point generated by a process that is different than the process generating the “nominal” data
- Given:
  - Training data: $\{x_1, x_2, \dots, x_N\}$
  - Case 1: All data come from D_0 the “nominal” distribution
  - Case 2: The data come from a mixture of D_0 and D_a the “anomaly” distribution
  - Test data: $\{x_{N+1}, \dots, x_{N+M}\}$ from a mixture of D_0 and D_a
- Find:
  - The data points in the test data that belong to D_a
- Note: D_a need not be a stationary distribution, but we general assume that D_0 is stationary.

Papers for Today

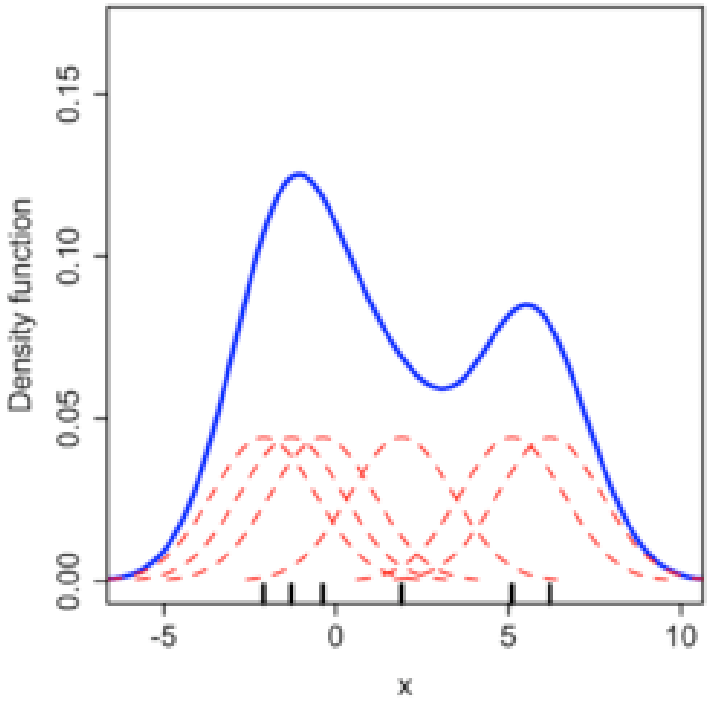
- Liu, F. T., Ting, K. M., & Zhou, Z.-H. (2012). Isolation-Based Anomaly Detection. ACM Transactions on Knowledge Discovery from Data, 6(1), 1–39. http://doi.org/10.1145/2133360.2133363
- Emmott, A., Das, S., Dietterich, T., Fern, A., & Wong, W.-K. (2015). Systematic construction of anomaly detection benchmarks from real data. https://arxiv.org/abs/1503.01158
- Siddiqui, A., Fern, A., Dietterich, T. G., & Das, S. (2016). Finite Sample Complexity of Rare Pattern Anomaly Detection. In *Proceedings of UAI-2016* (p. 10). http://auai.org/uai2016/proceedings/papers/226.pdf

Algorithms

- Density-Based Approaches
  - RKDE: Robust Kernel Density Estimation (Kim & Scott, 2008)
  - EGMM: Ensemble Gaussian Mixture Model (our group)
- Quantile-Based Methods
  - OCSVM: One-class SVM (Schoelkopf, et al., 1999)
  - SVDD: Support Vector Data Description (Tax & Duin, 2004)
- Neighbor-Based Methods
  - LOF: Local Outlier Factor (Breunig, et al., 2000)
  - ABOD: kNN Angle-Based Outlier Detector (Kriegel, et al., 2008)
- Projection-Based Methods
  - IFOR: Isolation Forest (Liu, et al., 2008)
  - LODA: Lightweight Online Detector of Anomalies (Pevny, 2016)

Robust Kernel Density Estimation

- Kernel Density Estimation
  - Let $k_\sigma(x, x')$ be a positive semi-definite kernel such as the Gaussian kernel or the Student-t-kernel
  - $\hat{p}(x) = \frac{1}{N}\sum_{i=1}^{N} k_\sigma(x, x_i)$
- Let $\Phi(x)$ be the feature function corresponding to k_σ
  - $k_\sigma(x, x') = \langle \Phi(x), \Phi(x') \rangle$
- Then the KDE is the solution to a least squares problem in Hilbert space:
  - $\hat{p} = \min_{g \in \mathcal{H}} \sum_{i=1}^{N} \|\Phi(x_i) - g(x_i)\|_{\mathcal{H}}^2$
- We can make this more robust by replacing the square loss with a robust loss

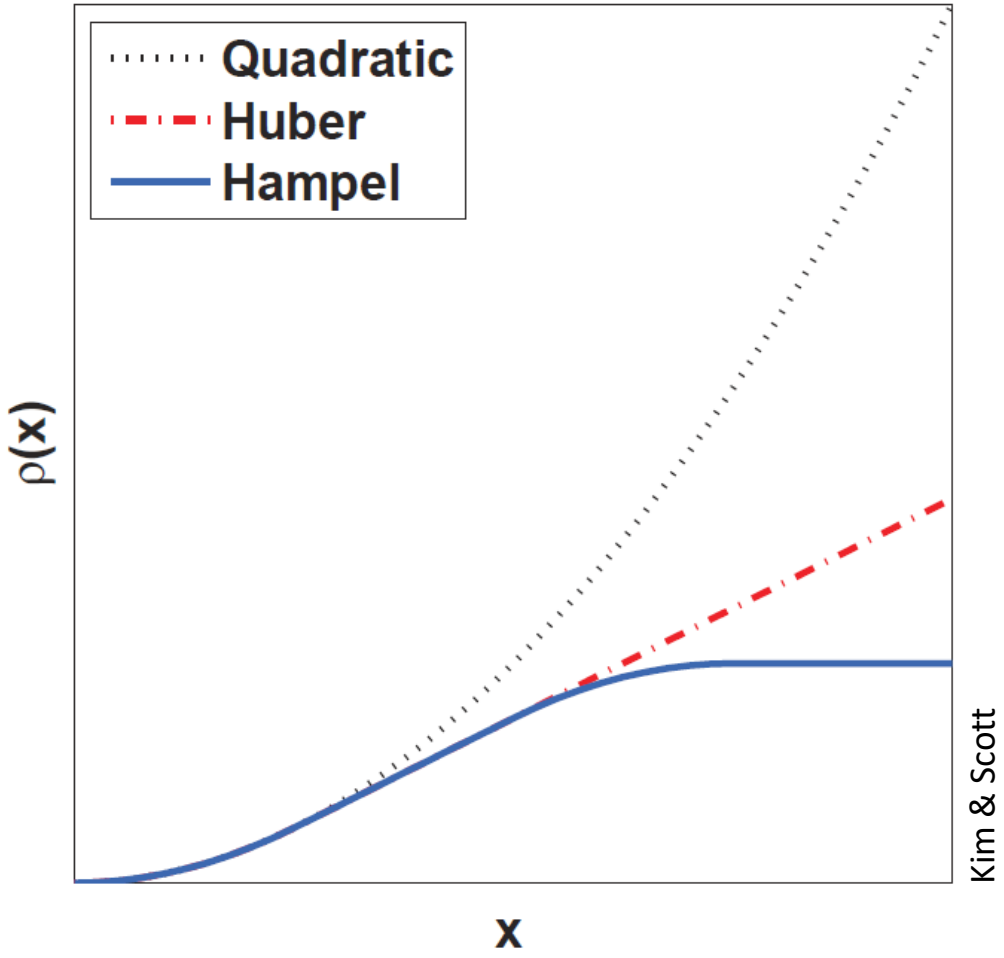


Wikipedia

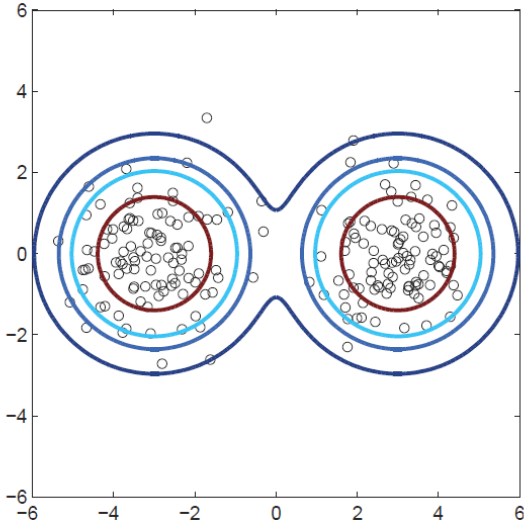
Robust Loss Functions

$$\hat{p} = \underset{g \in \mathcal{H}}{\operatorname{argmin}} \sum_{i=1}^{N} \rho(\|\Phi(x_i) - g(x_i)\|_{\mathcal{H}})$$

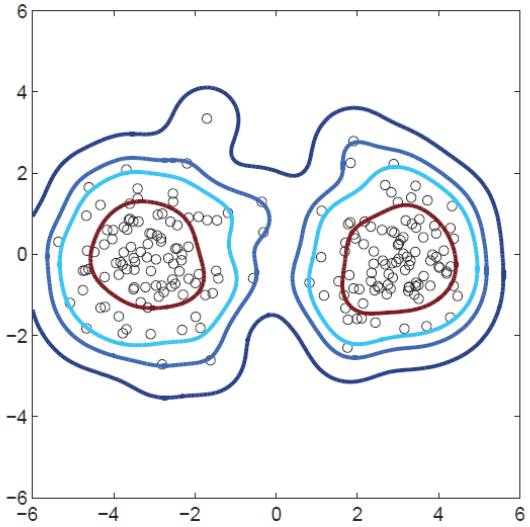
This can be solved by Iteratively Reweighted Least Squares



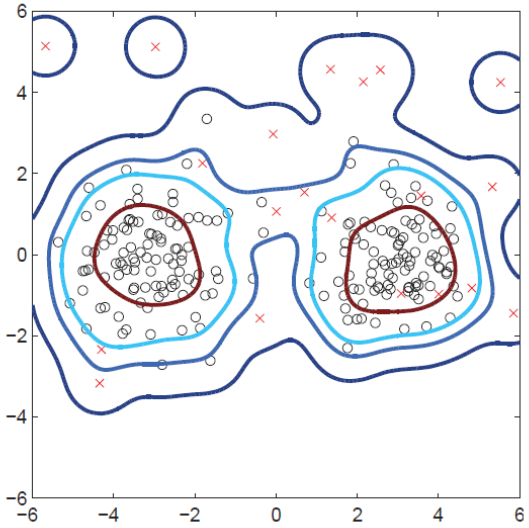
Example:
Mixture of 2 Gaussians



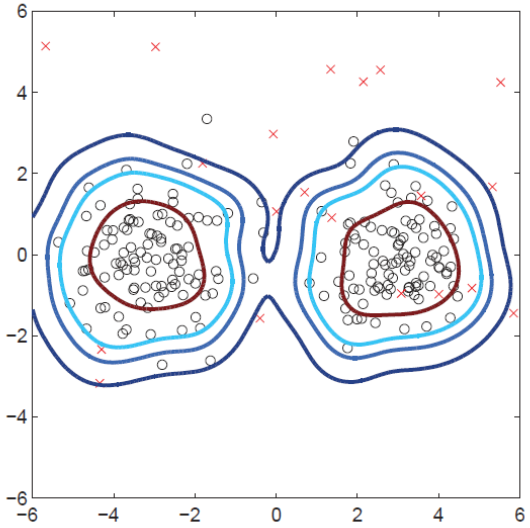
(a) True density



(b) KDE without outliers



(c) KDE with outliers

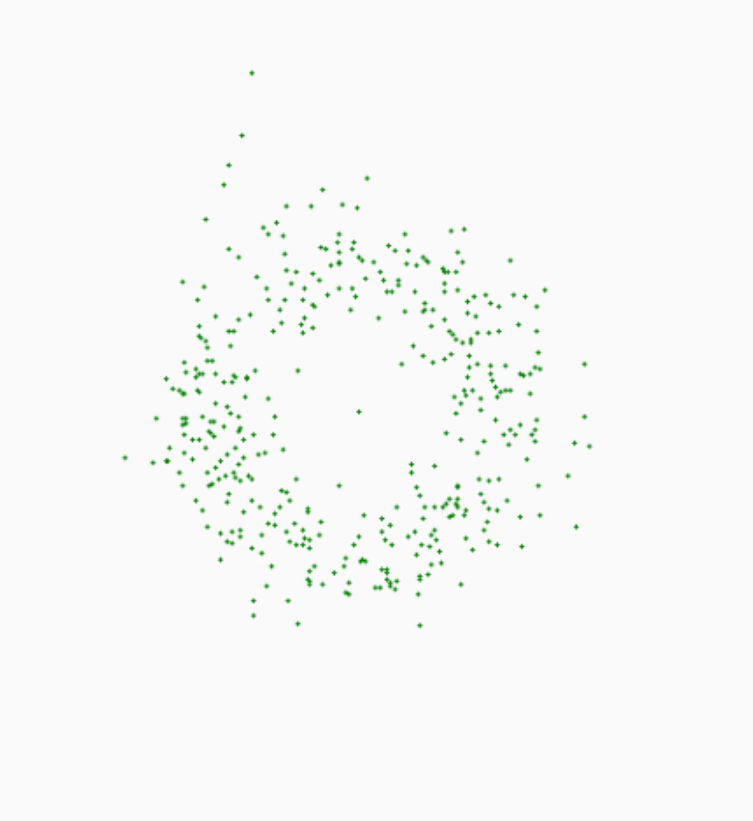


(d) RKDE with outliers

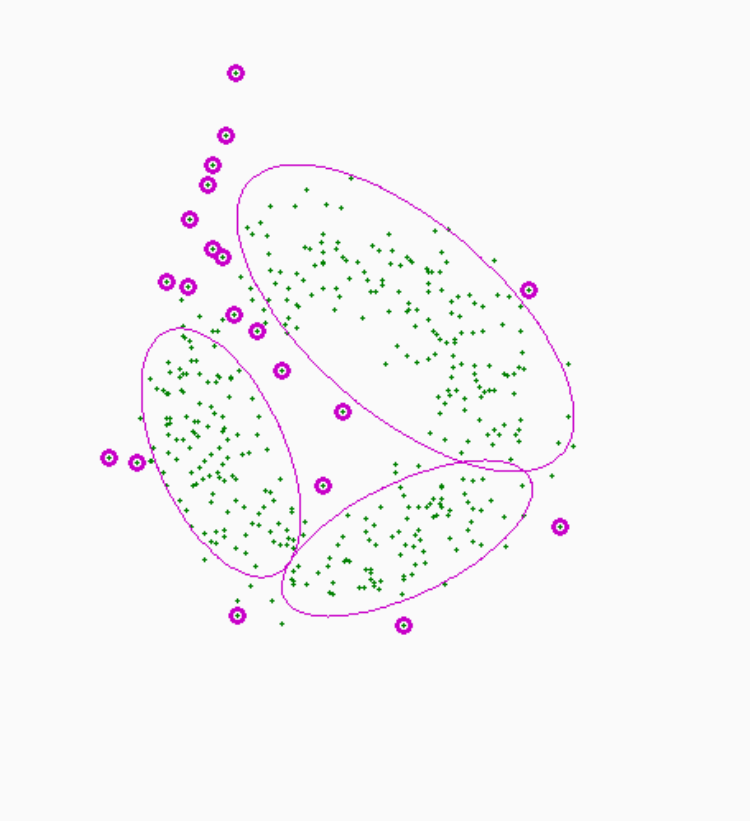
Kim & Scott

Ensemble of Gaussian Mixture Models

- $P(x) = \sum_{k=1}^{K} p_k \cdot \text{Normal}(x|\mu_k, \Sigma_k)$

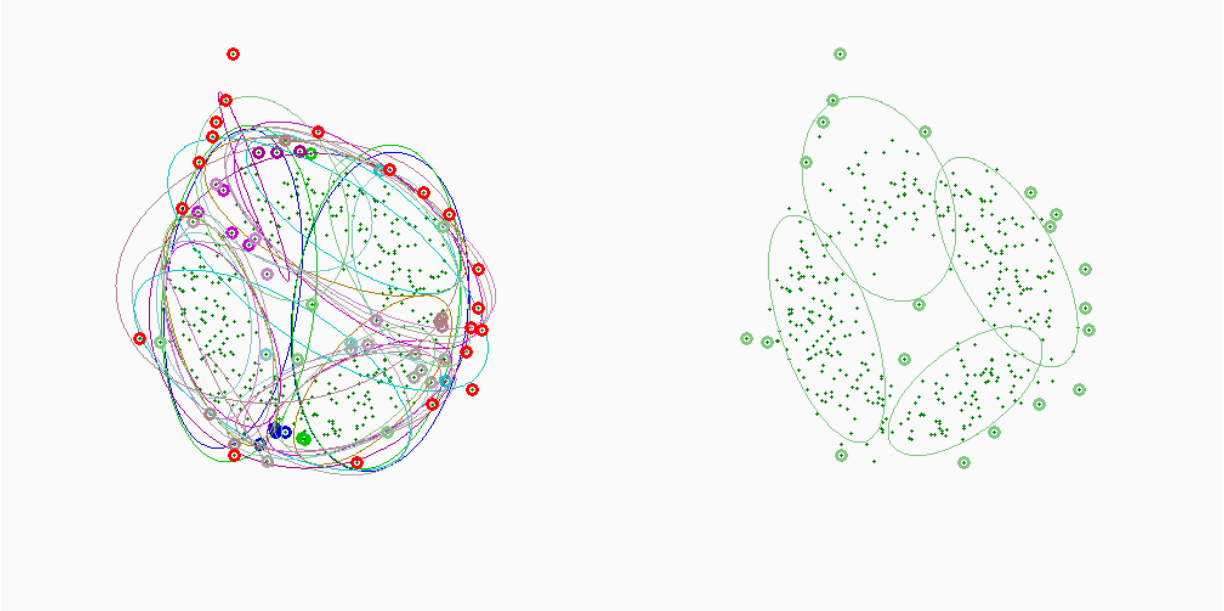


K=3



Ensemble of GMMs

- Train M independent Gaussian Mixture Models
- Train model $m = 1, \ldots, M$ on a bootstrap replicate of the data
- Vary the number of clusters K
- Delete any model with log likelihood < 70% of best model
- Compute average surprise: $-\frac{1}{M}\sum_m \log P_m(x_i)$



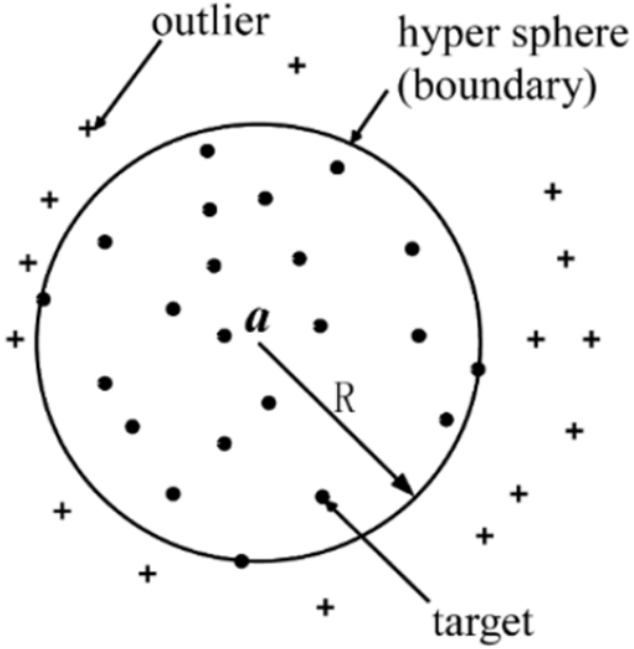
One-Class Support Vector Machine

(Schoelkopf, Williamson, Smola, Shawe-Taylor, Platt, NIPS 2000)

- Given a kernel $k(x, x')$, map the data into the feature space $\Phi(x)$ and find a hyperplane that is as far from the origin as possible and separates $1 - \nu$ of the data points from the origin
- Solution to the following
  - $\min_{w,\xi,\rho} \frac{1}{2}\|w\|^2 + \frac{1}{\nu N}\sum_{i=1}^{N}\xi_i - \rho$
  - Subject to $\left(w \cdot \Phi(x_i)\right) \geq \rho - \xi_i; \xi_i \geq 0$
- The discriminant function is
  - $f(x) = \sum_i \alpha_i k(x, x_i) - \rho$
  - It is positive for nominal points and negative for anomalies

Support Vector Data Description (Tax & Duin, 2004)

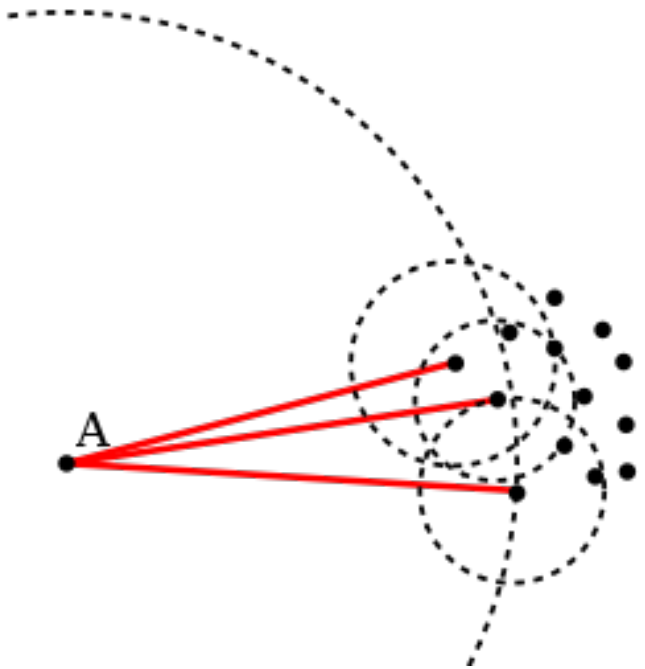
- Find the smallest hypersphere in feature space that contains $1 - \nu$ of the data points
- Solution to
  - $\min_{R,a} R^2 + C \sum_{i=1}^{N} \xi_i$
  - Subject to $\|x_i - a\|^2 \leq R^2 + \xi_i;\ \ \xi_i \geq 0$
- Generally only works well for the Gaussian kernel



Saeid Homayouni

LOF: Local Outlier Factor
(Breunig, et al., 2000)

- Distance from x to its k-th nearest neighbor divided by the average distance of each of those neighbors to their k-th nearest neighbors
- [The actual calculation is slightly more complex.]

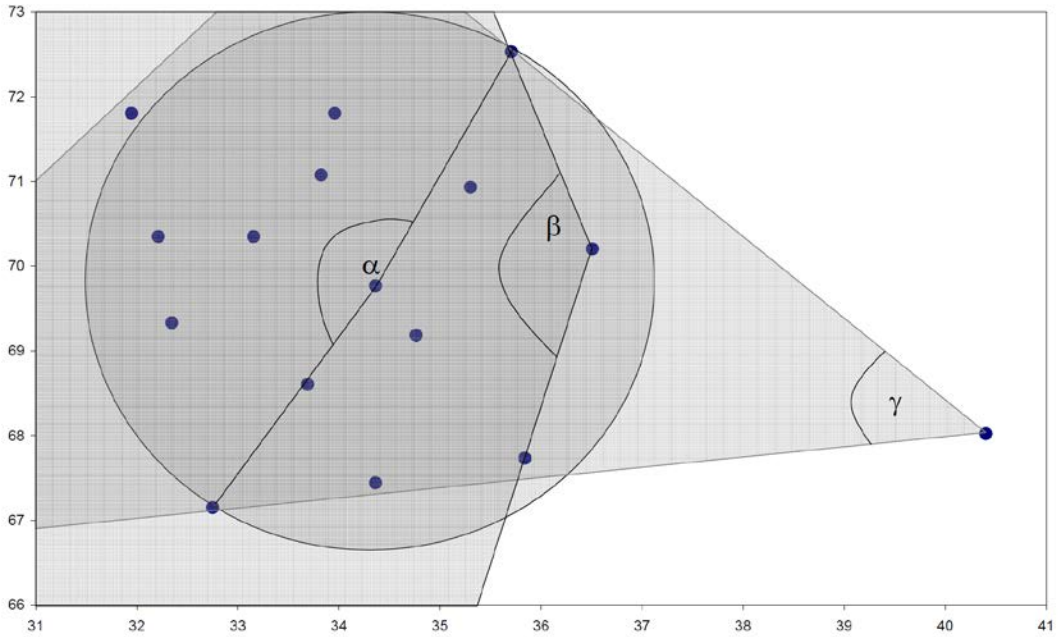


Breunig, et al.,

Angle-Based Outlier Detector (ABOD)

Kriegel, et al., 2008

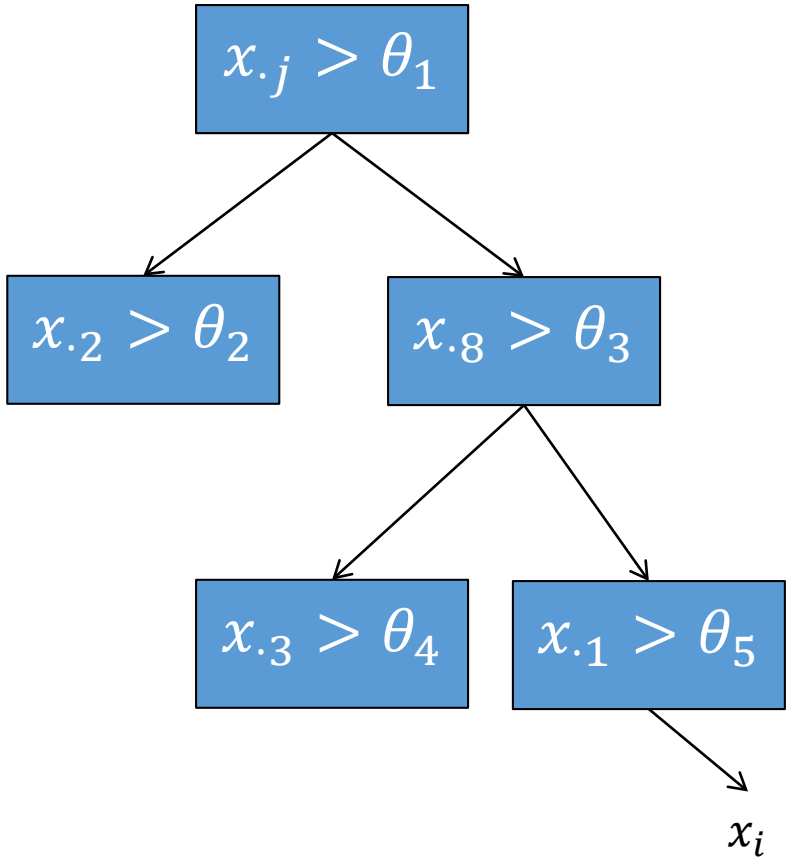
- Let $\angle(x, x_1, x_2)$ be the angle between x_1 and x_2 as viewed from x
- The anomaly score for x is the variance of $Var\left[\angle\left(x, x_i, x_j\right)\right]$ for all x_i, x_j in the training data



Breunig, et al.,

Isolation Forest [Liu, Ting, Zhou, 2011]

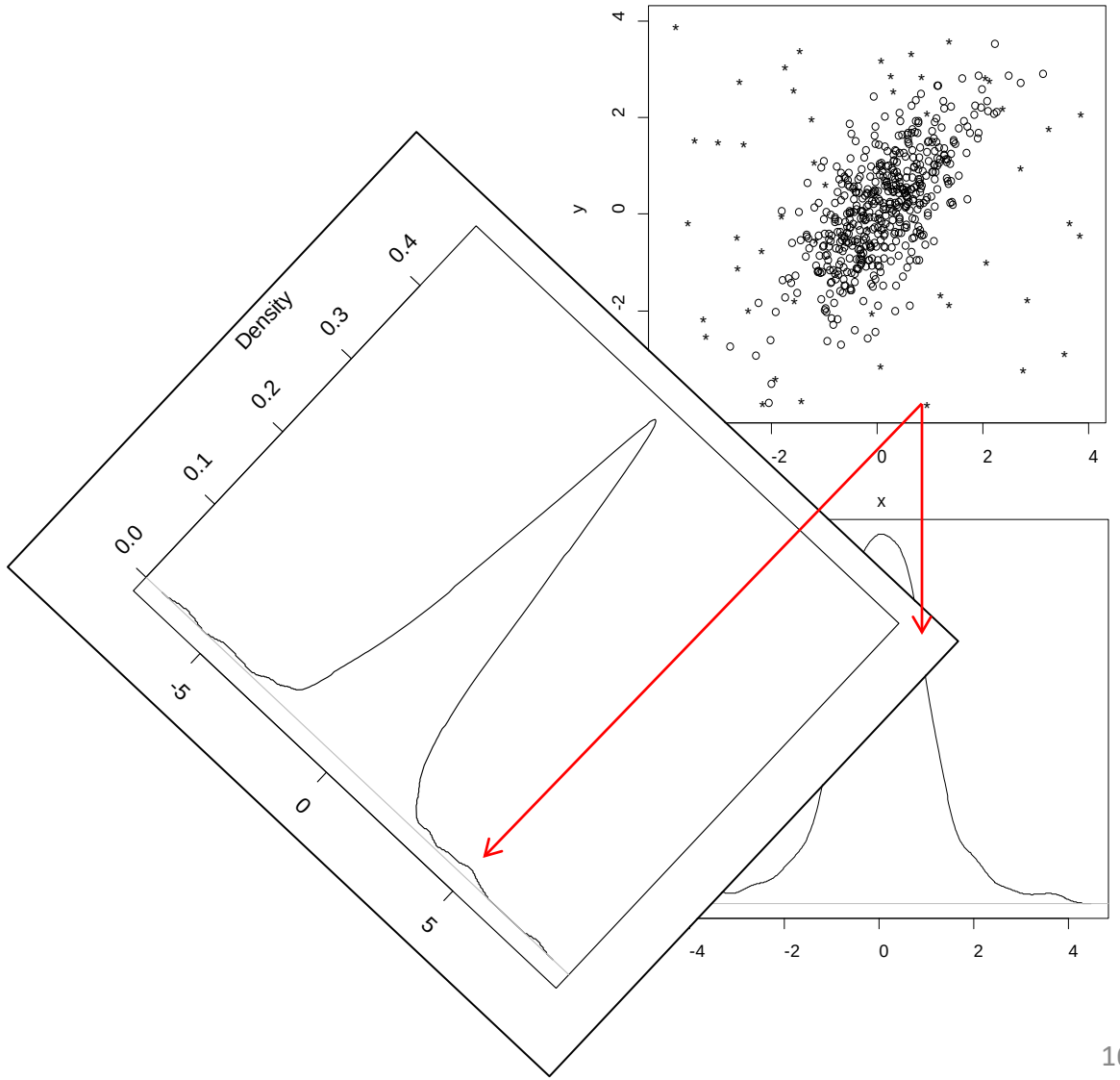
- Construct a fully random binary tree
  - choose attribute j at random
  - choose splitting threshold θ_1 uniformly from $[\min(x_{.j}), \max(x_{.j})]$
  - until every data point is in its own leaf
  - let $d(x_i)$ be the depth of point x_i
- repeat 100 times
  - let $\bar{d}(x_i)$ be the average depth of x_i
  - $score(x_i) = 2^{-\left(\frac{\bar{d}(x_i)}{r(x_i)}\right)}$
    - $r(x_i)$ is the expected depth



LODA: Lightweight Online Detector of Anomalies

[Pevny, 2016]

- $\Pi_1, \ldots, \Pi_M$ set of M sparse random projections
  - Let $w_m = (0, \ldots, 0)$
  - Choose $\sqrt{d}$ elements of w_m and set them to normal random variate
  - $\Pi_m(x) = w_m \cdot x$
- $f_1, \ldots, f_M$ corresponding 1-dimensional density estimators
  - Pevny uses optimal histograms
- $S(x) = -\frac{1}{M}\sum_m \log f_m(x)$ average "surprise"



Benchmarking Study

[Andrew Emmott]

- Most AD papers only evaluate on a few datasets
- Often proprietary or very easy (e.g., KDD 1999)
- Research community needs a large and growing collection of public anomaly benchmarks

[Emmott, Das, Dietterich, Fern, Wong, 2013; KDD ODD-2013]
[Emmott, Das, Dietterich, Fern, Wong. 2016; arXiv 1503.01158v2]

Benchmarking Methodology

- Select 19 data sets from UC Irvine repository
- Choose one or more classes to be “anomalies”; the rest are “nominals”
- Manipulate
  - Relative frequency
  - Point difficulty
  - Irrelevant features
  - Clusteredness
- 20 replicates of each configuration
- Result: 25,685 Benchmark Datasets

19 Selected Data Sets

Steel Plates Faults
Gas Sensor Array Drift
Image Segmentation
Landsat Satellite
Letter Recognition
OptDigits
Page Blocks
Shuttle
Magic Gamma
Skin

Waveform
Yeast
Abalone
Communities and Crime
Concrete Compressive Strength
Wine
Year Prediction
Spambase
Particle

Systematic Variation of Relevant Aspects

- Point difficulty: How deeply are the anomaly points buried in the nominals?
  - Fit supervised classifier (kernel logistic regression)
  - Point difficulty: $P(\hat{y} = "nominal"|x)$ for anomaly points
- Relative frequency:
  - sample from the anomaly points to achieve target values of α
- Clusteredness:
  - greedy algorithm selects points to create clusters or to create widely separated points
- Irrelevant features
  - create new features by random permutation of existing feature values
- Result: 25,685 Benchmark Datasets

Metrics

- AUC (Area Under ROC Curve)
  - ranking loss: probability that a randomly-chosen anomaly point is ranked above a randomly-chosen nominal point
  - transformed value: $\log \frac{AUC}{1-AUC}$
- AP (Average Precision)
  - area under the precision-recall curve
  - average of the precision computed at each ranked anomaly point
  - transformed value: $\log \frac{AP}{\mathbb{E}[AP]} = \log LIFT$

Filtering Out Impossible Benchmarks

- For each algorithm and each benchmark
  - Check whether we can reject the null hypothesis that the achieved AUC (or AP) is better than random guessing
  - If a benchmark dataset is too hard for all algorithms, then we delete it from the benchmark collection

Control Baselines

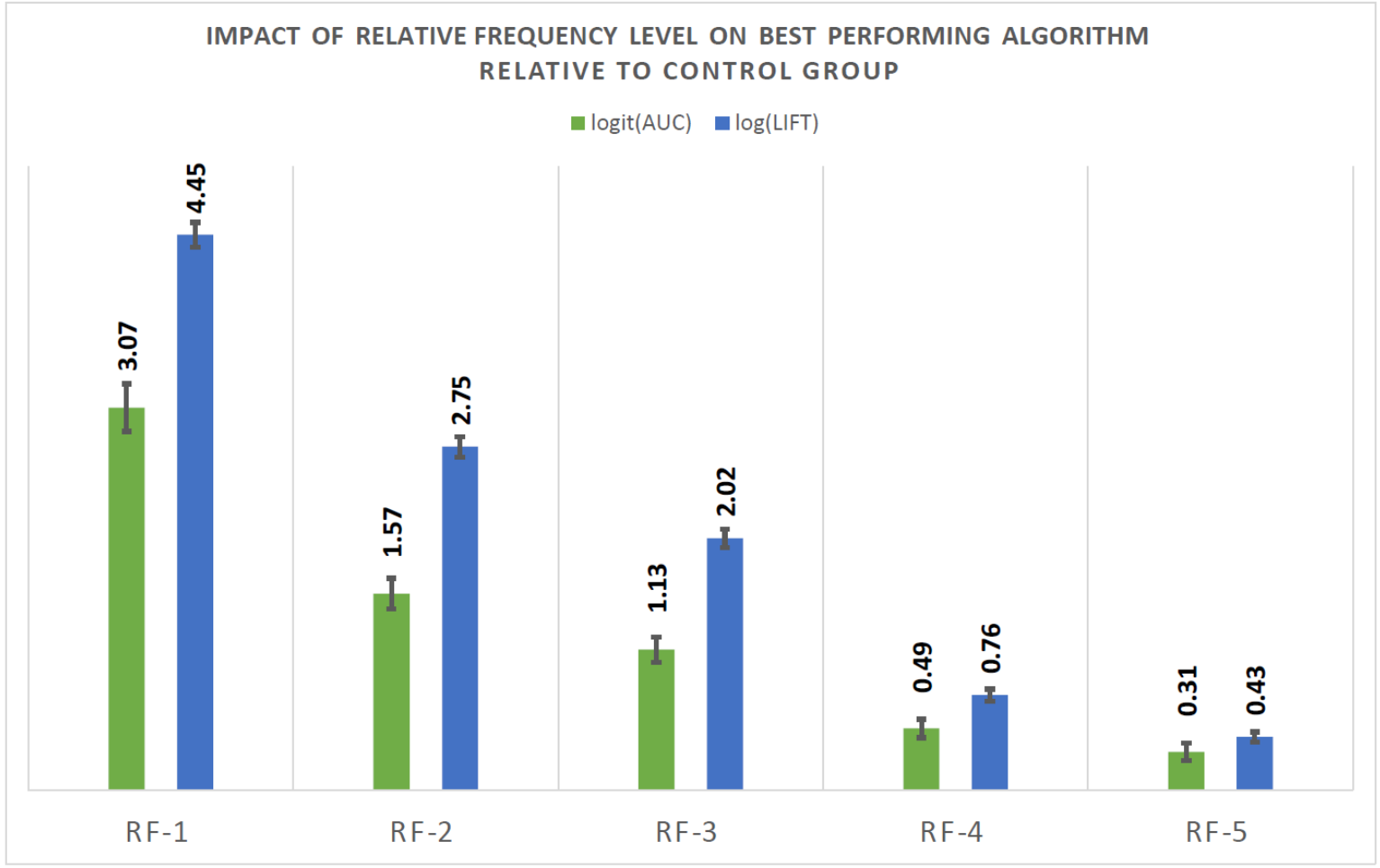
- Control Data Set
  - Nominals: standard d-dimensional multivariate Gaussian
  - Anomalies: uniform in the $[-4, +4]^d$ hypercube

- Control Algorithm
  - Distance to overall mean

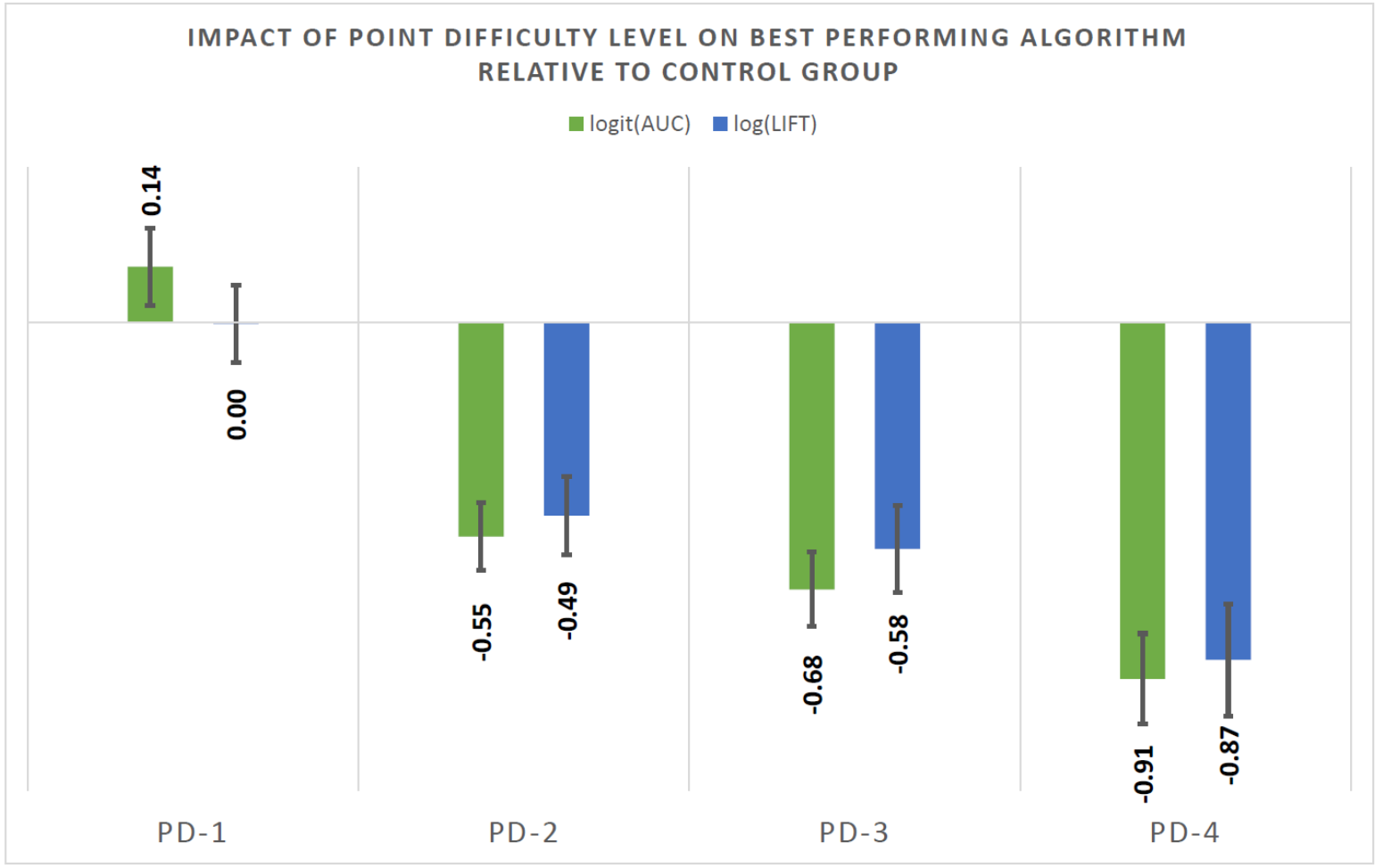
Analysis

- Linear ANOVA
  - $metric \sim rf + pd + cl + ir + mset + algo$
    - rf: relative frequency
    - pd: point difficulty
    - cl: normalized clusteredness
    - ir: irrelevant features
    - mset: “Mother” set
    - algo: anomaly detection algorithm
- Validate the effect of each factor
- Assess the *algo* effect while controlling for all other factors

Effect of Relative Frequency

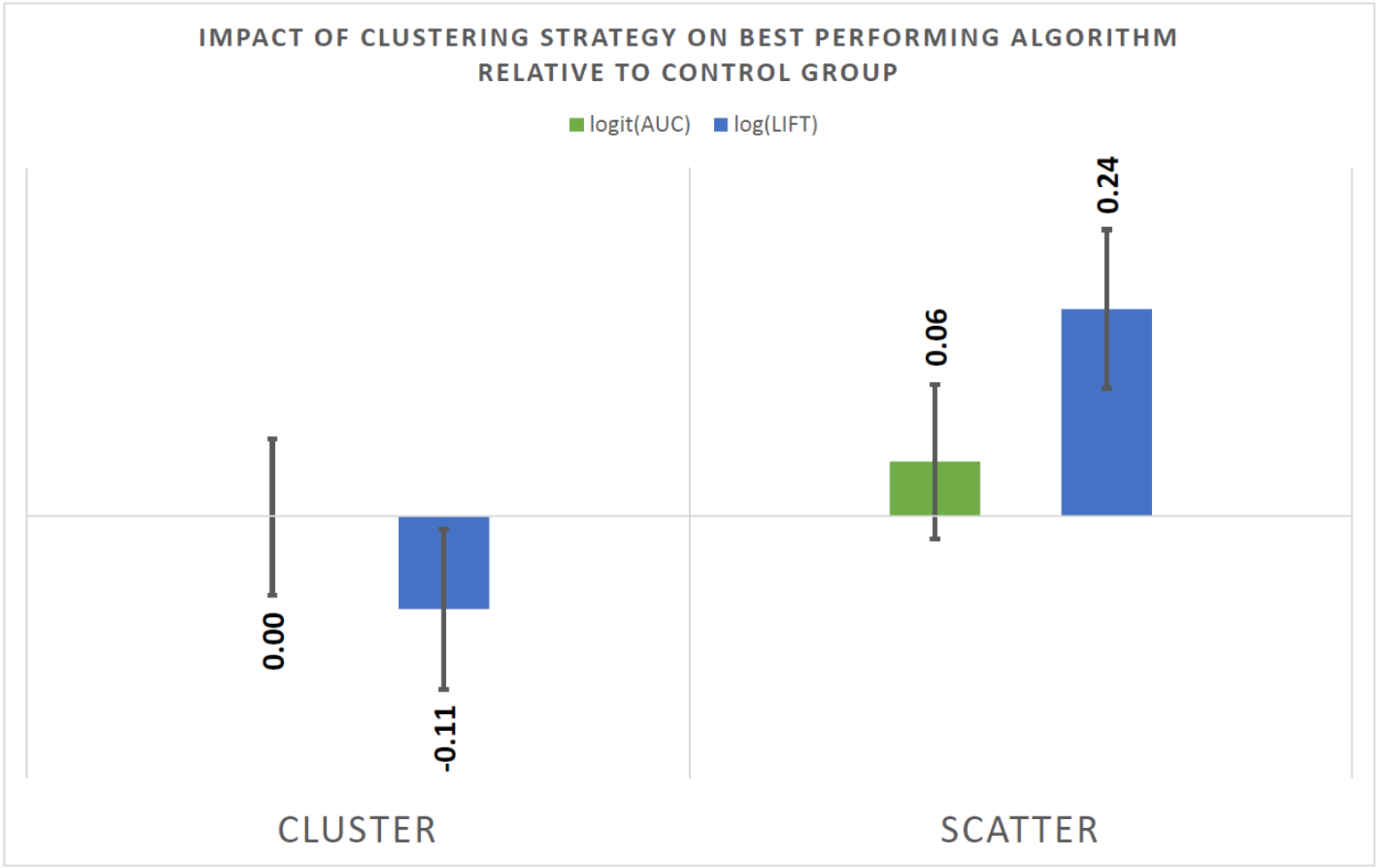


Effect of Point Difficulty



0-0.17 0.17-0.33 0.33-0.5 0.5-1.0

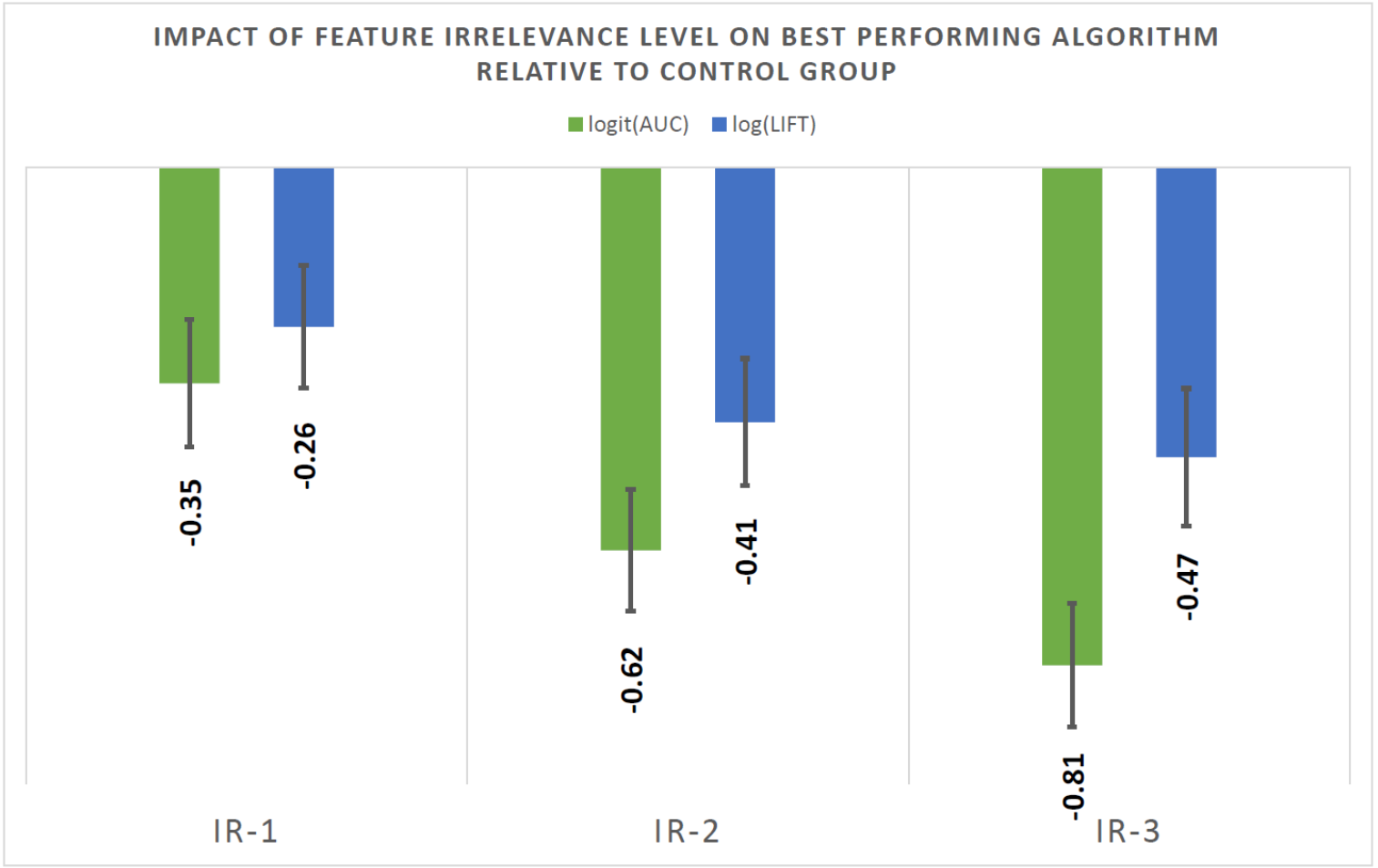
Effect of Clusteredness



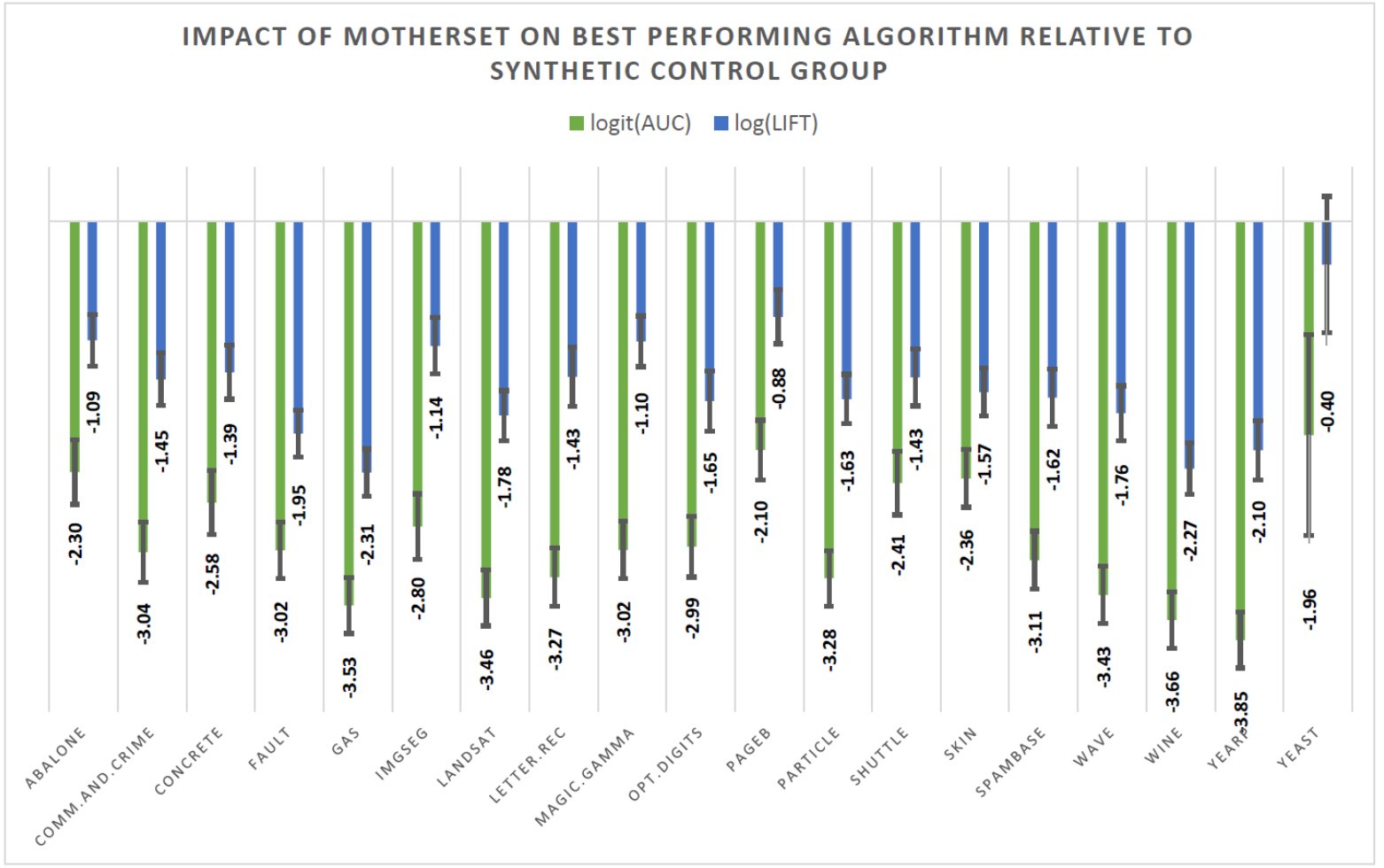
$\sigma^2_{anom} < \sigma^2_{nom}$

$\sigma^2_{anom} > \sigma^2_{nom}$

Effect of Irrelevant Features

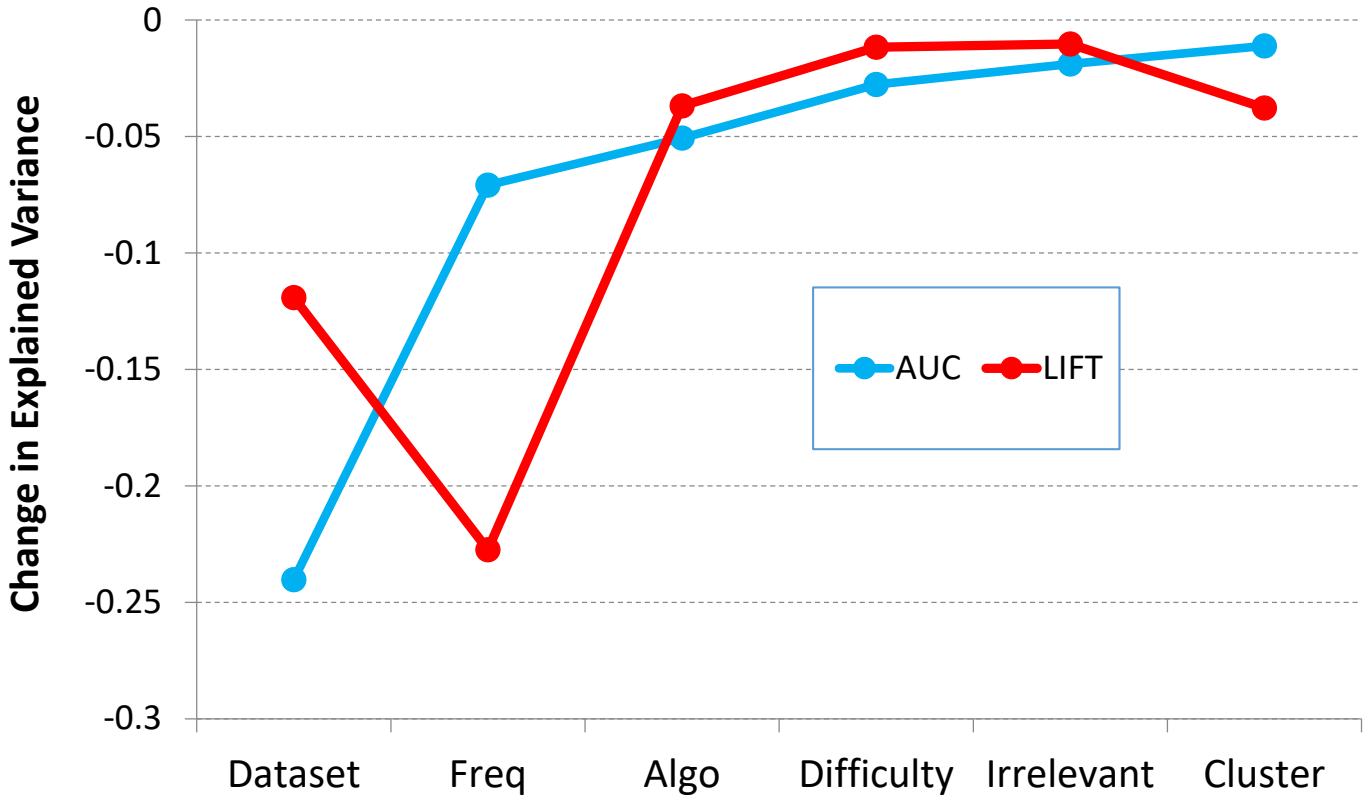


Choice of UCI Dataset



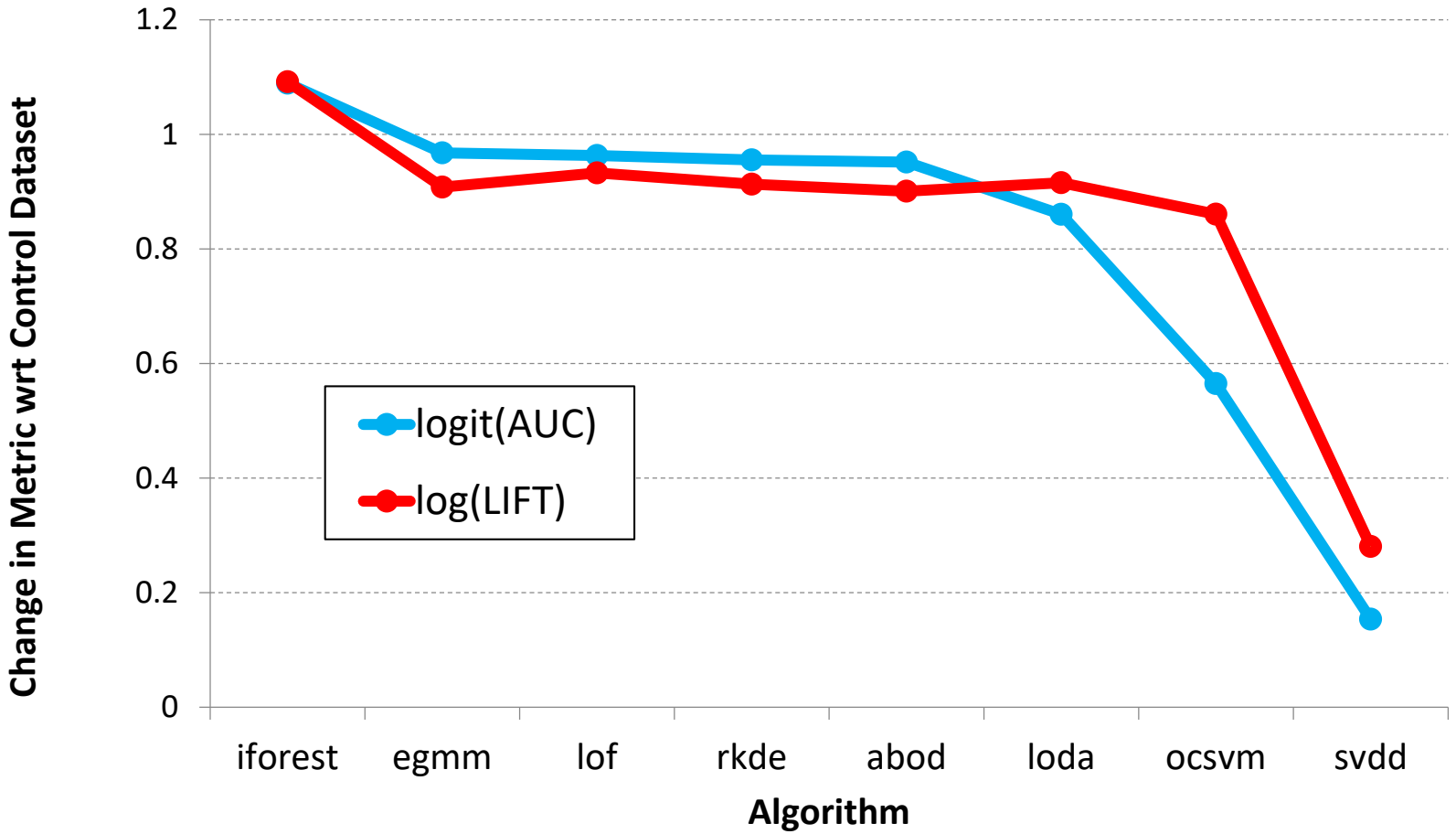
All datasets are more difficult than our synthetic control.

What Matters the Most?

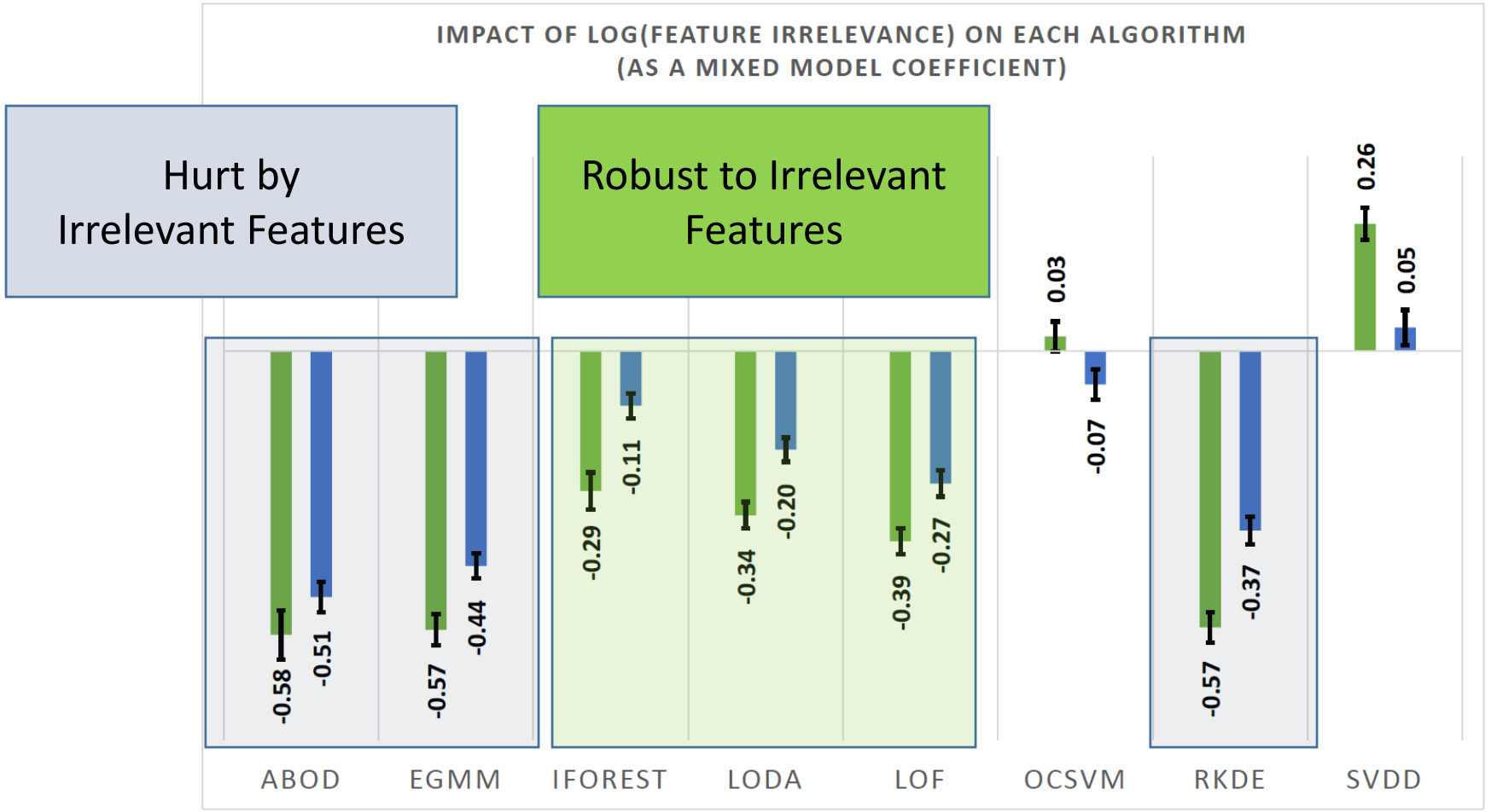


- Problem and Relative Frequency!
- Choice of algorithm ranks third

Algorithm Comparison

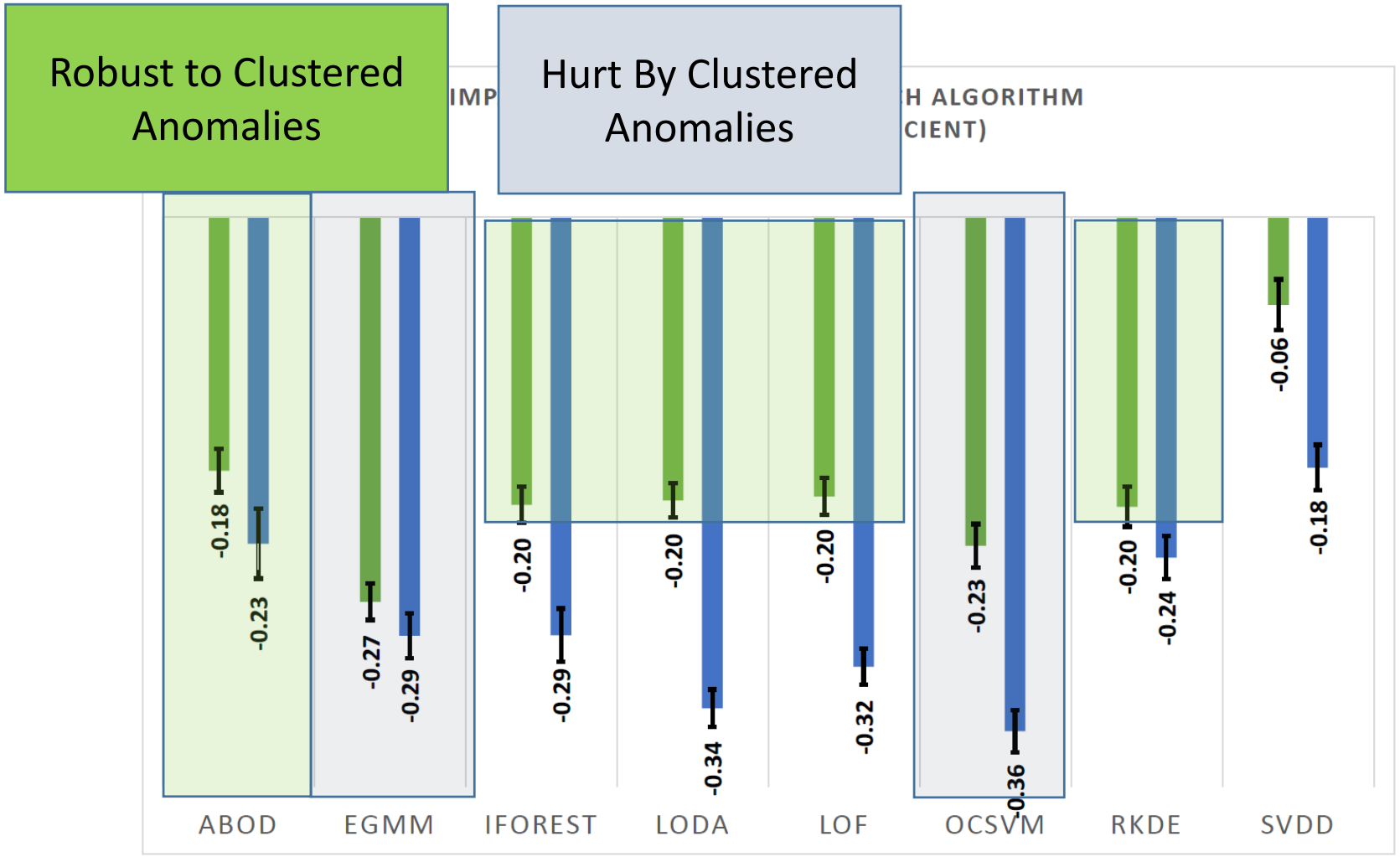


Algorithm Strengths and Weaknesses



iForest is least damaged by irrelevant features

Effect of Clusteredness



iForest Advantages

- Most robust to irrelevant features
  - for both AUC and LIFT
- Second most robust to clustered anomaly points
  - for AUC

Towards a Theory of Anomaly Detection

[Siddiqui, et al.; UAI 2016]

- Existing theory on sample complexity
  - Density Estimation Methods:
    - Exponential in the dimension d
  - Quantile Methods (OCSVM and SVDD):
    - Polynomial sample complexity

- Experimentally, many anomaly detection algorithms learn very quickly (e.g., 500-2000 examples)
- New theory: Rare Pattern Anomaly Detection

Pattern Spaces

- A pattern $h: \Re^d \rightarrow \{0,1\}$ is an indicator function for a measurable region in the input space
  - Examples:
    - Half planes
    - Axis-parallel hyper-rectangles in $[-1,1]^d$
- A pattern space $\mathcal{H}$ is a set of patterns (countable or uncountable)

Rare and Common Patterns

- Let μ be a fixed measure over $\Re^d$
  - Typical choices:
    - uniform over $[-1, +1]^d$
    - standard Gaussian over $\Re^d$
- $\mu(h)$ is the measure of the pattern defined by h
- Let p be the "nominal" probability density defined on $\Re^d$ (or on some subset)
- $p(h)$ is the probability of pattern h
- A pattern h is τ-rare if

$$f(h) = \frac{p(h)}{\mu(h)} \leq \tau$$

- Otherwise it is τ-common

Rare and Common Points

- A point x is τ-rare if there exists a τ-rare h such that $h(x) = 1$
- Otherwise a point is τ-common

- Goal: An anomaly detection algorithm should output all τ-rare points and not output any τ-common points

PAC-RPAD

- Algorithm $\mathcal{A}$ is PAC-RPAD for
  - pattern space $\mathcal{H}$,
  - measure μ,
  - parameters τ, ϵ, δ

  if for any probability density p and any τ, $\mathcal{A}$ draws a sample from p and with probability $1 - \delta$ detects all τ-rare points and rejects all $(\tau + \epsilon)$-commons in the sample

- ϵ allows the algorithm some margin for error
- If a point is between τ-rare and $(\tau + \epsilon)$-common, the algorithm can treat it arbitrarily
- Running time polynomial in $\frac{1}{\epsilon}$, $\frac{1}{\delta}$, and $\frac{1}{\tau}$, and some measure of the complexity of $\mathcal{H}$

RAREPATTERNDETECT

- Draw a sample of size $N(\epsilon, \delta)$ from p
- Let $\hat{p}(h)$ be the fraction of sample points that satisfy h
- Let $\hat{f}(h) = \frac{\hat{p}(h)}{\mu(h)}$ be the estimated rareness of h
- A query point x_q is declared to be an anomaly if there exists a pattern $h \in \mathcal{H}$ such that $h(x_q) = 1$ and $\hat{f}(h) \leq \tau$.

Results

- Theorem 1: For any finite pattern space $\mathcal{H}$, RAREPATTERNDETECT is PAC-RPAD with sample complexity

$$N(\epsilon, \delta) = O\left(\frac{1}{\epsilon^2}\left(\log|\mathcal{H}| + \log\frac{1}{\delta}\right)\right)$$

- Theorem 2: For any pattern space $\mathcal{H}$ with finite VC dimension $\mathcal{V}_{\mathcal{H}}$, RAREPATTERNDETECT is PAC-RPAD with sample complexity

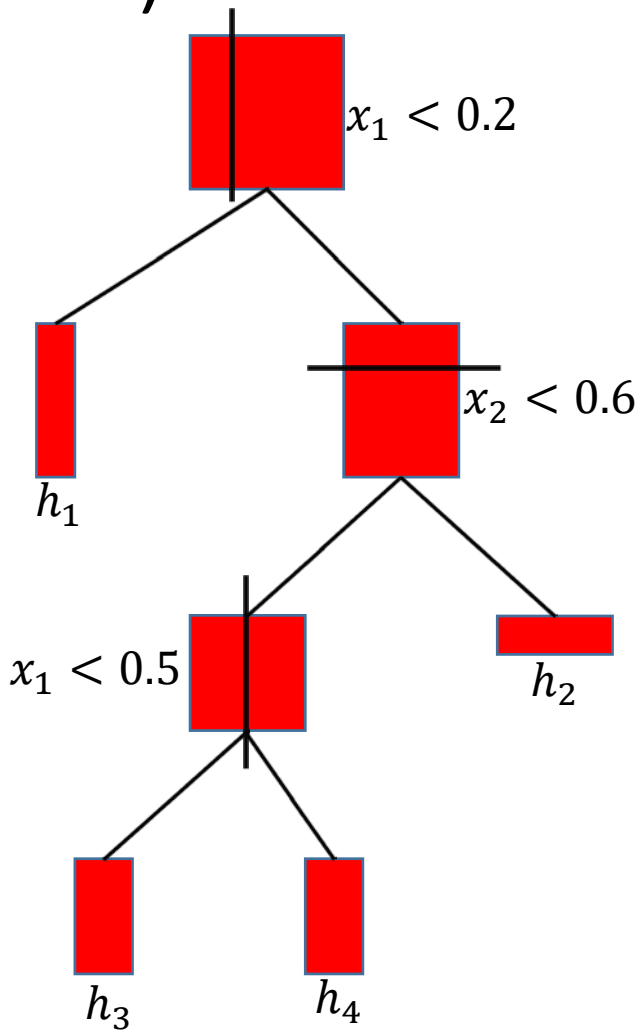
$$N(\epsilon, \delta) = O\left(\frac{1}{\epsilon^2}\left(\mathcal{V}_{\mathcal{H}} \log\frac{1}{\epsilon^2} + \log\frac{1}{\delta}\right)\right)$$

Examples of PAC-RPAD $\mathcal{H}$

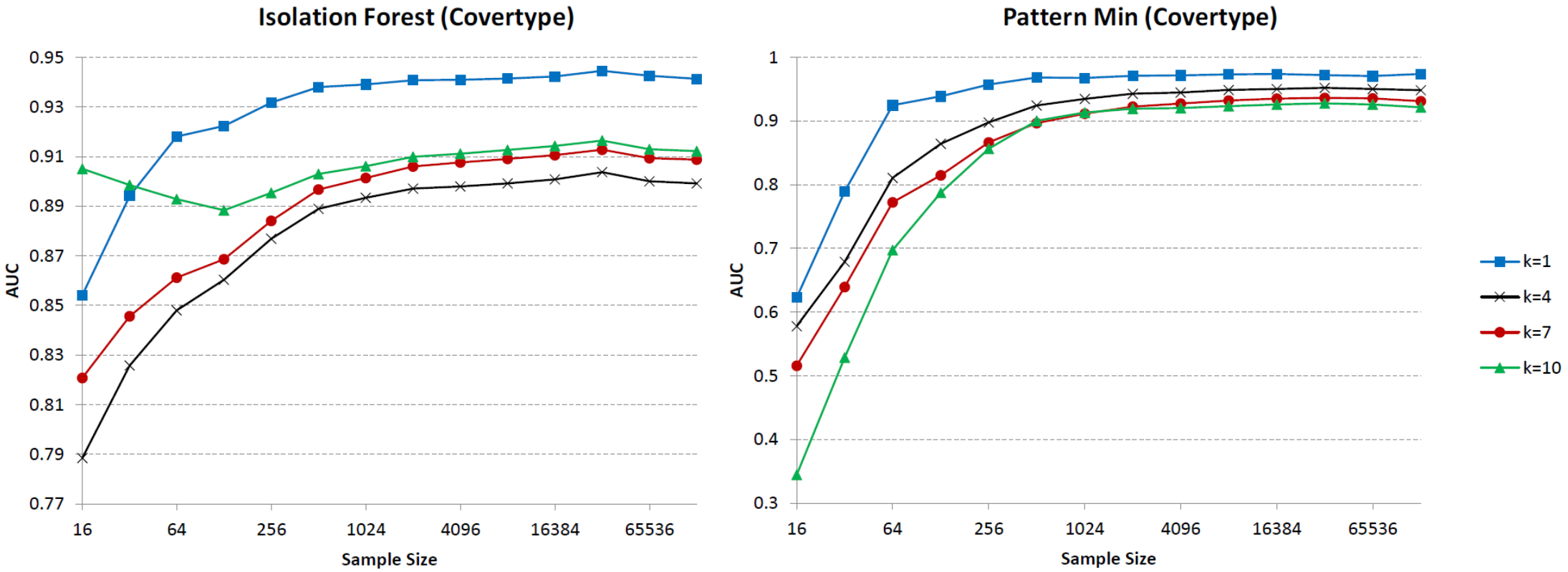
- Half spaces
- Axis-aligned hyper-rectangles (related to iForest leaves)
- Stripes (equivalent to LODA's histogram bins)
- Ellipsoids
- Ellipsoidal shells (difference of two ellipsoidal level sets)

Isolation RPAD (aka Pattern Min)

- Grow an isolation forest
  - Each tree is only grown to depth k
  - Each leaf defines a pattern h
  - μ is the volume (Lebesgue measure)
  - Compute $\hat{f}(h)$ for each leaf
- Details
  - Grow the tree using one sample
  - Estimate $\hat{f}$ using a second sample
  - Score query point(s)

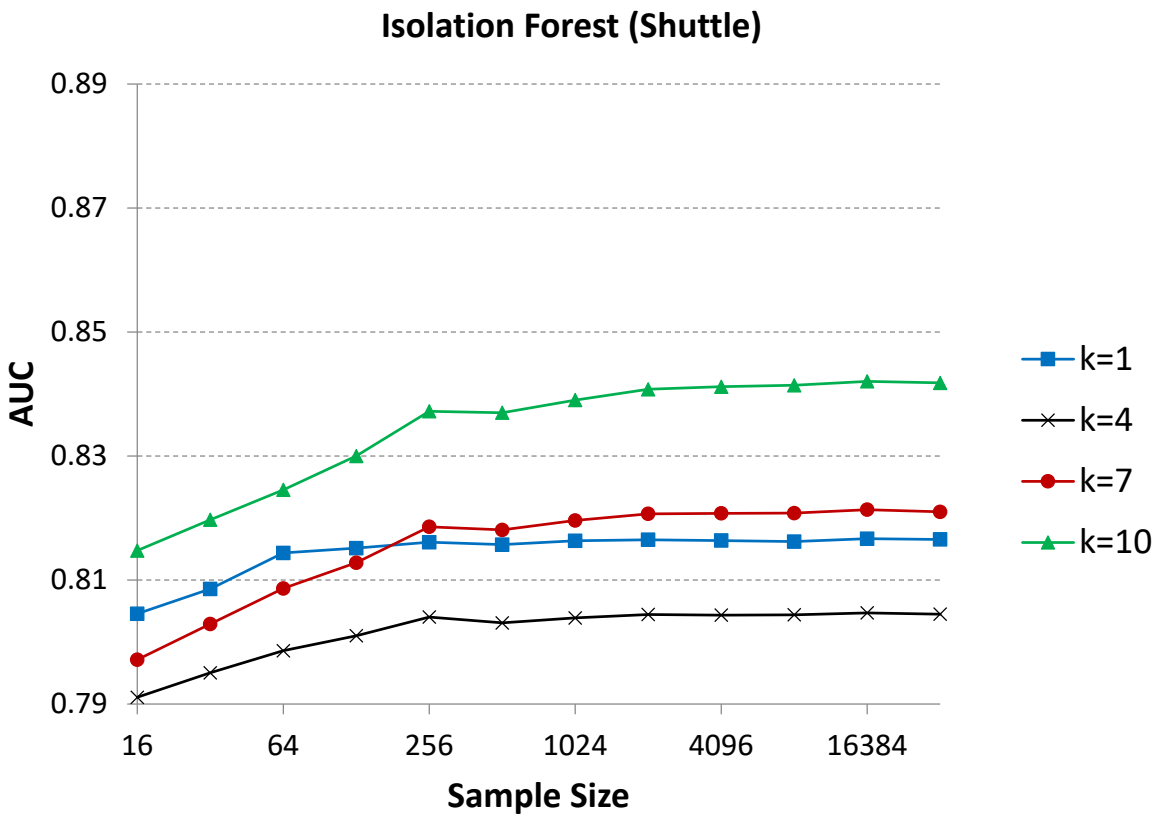


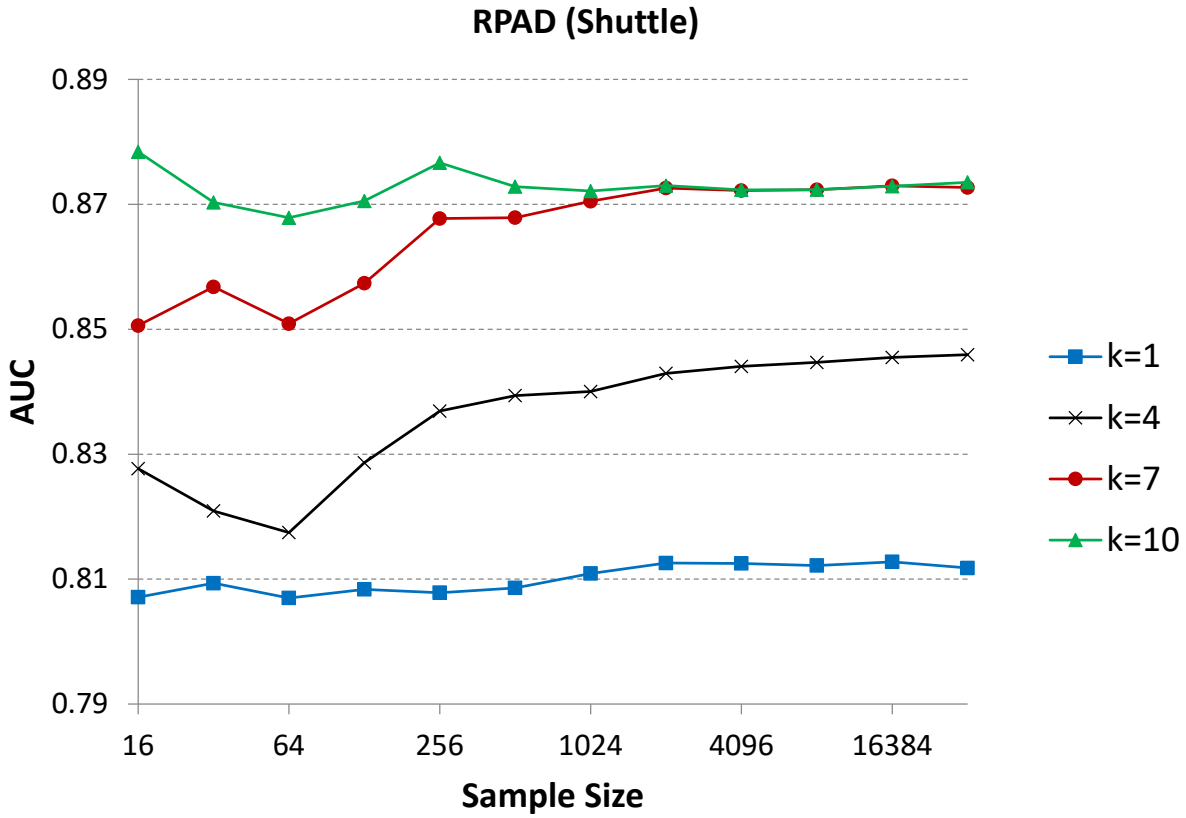
Results: Covertype



- PatternMin is slower, but eventually beats IFOREST

Results: Shuttle





- PatternMin is consistently beats iForest for $k > 1$

RPAD Conclusions

- The PAC-RPAD theory seems to capture the behavior of algorithms such as IFOREST
- It is easy to design practical RPAD algorithms
- Theory requires extension to handle sample-dependent pattern spaces $\mathcal{H}$

Discussion

- RPAD theory does not explain the good AUC results
  - Can we develop a PAC theory for anomaly ranking?
- iForest trains on small subsamples of the data
  - This gives better performance. Why?
- How large should the ensembles be?
  - iForest, LODA, EGMM
- Deep Learning for Anomaly Detection
  - Deep Density Estimation (e.g., Masked Autoregressive Flow)
  - Generalized Fisher Discriminants
  - GANs
  - Need to compare against simple baselines

Citations

- Breunig, M., Kriegel, H-P., Ng, R., Sander, J. (2000). LOF: Identifying Density-Based Local Outliers. SIGMOD 2000.
- Emmott, A. F., Das, S., Dietterich, T. G., Fern, A., Wong, W.-K. (2013). Systematic construction of anomaly detection benchmarks from real data. *ODD '13 Proceedings of the ACM SIGKDD Workshop on Outlier Detection and Description*. (pp. 16-21). Later draft at arXiv 1503.01158.
- Kim, J., Scott, C., (2008). Robust Kernel Density Estimation. JMLR 13, 2529-2565.
- Kriegel, H-P., Schubert, M., Zimek, A. (2008). Angle-based outlier detection in high-dimensional data. *KDD 2008.* 444-452.
- Liu, F., Ting, K. M., Zhou, Z-H. (2008). Isolation Forest. *ICDM 2008.*
- Pevny, T. (2016). Loda: Lightweight Online Detector of Anomalies. *Machine Learning*, 102(2), 275–304
- Scholkopf, B., Williamson, R., Smola, A., Shawe-Taylor, J., Platt, J. (1999). Support Vector Method for Novelty Detection. NIPS 1999.
- Siddiqui, M. A., Fern, A., Dietterich, T. G., Das, S. (2016). Finite Sample Complexity of Rare Pattern Anomaly Detection**.** *Uncertainty in Artificial Intelligence (UAI-2016)*
- Tax, D., Duin, R. (2004). Support Vector Data Description. *Machine Learning*, 54, 45-66.