ROBUST KERNEL DENSITY ESTIMATION

JooSeuk Kim and Clayton Scott

Dept. of EECS, University of Michigan, Ann Arbor, MI, USA
E-mail: {stannum, clayscot}@umich.edu

ABSTRACT

In this paper, we propose a method for robust kernel density estimation. We interpret a KDE with Gaussian kernel as the inner product between a mapped test point and the centroid of mapped training points in kernel feature space. Our robust KDE replaces the centroid with a robust estimate based on M-estimation [1]. The iteratively re-weighted least squares (IRWLS) algorithm for M-estimation depends only on inner products, and can therefore be implemented using the kernel trick. We prove the IRWLS method monotonically decreases its objective value at every iteration for a broad class of robust loss functions. Our proposed method is applied to synthetic data and network traffic volumes, and the results compare favorably to the standard KDE.

Index Terms— kernel density estimation, M-estimator, outlier, kernel feature space, kernel trick

1. INTRODUCTION

Kernel density estimators (KDEs) are perhaps the most common nonparametric density estimators for multivariate data. They are essential ingredients in the toolbox of researchers in statistical data analysis, data mining, and machine learning [2, 3], and form the backbone of numerous methods for classification, clustering, and level set estimation. In addition, they are well-known to be consistent density estimators under suitable conditions on the bandwidth tending to zero [2, 4]. In this paper we propose a *robust* kernel density estimator (RKDE) that is resistant to outlying data points.

The need for a robust KDE arises when analyzing either contaminated or uncontaminated data. Contaminated data refers to data consisting of realizations from both a nominal or "clean" distribution in addition to outlying or anomalous measurements. In an increasing number of applications, data arise from high-dimensional or high-throughput systems where the nominal distribution itself may be quite complex and not amenable to parametric modeling. Robust nonparametric estimation of the nominal distribution is therefore relevant for problems such as anomaly detection, where an anomaly detector may be defined as a level set of the nominal distribution, and outlier ranking, where outliers are ordered with respect to the nominal density contour on which they lie.

Analysis of uncontaminated data also benefits from a robust KDE. For example, in one common approach to clustering, contours are defined to be the connected components of a level set of an underlying density. As the level is varied, a hierarchical clustering or cluster tree is swept out [5]. Since KDEs are averages of kernels centered at the data points, for sufficiently low levels of a KDE, outliers will eventually form isolated clusters which is contrary to our intuition. Alternatively, clusters may be defined as the basins of attraction (with respect to a hill-climbing algorithm) of modes of a density, which will again give rise to one-point clusters near outliers. We may summarize this phenomenon by saying that the standard KDE tends to overfit the data in the vicinity of isolated points.

Our work focuses on KDEs based on the Gaussian kernel with isotropic covariance. This kernel is a kernel in both popular uses of the word: it is nonnegative and integrates to one and is therefore appropriate for KDEs, but it is also an inner product kernel, meaning it may be viewed as evaluating an inner product between points in a high-dimensional Hilbert space [6]. This allows us to write the KDE as an inner product between a test point and a sample average (or centroid) of training points in feature space. We achieve a RKDE by estimating this centroid with an M-estimator, a technique developed for robust estimation of centroids in robust parametric statistics. The standard algorithm for M-estimators, known as the iterative re-weighted least squares (IRWLS), depends only on inner products and may therefore be implemented efficiently using the kernel function. The RKDE has the form

$$\widehat{f}(\mathbf{x}) = \sum_{i=1}^{n} w_i k(\mathbf{x}, \mathbf{x}_i)$$

where k is a kernel function, $w_i \geq 0$, $\sum_{i=1}^{n} w_i = 1$, and w_i tends to be downweighted for outlying data points. We present simulations that demonstrate significant improvement of RKDEs over standard KDEs, as assessed by the Kullback-Liebler divergence, and we also illustrate the application of the RKDE to a problem in Internet traffic analysis.

2. MULTIVARIATE M-ESTIMATOR

Given i.i.d samples $\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_n \in \mathbb{R}^d$ generated from a multivariate Gaussian pdf $f(\mathbf{x}; \boldsymbol{\theta})$ with unknown mean $\boldsymbol{\theta}$ and covariance matrix $\sigma^2 I$, the maximum likelihood (ML) estimator of $\boldsymbol{\theta}$ is

$$\begin{aligned}\hat{\boldsymbol{\theta}} &= \arg\max_{\boldsymbol{\theta}} \sum_{i=1}^{n} \ln f(\mathbf{x}_i; \boldsymbol{\theta}) \\ &= \arg\min_{\boldsymbol{\theta}} \sum_{i=1}^{n} \rho(\|\mathbf{x}_i - \boldsymbol{\theta}\|) \end{aligned} \quad (1)$$

where $\rho(x) = x^2/2$. We can find $\widehat{\boldsymbol{\theta}}$ by taking the derivative of (1) with respect to $\boldsymbol{\theta}$ and setting it equal to zero, i.e,

$$-\sum_{i=1}^{n} \left(\mathbf{x}_i - \widehat{\boldsymbol{\theta}}\right) \cdot \frac{\psi(\|\mathbf{x}_i - \widehat{\boldsymbol{\theta}}\|)}{\|\mathbf{x}_i - \widehat{\boldsymbol{\theta}}\|} = 0 \quad (2)$$

where $\psi = \rho'$ and we define $\psi(0)/0 := \lim_{x \to 0} \psi(x)/x$. For $\rho(x)$ function in the ML estimation case, $\psi(x) = x$ and (2) has a closed form solution $\widehat{\boldsymbol{\theta}} = \frac{1}{n}\sum_{i=1}^{n} \mathbf{x}_i$, the sample mean.

The sample mean is seriously affected by the presence of outliers. In 1964, Huber proposed a robust estimator called an M-estimator, M for maximum likelihood [7]. An M-estimator is the

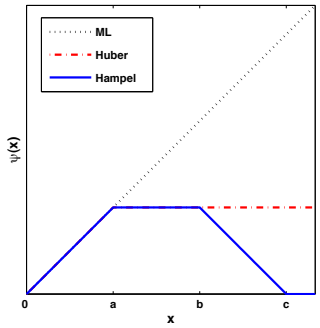


Fig. 1. The comparison between three different ψ functions: ML case with Gaussian, Huber's, and Hampel's.

solution of (1) but where different *loss* functions are used. For example, Huber's loss function is defined as

$$\rho(x) = \begin{cases} x^2/2 & , |x| \le a \\ a|x| - a^2/2 & , a < |x| \end{cases}$$

so that the ψ function is

$$\psi(x) = \begin{cases} -a & , x < -a \\ x & , a \le x \le a \\ a & , a < x. \end{cases} \quad (3)$$

For outliers such that $\|\mathbf{x} - \boldsymbol{\theta}\| > k$, $\psi(x)$ in (3) is constant while $\psi(x) = x$ is monotonically increasing. Therefore, we can expect M-estimators to be less affected by outliers than the ML estimator. Another well known ψ function is Hampel's ψ function which is defined as

$$\psi(x) = (\text{sgn } x) \begin{cases} |x| & , 0 \le |x| < a \\ a & , a \le |x| < b \\ a\frac{c-|x|}{c-b} & , b \le |x| < c \\ 0 & , c \le |x|. \end{cases} \quad (4)$$

$\psi(x)$ in (4) gives more penalty to outliers than that in (3). The plot of these ψ functions are shown in Fig 1.

For ψ functions in (3) and (4), the equation (2) does not have closed form solution, but can be solved by the iteratively re-weighted least squares (IRWLS) method. We extend the result of [1] for univariate M-estimators to the multivariate case. Because of space limitations, we present only this initial result. Future work will develop local and global optimality properties.

Theorem 1 *Consider the optimization problem (1). Assume that $\psi(x)/x$ is bounded and monotone decreasing for $x > 0$. Then, for any initial guess $w^{(0)}$, the following iteratively re-weighted least squares method produces a sequence $\{\boldsymbol{\theta}^{(k)}\}_{k=1}^{\infty}$ such that the objective function value monotonically decreases at every iteration.*

$$\boldsymbol{\theta}^{(k)} = \sum_{i=1}^{n} w_i^{(k-1)} \mathbf{x}_i$$

$$\widetilde{w}_i^{(k)} = \frac{\psi(\|\mathbf{x}_i - \boldsymbol{\theta}^{(k)}\|)}{\|\mathbf{x}_i - \boldsymbol{\theta}^{(k)}\|}, \quad w_i^{(k)} = \frac{\widetilde{w}_i^{(k)}}{\sum_{i=1}^{n} \widetilde{w}_i^{(k)}}$$

Proof For notational convenience, we sometimes omit the superscript k. At step k+1, given $\boldsymbol{\theta}^{(k)}$, define the surrogate functions

$$U_i(x) = \rho(r_i) - \frac{1}{2} r_i \psi(r_i) + \frac{\psi(r_i)}{2r_i} x^2$$

where $r_i = \|\mathbf{x}_i - \boldsymbol{\theta}^{(k)}\|$. We can show that

$$U_i(x) \ge \rho(x) \quad \forall x, \quad U_i(r_i) = \rho(r_i).$$

To see this, let $z(x) = U_i(x) - \rho(x)$. Note that

$$z'(x) = \frac{\psi(r_i)}{r_i} x - \psi(x).$$

Since $\psi(x)/x$ is monotone decreasing for $x > 0$,

$$z'(x) \begin{cases} \le 0 & , \quad \text{for} \quad 0 < x \le r_i \\ \ge 0 & , \quad \text{for} \quad r_i \le x \end{cases}.$$

Therefore, $z(x) \ge z(r_i) = 0$ for all x.

By letting $\widetilde{w}_i^{(k)} = \psi(r_i)/r_i$, the next iterate $\boldsymbol{\theta}^{(k+1)}$ is

$$\begin{aligned} \boldsymbol{\theta}^{(k+1)} &= \sum_{i=1}^{n} w_i^{(k)} \mathbf{x}_i = \frac{\sum_{i=1}^{n} \widetilde{w}_i^{(k)} \mathbf{x}_i}{\sum_{i=1}^{n} \widetilde{w}_i^{(k)}} \\ &= \arg\min_{\boldsymbol{\theta}} \frac{1}{2} \sum_{i=1}^{n} \widetilde{w}_i^{(k)} \|\mathbf{x}_i - \boldsymbol{\theta}\|^2 \\ &= \arg\min_{\boldsymbol{\theta}} \sum_{i=1}^{n} U_i(\|\mathbf{x}_i - \boldsymbol{\theta}\|). \end{aligned}$$

Then, the objective value at $\boldsymbol{\theta}^{(k+1)}$ is

$$\begin{aligned} &\sum_{i=1}^{n} \rho(\|\mathbf{x}_i - \boldsymbol{\theta}^{(k+1)}\|) \le \sum_{i=1}^{n} U_i(\|\mathbf{x}_i - \boldsymbol{\theta}^{(k+1)}\|) \\ &\le \sum_{i=1}^{n} U_i(\|\mathbf{x}_i - \boldsymbol{\theta}^{(k)}\|) = \sum_{i=1}^{n} \rho(\|\mathbf{x}_i - \boldsymbol{\theta}^{(k)}\|). \end{aligned}$$

Since U_i is quadratic, the second inequality is strict if $\boldsymbol{\theta}^{(k+1)} \ne \boldsymbol{\theta}^{(k)}$. Therefore, the objective function value monotonically decreases at every iteration.

3. KERNEL DENSITY ESTIMATION: A VIEW FROM KERNEL FEATURE SPACE

Let $\mathbf{x}_1, \ldots, \mathbf{x}_n \in \mathbb{R}^d$ be a random sample from a distribution with density $f(\mathbf{x})$. A kernel density estimate of f, also called a Parzen window estimate, is a nonparametric estimate given by

$$\widehat{f}(\mathbf{x}) = \frac{1}{n} \sum_{i=1}^{n} k(\mathbf{x}, \mathbf{x}_i)$$

where $k(\mathbf{x}, \mathbf{x}_i)$ is a kernel function. The most commonly used kernel function is a Gaussian kernel

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

We can view the KDE as performing operations in a high dimensional feature space. For the Gaussian kernel, there exists a mapping $\Phi : \mathbb{R}^d \to \mathcal{H}$, where $\mathcal{H}$ is an infinite dimensional Hilbert space, such that $k(\mathbf{x}, \mathbf{x}_i) = \langle \Phi(\mathbf{x}), \Phi(\mathbf{x}_i) \rangle$ [6].

From this point of view, the KDE can be expressed as

$$\widehat{f}(\mathbf{x}) = \frac{1}{n} \sum_{i=1}^{n} \langle \Phi(\mathbf{x}), \Phi(\mathbf{x}_i) \rangle = \left\langle \Phi(\mathbf{x}), \frac{1}{n} \sum_{i=1}^{n} \Phi(\mathbf{x}_i) \right\rangle. \quad (5)$$

Therefore, given an input $\mathbf{x}$, the KDE outputs the inner product between $\Phi(\mathbf{x})$ and the sample mean, or the centroid, of the $\Phi(\mathbf{x}_i)$.

4. ROBUST KERNEL DENSITY ESTIMATION

As mentioned earlier, the sample mean can be drastically influenced by outliers. Therefore, we replace the sample mean in (5) with a robust mean estimator $\widehat{\mathbf{m}}$. By adopting the M-estimator criterion explained in Section 2, $\widehat{\mathbf{m}}$ is

$$\hat{\mathbf{m}} = \arg\min_{\mathbf{m}\in\mathcal{H}} \sum_{i=1}^{n} \rho\left(\|\Phi(\mathbf{x}_i) - \mathbf{m}\|\right)$$

where ρ is a robust loss function. The only remaining issue is how to compute $\|\Phi(\mathbf{x}_i) - \mathbf{m}^{(k)}\|$ in the IRWLS method. This can be done by observing

$$\begin{aligned}\|\Phi(\mathbf{x}_i) - \mathbf{m}^{(k)}\|^2 &= \left\langle \Phi(\mathbf{x}_i) - \mathbf{m}^{(k)}, \Phi(\mathbf{x}_i) - \mathbf{m}^{(k)} \right\rangle \\ &= \left\langle \Phi(\mathbf{x}_i), \Phi(\mathbf{x}_i) \right\rangle - 2\left\langle \Phi(\mathbf{x}_i), \mathbf{m}^{(k)} \right\rangle + \left\langle \mathbf{m}^{(k)}, \mathbf{m}^{(k)} \right\rangle.\end{aligned}$$

Since $\mathbf{m}^{(k)} = \sum_{j=1}^{n} w_j^{(k-1)} \Phi(\mathbf{x}_j)$, we have

$$\begin{aligned}\left\langle \Phi(\mathbf{x}_i), \Phi(\mathbf{x}_i) \right\rangle &= k(\mathbf{x}_i, \mathbf{x}_i) \\ \left\langle \Phi(\mathbf{x}_i), \mathbf{m}^{(k)} \right\rangle &= \sum_{j=1}^{n} w_j^{(k-1)} k(\mathbf{x}_i, \mathbf{x}_j) \\ \left\langle \mathbf{m}^{(k)}, \mathbf{m}^{(k)} \right\rangle &= \sum_{j=1}^{n}\sum_{l=1}^{n} w_j^{(k-1)} w_l^{(k-1)} k(\mathbf{x}_j, \mathbf{x}_l).\end{aligned}$$

The IRWLS algorithm for RKDE is summarized below.

1. Initialize $w_i^{(0)}$. Let $k = 1$.
2. Compute $\|\Phi(\mathbf{x}_i) - \mathbf{m}^{(k)}\|$ using the equations above.
3. Update $\widetilde{w}_i^{(k)} = \psi\left(\|\Phi(\mathbf{x}_i) - \mathbf{m}^{(k)}\|\right) / \|\Phi(\mathbf{x}_i) - \mathbf{m}^{(k)}\|$.
4. Normalize $\widetilde{w}_i^{(k)}$ to get $w_i^{(k)}$.
5. If the algorithm converges, $\widehat{\mathbf{m}} = \sum_{i=1}^{n} w_i^{(k)} \Phi(\mathbf{x}_i)$. Otherwise, let $k \leftarrow k+1$ and go to step 2.

The resulting RKDE is

$$\begin{aligned}\widehat{f}(\mathbf{x}) = \langle \Phi(\mathbf{x}), \widehat{\mathbf{m}} \rangle &= \left\langle \Phi(\mathbf{x}), \sum_{i=1}^{n} w_i^{(k)} \Phi(\mathbf{x}_i) \right\rangle \\ &= \sum_{i=1}^{n} w_i^{(k)} \langle \Phi(\mathbf{x}), \Phi(\mathbf{x}_i) \rangle = \sum_{i=1}^{n} w_i^{(k)} k(\mathbf{x}, \mathbf{x}_i).\end{aligned}$$

If the ρ function is not convex, we may be stuck in a local minimum, and thus good initialization is required. We initialize $w_i^{(0)}$ such that $\widehat{\mathbf{m}}^{(1)}$ is the geometric median of $\Phi(\mathbf{x}_i)$, i.e.,

$$\widehat{\mathbf{m}}^{(1)} = \arg\min_{\mathbf{m}} \sum_{i=1}^{n} \|\Phi(\mathbf{x}_i) - \mathbf{m}\|$$

since the geometric median is a more robust initializer than the centroid. Note that the geometric median is the solution of (1) with $\rho(x) = x$, or equivalently, the solution of (2) with $\psi(x) = \text{sgn}(x)$. The geometric median can be found by Weiszfeld's algorithm [8], which is equivalent to the above IRWLS method with $\psi(x) = \text{sgn}(x)$.

5. EXPERIMENTS

5.1. Experimental setting

We demonstrate our algorithm on 1D and 2D synthetic data and real world data. For all experiments, we use the Gaussian kernel as the kernel function and the bandwidth is chosen as the least square cross-validation estimator [9]. Hampel's ψ function in (4) is used and the parameters a, b, and c are selected by the following heuristic.

First, we compute $d_i = \|\Phi(\mathbf{x}_i) - \mathbf{m}^{(1)}\|$, the distance between the geometric median $\mathbf{m}^{(1)}$ and $\Phi(\mathbf{x}_i)$. Then, a is set to the median of $\{d_i\}$ (the median absolute deviation), b is the 95th percentile of $\{d_i\}$, $c = \max\{d_i\}$.

5.2. Synthetic data

In the first example, we experiment with 1- dimensional data. The pdf is a Gaussian mixture given by

$$f_1(x) = 0.5\phi(x; 0, 1) + 0.5\phi(x; 10, 1)$$

where $\phi(x; \mu, \sigma)$ is a univariate Gaussian pdf with mean μ and variance σ^2. Different numbers of outliers are generated from a uniform distribution from -5 to 15. The number of data samples, n, is 200 and the numbers of outliers m are $0, 10, 20, 40$. For $m = 40$, the results are shown in Fig 2. From the figure, we can see that KDEs are affected by outliers such that the density estimates have small bumps over the regions where outliers exist. On the other hand, the RKDE method gives better density estimates which do not have such features, and thus can be considered less affected by outliers.

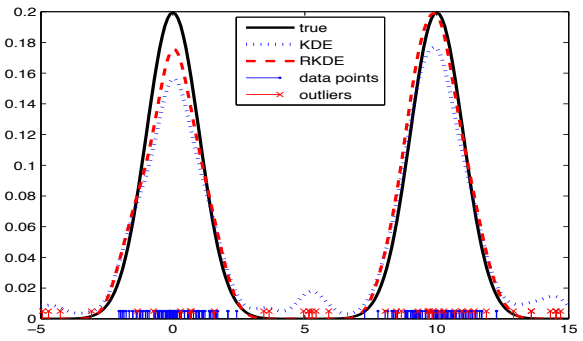


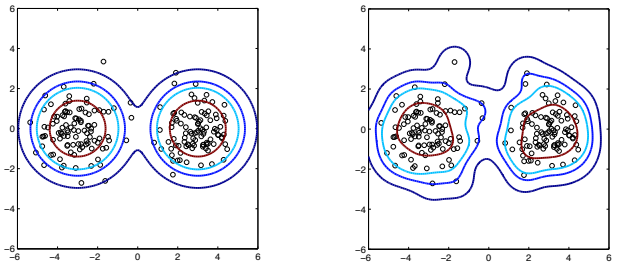
Fig. 2. The comparison of density estimates. True density (solid line), KDE (dotted line), and RKDE (dashed line).

For a 2-dimensional example, 200 data samples are generated from a Gaussian mixture given by

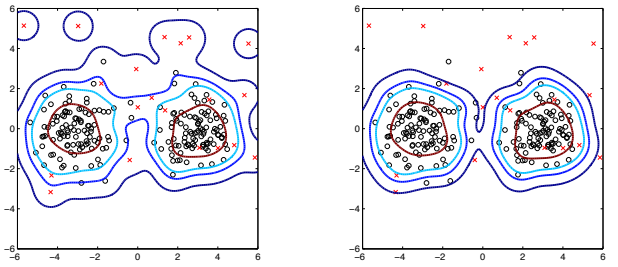
$$f_2(\mathbf{x}) = 0.5\phi(\mathbf{x}; \boldsymbol{\mu}_1, \Sigma_1) + 0.5\phi(\mathbf{x}; \boldsymbol{\mu}_2, \Sigma_2)$$

where $\boldsymbol{\mu}_1 = [-3, 0]^T$, $\boldsymbol{\mu}_2 = [3, 0]^T$ and $\Sigma_1 = \Sigma_2 = I$. The numbers of outliers from a uniform distribution over $[-6, 6] \times [-6, 6]$ are $m = 0, 10, 20$, and, 40. The results for $m = 20$ are shown in Fig 3. From (c), we can see that the lower level contour of the KDE encloses the outliers. However, in (d), the contours of the RKDE are much closer to that of the true density.

We compare RKDE with KDE quantitatively, using the Kullback-Leibler (KL) divergence as the performance measure. We compute both $D_{KL}(f||\hat{f})$ and $D_{KL}(\hat{f}||f)$ where $\hat{f}$ is either the KDE or RKDE and f is the true density. For each 1d and 2d example, the average KL divergence over 100 simulations is shown in Table 1. For both cases, $D_{KL}(\hat{f}||f)$ for RKDE is always better than



(a) True density (b) KDE without outliers



(c) KDE with outliers (d) Robust KDE with outliers

Fig. 3. Contours of densities along with data samples from true density (o) and outliers (x).

Table 1. KL divergence between true density f_{true} and f_{KDE} and f_{RKDE} for simulated data

1d data	n = 200			
	m = 0	m = 10	m = 20	m = 40
$D_{KL}\left(f_{true}\|\|f_{KDE}\right)$	0.0333	0.0507	0.0698	0.1065
$D_{KL}\left(f_{true}\|\|f_{RKDE}\right)$	0.0906	0.0529	0.0509	0.0695
$D_{KL}\left(f_{KDE}\|\|f_{true}\right)$	0.0335	0.1122	0.1919	0.3390
$D_{KL}\left(f_{RKDE}\|\|f_{true}\right)$	0.0331	0.0290	0.0330	0.0509

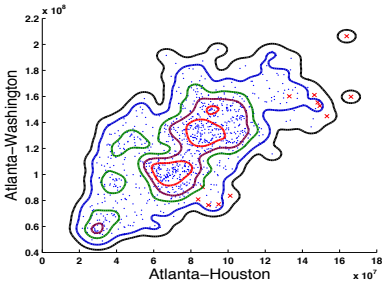
2d data	n = 200			
	m = 0	m = 10	m = 20	m = 40
$D_{KL}\left(f_{true}\|\|f_{KDE}\right)$	0.0745	0.0968	0.1228	0.1795
$D_{KL}\left(f_{true}\|\|f_{RKDE}\right)$	0.0868	0.0756	0.0702	0.0883
$D_{KL}\left(f_{KDE}\|\|f_{true}\right)$	0.0878	0.2525	0.4405	0.7536
$D_{KL}\left(f_{RKDE}\|\|f_{true}\right)$	0.0670	0.0707	0.0789	0.1060

that for KDE. On the other hand, for $D_{KL}(f||\hat{f})$, KDE is somewhat better than robust KDE when outliers do not exist. This can be considered as a tradeoff between efficiency and robustness. However, $D_{KL}(\hat{f}||f)$ is the one that reflects the robustness to outliers and we see the biggest improvement here.

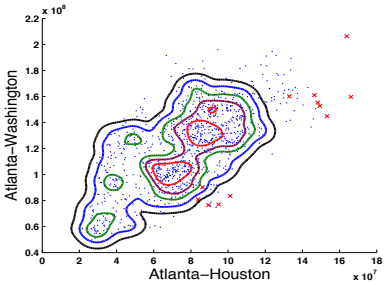
5.3. Network anomaly detection

We also experiment with Internet traffic flowing over links in the Abilene network [10]. In Fig 4, each point corresponds to the total traffic volume (measured in bytes) for a given ten minute interval over a pair of links, from Houston and Washington to Atlanta. In this setting, we may want to determine which points represent potentially anomalous behavior, such as might be caused by malicious activities (e.g., denial of service attacks). We could do this by estimating the underlying nominal density and thresholding the estimated density value at the given level [11]. Thus, the RKDE is applicable here.

The contour plots of the KDE and RKDE are shown in Fig 4 (a) and (b), respectively. While the KDE overfits the data, the RKDE method finds a more reasonable estimate of the nominal density. Furthermore, outliers detected using a global method based on principal components analysis (PCA) (having access to all data on all links in the network) [12, 13] are marked as 'x' and the result shows that the RKDE downweights these points in estimating the nominal density relative to the KDE.



(a) KDE



(b) RKDE

Fig. 4. Contours of density estimates for Abilene network traffic (thresholds at the same quantiles: 20%, 50%, 70%, 95%, 99.5%). Outliers detected by PCA method are marked 'x'.

6. REFERENCES

[1] P. Huber, *Robust Statistics*, Wiley, New York, 1981.

[2] D. W. Scott, *Multivariate Density Estimation*, Wiley, New York, 1992.

[3] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning*, Springer, New York, 2001.

[4] Luc Devroye and Gabor Lugosi, "Combinatorial methods in density estimation," 2001.

[5] J. A. Hartigan, "Consistency of single linkage for high-density clusters," *Journal of the American Statistical Association*, vol. 76, pp. 388–394, 1981.

[6] B. Schölkopf and A. J. Smola, *Learning with Kernels*, MIT Press, Cambridge, MA, 2002.

[7] P. J. Huber, "Robust estimation of a location parameter," *Ann. Math. Statist*, vol. 35, pp. 73–101, 1964.

[8] E Weiszfeld, "Sur le point pour lequel la somme des distances de n points donnes est minimum," *Tohoku Math. Journal*, pp. 355–386, 1937.

[9] B.A. Turlach, "Bandwidth selection in kernel density estimation: A review," *Technical Report 9317, C.O.R.E. and Institut de Statistique, Université Catholique de Louvain*, 1993.

[10] "http://www.internet2.org" .

[11] P. Chhabra, C. Scott, E. Kolaczyk, and M. Crovella, "Distributed spatial anomaly detection," to appear at *IEEE INFOCOM 2008*.

[12] A. Lakhina, K. Papagiannaki, M. Crovella, C. Diot, E. Kolaczyk, and N. Taft, "Structural analysis of network traffic flows," *Proc. ACM SIGMETRICS/Performance*, 2004.

[13] A. Lakhina, M. Crovella, and C. Diot, "Diagnosing network-wide traffic anomalies," *Proc. ACM SIGCOMM*, 2004.