Nonparametric Assessment of Contamination in Multivariate Data Using Generalized Quantile Sets and FDR

Clayton Scott* and Eric Kolaczyk†

January 19, 2010

Abstract

Large, multivariate datasets from high-throughput instrumentation have become ubiquitous in the sciences. Frequently, it is of interest to characterize the measurements in these datasets by the extent to which they represent 'nominal' versus 'contaminated' instances. However, often the nature of even the nominal patterns in the data are unknown and potentially quite complex, making their explicit parametric modeling a daunting task. In this paper, we introduce a nonparametric method for the simultaneous annotation of multivariate data (called *MN-SCAnn*), by which one may produce an annotated ranking of the observations, indicating the relative extent to which each may or may not be considered nominal, while making minimal assumptions on the nature of the nominal distribution. In our framework each observation is linked to a corresponding generalized quantile set and, implicitly adopting a hypothesis testing perspective, each set is associated with a test, which in turn is accompanied by a certain false discovery rate. The combination of generalized quantile set methods with false discovery rate principles, in the context of contaminated data, is new, and estimation of the key underlying quantities requires that a number of issues be addressed. We illustrate *MN-SCAnn* through examples in two contexts: the pre-processing of flow cytometry data in bioinformatics, and the detection of anomalous traffic patterns in Internet measurement studies.

*Department of Electrical Engineering and Computer Science, University of Michigan, 1301 Beal Avenue, Ann Arbor, MI 48105 (email: cscott-at-eecs-dot-umich-dot-edu).

†Department of Mathematics and Statistics, Boston University, 111 Cummington Street, Boston, MA 02215 (email: kolaczyk-at-math-dot-bu-edu).

Keywords: Generalized quantile sets, false discovery rate, nonparametric multivariate outlier detection, monotone density estimation

1 Introduction

High-throughput data collection has become a prominent measurement paradigm across the sciences. Examples include DNA microarray technology and similar in biology, remote-sensing imaging in geography and the earth sciences, computer network traffic monitoring in the Internet, and the collection of consumer purchasing information in marketing and business. Given the often massive, automated and instrument-based nature of these methods of data collection, frequently it is the case that there is 'contamination' of some sort among the otherwise 'nominal' measurements, and it is desirable to be able to characterize observations by the extent to which they may be one or the other. Such characterizations may be used, for example, to separate out 'reliable' measurements from 'unreliable' ones, or to detect 'anomalous' observations amidst a background of otherwise 'typical' data. In many cases, the data are multivariate and sufficiently complicated in distribution, even under 'nominal' conditions, that the painstaking construction of an accurate parametric model is quite difficult. It is therefore desirable that techniques for the assessment of contamination in such data be nonparametric.

By way of motivation, consider the data in Fig. 1, representing measurements gathered on Internet traffic flowing over links in the Abilene network described in Sec. 5. Each point corresponds to the total traffic volume (measured in bytes) for a given ten minute interval over a pair of links to a given node in the network. In Fig. 1(a), for example, the node corresponds to Atlanta, and the links correspond to routes in the Abilene network to Atlanta from Houston and Washington. A useful goal in this setting is to design a system that takes this collection of roughly 1000 measurements and identifies the extent to which each point may represent potentially anomalous behavior, such as might be caused by malicious activities (e.g., denial of service (DoS) attacks). The output of such a system would be transmitted, for instance, to a network operator who might then conduct follow-up examinations on the nature of the most suspect data. An essential feature of this system

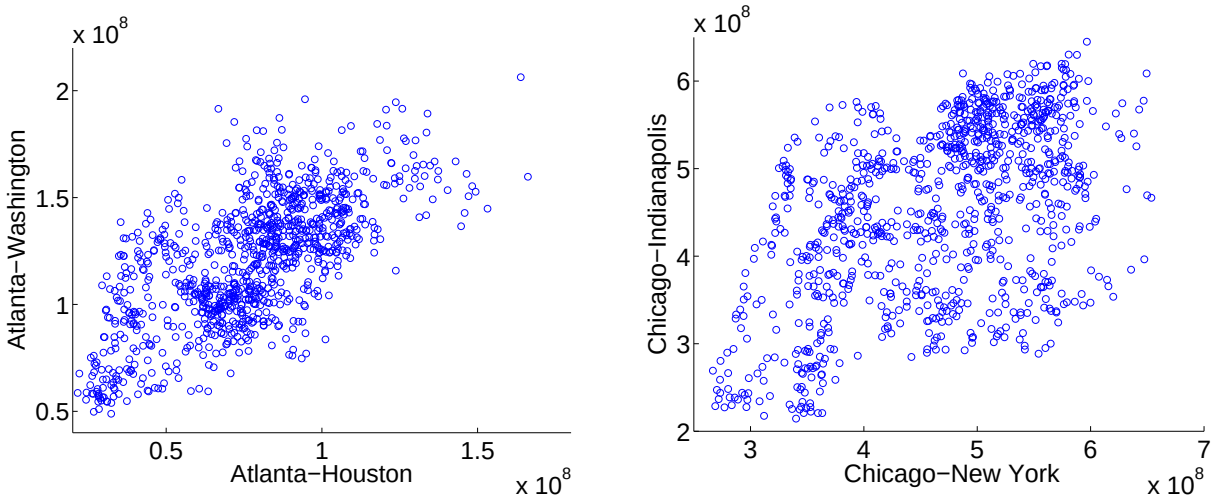


Figure 1: Examples of Contaminated Multivariate Data. Scatterplot of volume levels for traffic passing through select pairs of links at (a) Atlanta and (b) Chicago, in the Abilene network of Fig. 4 (a), over consecutive 10 minute intervals.

is that it make minimal assumptions about the nature of the typical data, as it would be required to apply equally well to data at any node in the network, such as the Chicago node, as shown in Fig. 1 (b), whose distributional characteristics are clearly different from those of the data for the Atlanta node.

In this paper we propose a framework well-suited to accomplish goals like the one just described. In particular, we propose a multivariate, nonparametric method for simultaneous contamination annotation, which we call *MN-SCAnn*. Formally, we suppose we observe independent and identically distributed measurements $X_i \in \mathbb{R}^d, i = 1, \ldots, n$ from a mixture distribution i.e.,

$$X_i \sim Q = (1 - \pi)P + \pi\mu \ , \tag{1}$$

where P is the distribution of the nominal data, μ is the distribution of the contaminating data (e.g., such as anomalies), and π is the *a priori* probability of obtaining a contaminated observation. The challenge here is that we assume P is unknown, as is π as well. However, we will allow that the user be willing or able to specify μ, which we will therefore consider known. Precise distributional assumptions are stated below.

We take as our goal to produce an annotated ranking of the observations X_i. We have in mind that the ranking serve to impose priorities through a basic ordering, while the annotations should provide some further indication as to the extent to which observations are actually likely to be contaminated. Our approach in this paper is to link each observation to a corresponding generalized quantile (GQ) set and, implicitly adopting a hypothesis testing perspective, to associate each set with a test. An inherent ordering of these sets yields a natural ranking, while the association of each test with a certain false discovery rate (FDR) yields an appropriate annotation.

GQ sets have a long history in statistics, going back at least to the 1970s, where they have been used to obtain robust estimates of location and scale (e.g., Sager (1978, 1979)) and in the study of the modality of a distribution (e.g., Hartigan (1987); Müller and Sawitzki (1992)). Consistency and rates of convergence have been established in Polonik (1997); Walther (1997). More closely related to our usage, they have been used to quantify multivariate data depth, with an eye towards assessing the outlying-ness of observations (e.g., see Liu et al. (1999) and references therein), and for the construction of classifiers (using uncontaminated observations from P) for predicting anomalies (Theiler and Cai, 2003; Schölkopf et al., 2001; Lanckriet et al., 2003; Steinwart et al., 2005; Scott and Nowak, 2006). But our combination of GQ set methods with FDR principles is new, and is motivated by the fact that, by incorporating a hypothesis testing element into our assessment of contamination, we are implicitly faced with conducting a large number of such tests simultaneously. FDR methods have received a great deal of attention in the statistics literature over the past decade (e.g., Benjamini and Hochberg (1995); Storey (2002, 2003); Genovese and Wasserman (2002); Efron et al. (2001)), and have emerged as the method of choice for quantifying error rates meaningfully in multiple testing situations, with applications now found in contexts ranging from wavelet denoising to the analysis of DNA microarrays. However, it is typically the case in such settings that the null distribution (i.e., P, in our notation) is assumed known, which it is important to note is *not* the case here. We show that FDR probabilities may nevertheless be estimated in the present context through our use of GQ sets and known μ. Additional discussion of related work is given in Sec. 6.

The assumption of known μ, of one sort or another, can be justified in various settings. For

example, gamma ray spectrometers, such as are found in high-energy astrophysics, are commonly used to detect radioactive isotopes. Often, particular isotopes having known energy spectra are targeted. The challenge is to detect such signals in the presence of significant and sometimes highly unpredictable background radiation. As another example, in the Internet traffic problem described above, if the basic measurements derive from data on traffic volume, network operators generally can at least provide a rough characterization of potentially anomalous levels of volume by considering network bandwidth and traffic history. Finally, in many situations, the assumption of *uniform* μ is intuitive and natural. For instance, this choice recently has been shown to optimize the worst-case detection rate among all choices for the unknown contamination distribution (El-Yaniv and Nisenson, 2007). Furthermore, all of the above cited works that apply GQ sets to anomaly prediction implicitly assume μ is a known distribution on anomalies (Theiler and Cai, 2003; Schölkopf et al., 2001; Lanckriet et al., 2003; Steinwart et al., 2005; Scott and Nowak, 2006). That is, these predictors are only optimal when in fact the anomalies truly follow the volume-defining measure μ. Ranking with respect to GQ sets and uniform μ coincides with the ranking determined by the so-called likelihood data depth (Fraiman and Meloche, 1999; Liu et al., 1999). The connection to data depth is discussed further in the concluding section. Finally, the validity of assuming μ known is confirmed through its practical utility. In addition to the experiments presented later, in a separate empirical study we have successfully applied *MN-SCAnn* to the distributed detection of anomalous events in network traffic data (Chhabra et al., 2008).

We also note that our experiments do not validate *MN-SCAnn* for dimensions greater that 10. Because of the difficulty in modeling μ in such cases, it may be necessary to reduce the dimension by first applying a standard method for dimensionality reduction.

The rest of this paper is organized as follows. In Sec. 3 we introduce the basic elements of our proposed methodology. Estimation of the corresponding GQ sets and FDR probabilities is addressed in Sec. 4. Some numerical illustrations are presented in Sec. 5, using both synthetic data and data from two different real-world contexts: the pre-processing of flow cytometry data in bioinformatics and the detection of anomalies in computer network traffic data (as described above). Finally, we close with a brief discussion in Sec. 6.

2 Supplemental Material

A full Matlab implementation of our software is available as supplemental material on the JCGS website. All proofs are gathered in an appendix that is also available as supplemental material on the JCGS website.

3 Methodology

Recall the model in Eqn. (1). To formalize this mixture model, we assume the existence of unobserved variables $Y_1, \ldots, Y_n$ such that (X_i, Y_i) are iid realizations of a random pair (X, Y), $Y \in \{0, 1\}$. Thus Q is the marginal of X, P and μ are the conditional distributions of X given $Y = 0$ and 1, respectively, and $\pi = \Pr(Y = 1)$.

Define $G_{P,\beta}$, the *generalized quantile (GQ) set of P at level β*, to be the P-measurable set with minimal μ-measure containing at least $\beta \in [0, 1]$ probability mass under P i.e.,

$$G_{P,\beta} := \arg\min\{\mu(G) : P(G) \geq \beta\} \ . \tag{2}$$

GQ sets coincide with level sets of the density of P with respect to μ when this density exists. Each mass β corresponds to a certain level of the density of P, and as β ranges from 1 to 0, the density level ranges from 0 to the maximum value of the density. When μ is proportional to Lebesgue measure, then GQ sets are also called *minimum volume sets*. We do not require μ to be uniform.

We make the following assumptions on the model in Eqn. (1).

[A] $\pi < 1$.

If the data are entirely from the contamination distribution μ, and we make no assumptions on the nominal component P, the problem is intractable. Generally, we expect $\pi \ll 1$.

[B] μ is absolutely continuous with respect to Lebesgue measure, and has the form

$$\mu(G) \propto \nu(G \cap G_0),$$

where ν is a *known* positive measure (not necessarily a probability measure), and G_0 is the unknown support of μ.

For example, ν could be Lebesgue measure, in which case μ is uniform, or ν could be a Gaussian measure, in which case μ is a truncated Gaussian. By the following assumption, which implies that the supports of μ and Q coincide, the set G_0 can be estimated consistently from the measurements using nonparametric support estimation.

[C] P is absolutely continuous with respect to μ (and therefore Lebesgue measure), and its density f with respect to μ has no plateaus: for all $\lambda > 0$, $\mu(\{x : f(x) = \lambda\}) = 0$.

Absolute continuity implies that $P(G_{P,\beta}) = \beta$ for all β, and that the mapping $\beta \mapsto \mu(G_{P,\beta})$ is continuous and nondecreasing. It is also implies concavity of the receiver operating characteristic (ROC) curves of optimal tests of μ against either P or Q. The no-plateau assumption implies uniqueness of P-GQ sets, for all $0 \leq \beta \leq 1$, up to μ-measure zero.

The assumption of a density implies that the support of P is contained in the support of μ. If the support of μ is known *a priori*, this condition can be relaxed.

[D] The density f of P is not bounded away from zero.

This means it is impossible to express P as a nontrivial mixture of μ and another distribution, and ensures that both π and P in Eqn. (1) are unique.

As stated previously, our goal is to produce an annotated ranking of the observations $X_1, \ldots, X_n$. We consider the task of ranking first. Define

$$\beta_P(X) := \sup\{\beta \in [0,1] : X_i \notin G_{P,\beta}\}. \tag{3}$$

and for $i = 1, \ldots, n$, set

$$\beta_i := \beta_P(X_i).$$

Essentially, each X_i is assigned to the smallest P-GQ set that contains X_i. Ordering the β_i as $\{\beta_{(n)}, \ldots, \beta_{(1)}\}$, from largest to smallest, naturally induces a ranking $\{X_{(1*)}, \ldots, X_{(n*)}\}$ of the

observations, where $(i*)$ denotes the index of the i-th most potentially anomalous observation.

Our choice of approach here may be motivated by considering the problem of formally testing the null hypothesis $H_0 : X_i \sim P$ versus the alternative hypothesis $H_1 : X_i \sim \mu$, for each $i = 1, \ldots, n$. If we choose to use a test of size α i.e., with a Type I error rate $\Pr(\text{Reject } H_0 | H_0 \text{ True}) = \alpha$, then the set $G^c_{P,1-\alpha}$ is in fact the rejection region for the most powerful test of this size. Then, if instead of making a hard decision of H_0 versus H_1, we report the corresponding statistical p-value under this class of tests, that p-value is simply $1-\beta_i$. Therefore, our proposed ranking follows the ordering of the observed p-values, from smallest to largest.

Now consider the issue of annotation of our ranked observations. The values β_i are themselves an obvious, and indeed not unreasonable, candidate for such an annotation. However, there is the need to interpret these values and, although the values β_i are well-defined probabilities in the context of the individual hypothesis tests for their corresponding observations X_i, they are not designed to be meaningfully interpreted *en masse* when simultaneously conducting multiple hypothesis tests. This observation is a variation on the issue at the heart of the well-known 'multiple testing problem' in statistics. Recall that, stated simply, the problem is that, whereas standard testing theory dictates that one should choose the size α of a single test to control the chance of an incorrectly rejected null hypothesis i.e., a 'false discovery', in contexts where a large number of such tests are to be conducted, one expects to end up with a correspondingly large number of false discoveries purely by chance. Such an outcome is often unsatisfactory, particularly when nontrivial amounts of energy are expected to be used to follow up on discoveries, as is often the case in, for example, anomaly detection problems.

The multiple testing problem has received a great deal of attention in the statistical literature over the past decade, since the seminal paper of Benjamini and Hochberg (1995). Their proposal for this problem effectively boils down to focusing attention not on the size α of individual tests, but rather the *rate* of false discoveries across tests. Since their paper, an entire sub-literature has evolved on the topic of FDR's, including a number of extensions in which analogues of the model in Eqn. (1) are assumed (e.g., Storey (2002, 2003); Genovese and Wasserman (2002); Efron et al. (2001)). From among these various contributions, we choose to adopt the so-called positive FDR

statistic of Storey (2002) as a natural one for our problem. The positive FDR (pFDR) is so named because it happens to be equal to the expected fraction of false discoveries, conditional on a positive number of discoveries having been made. In our context, this statistic can be written (Storey, 2003) as a probability

$$\mathrm{pFDR}(G) = \Pr\left(Y = 0 \,|\, X \notin G\right) \quad , \tag{4}$$

where G denotes an arbitrary set. This is just the probability that, given a 'discovery' is made i.e., H_0 is rejected due to X not being in G, that in fact this discovery is false.

Storey (2003) also proposes a corresponding analogue of the p-value, which he calls a q-value. This statistic, in our context, takes the form $\mathrm{pFDR}(G_{P,\beta_i})$. We therefore propose, as a more meaningful alternative to the values β_i, to annotate our ranked observations by the values

$$\gamma_i := 1 - \mathrm{pFDR}(G_{P,\beta_i}). \tag{5}$$

Two questions immediately arise in the context of this proposal. First, are the values of the γ_i's consistent with the ranking arising from the β_i's? And second, if so, since we do not know the actual values γ_i, how might they be estimated, given that they are formulated in terms of the unknown measure P? The first question may be answered in the affirmative under fairly general conditions, as summarized in the following result.

Proposition 1. *Let $C(s) = 1 - \mu(G_{P,1-s})$, for $s \in [0,1]$. Assume $C(s)$ is such that for all $s, s' \in [0,1]$, $s \geq s'$ implies $C(s)/s \leq C(s')/s'$. Then the ordered sequences $\{\beta_{(n)}, \ldots, \beta_{(1)}\}$ and $\{\gamma_{(n)}, \ldots, \gamma_{(1)}\}$ produce the same rank ordering $\{X_{(1*)}, \ldots, X_{(n*)}\}$ of the observations $X_1, \ldots, X_n$.*

The proof of this and all other such results in this paper may be found in the appendix. There we note that the function $C(s)$ is the receiver operating characteristic (ROC) curve for the optimal test of $X \sim P$ against $X \sim \mu$ (see Appendix). The assumption that $s \geq s'$ implies $C(s)/s \leq C(s')/s'$ says that the slope of the line connecting the origin to a point on the ROC is monotone decreasing as the point moves up the ROC. Equivalently, $(1 - \mu(G_{P,\beta}))/(1 - \beta)$ is monotone decreasing as β decreases. This assumption is satisfied when $C(s)$ is concave, which occurs, for example, under

assumption **[C]**. For another instance of a condition similar to the one assumed here, see Proposition 1 of Storey (2003).

Regarding the second question raised above, as to the estimation of the γ_i's, we address that in detail in the next section.

4 Estimation

4.1 A Fundamental Relation

The γ_i, and even the rankings as determined through the β_i, depend on P, which we assume unknown. Instead, all we have at our disposal are the observations $X_1, \ldots, X_n$, which are from the mixture distribution Q defined in (1), and our assumed knowledge about the contaminating distribution μ. In analogy to Eqn. (2), for $0 \leq \tilde{\beta} \leq 1$, define the GQ set under Q at level $0 \leq \tilde{\beta} \leq 1$ as

$$G_{Q,\tilde{\beta}} = \arg\min\{\mu(G) : Q(G) \geq \tilde{\beta}\} \ .$$

The following result is fundamental to the practical implementation of our proposed methodology, in that it relates the GQ sets under P to those under Q.

Proposition 2. *Assume conditions* **[A]**-**[C]** *hold. If* $0 \leq \beta \leq 1$ *and we define*

$$\tilde{\beta} \equiv \tilde{\beta}_{P,\beta} := \pi\mu(G_{P,\beta}) + (1-\pi)\beta \ ,$$

then $G_{Q,\tilde{\beta}} = G_{P,\beta}$. *Conversely, suppose* $0 \leq \tilde{\beta} \leq \tilde{\beta}_{\max} := \pi\mu(G_{P,1}) + 1 - \pi$. *Then* $G_{Q,\tilde{\beta}}$ *is unique, and if we define*

$$\beta \equiv \beta_{Q,\tilde{\beta}} := \frac{\tilde{\beta} - \pi\mu(G_{Q,\tilde{\beta}})}{1-\pi} \ , \tag{6}$$

then $G_{P,\beta} = G_{Q,\tilde{\beta}}$.

This result links P-GQ sets to Q-GQ sets in an explicit fashion. Every P-GQ set is a Q-GQ set, and every Q-GQ set with Q-mass $\leq \tilde{\beta}_{\max}$ is a P-GQ set. In particular, it implies that the ordering given by the P-GQ sets coincides with that of the Q-GQ sets. Intuitively, the smallest P-GQ set

containing X_i is also the smallest Q-GQ set containing X_i.

Note that $\tilde{\beta}_{\max}$ is the Q-mass of the support of P. Points X_i outside of $G_{P,1}$ all have $\beta_i = 1$, but they may have distinct values of $\tilde{\beta}_i$. In addition, $G_{Q,\tilde{\beta}}$ is not unique for $\tilde{\beta} > \tilde{\beta}_{\max}$. Nonetheless, the annotation γ_i in this case is still 1, as the following result shows. This result is the main result that facilitates the estimation of the annotations. Define

$$\tilde{\beta}_Q(X) := \sup\{\tilde{\beta} \in [0,1] : X \notin G_{Q,\tilde{\beta}}\}. \tag{7}$$

and $\tilde{\beta}_i := \tilde{\beta}_Q(X_i)$, in analogy to the β_i defined for P in Eqn. (3).

Corollary 1. *Under conditions* **[A]**-**[C]**,

$$\gamma_i = \frac{\pi(1 - \mu(G_{Q,\tilde{\beta}_i}))}{1 - \tilde{\beta}_i}. \tag{8}$$

Hence, we have the key insight that to estimate γ_i, we need only estimate $\mu(G_{Q,\tilde{\beta}_i})$, $Q(G_{Q,\tilde{\beta}_i})$, and π.

4.2 Estimating the Components of γ_i

Here we describe strategies for estimating each of the components of γ_i in (8). In this section, we use the notation $G(X) = G_{Q,\beta_Q(X)}$, the smallest Q-GQ set containing X. In addition, set $G_i = G(X_i) = G_{Q,\tilde{\beta}_i}$ and let $\hat{G}_i$ denote an estimate of G_i.

4.2.1 Estimation of $\mu(G_{Q,\tilde{\beta}_i})$

By assumption **[B]**, μ is known except for its support. By **[C]**, the support of μ is the support of Q. Thus, we may estimate $G_{Q,1}$, the support of μ, using standard methods for support estimation. With an estimate of the support, we can then properly normalize μ and compute an estimate $\hat{\mu}(G)$ for arbitrary G. In particular, we need to estimate the μ-measure of the sets G_i. Given $\hat{G}_i$, in our implementation, we estimate $\mu(\hat{G}_i)$ with a Monte Carlo approach based on simulation from μ.

4.2.2 Estimation of G_i

Suppose $\{\hat{G}_\lambda\}_{\lambda\in\Lambda}$ is a family of set estimates such that (a) each $\hat{G}_\lambda$ estimates some Q-GQ set, and (b) Λ is such that the range of GQ sets estimated is sufficiently rich to reasonably approximate $G_{Q,\tilde{\beta}}$ for any $0 \leq \tilde{\beta} \leq 1$. Then a natural estimator for G_i is $\hat{G}_i := \hat{G}_{\hat{\lambda}_i}$, where $\hat{\lambda}_i := \arg\max\{\mu(\hat{G}_\lambda) : \lambda \in \Lambda, X_i \notin \hat{G}_\lambda\}$.

In our experiments, we use thelevel sets

$$\hat{G}_\lambda = \{x : \frac{1}{n}\sum_{i=1}^{n} K_\sigma(x - X_i) \geq \lambda\}$$

of a kernel density estimate having a Gaussian kernel with bandwidth σ. The support of μ is estimated by taking the largest λ such that $\hat{G}_\lambda$ contains all the data. The bandwidth is selected by maximizing a cross-validation-based estimate of the area under the ROC curve $\tilde{C}(s)$, which is discussed in the appendix.

Other possibilities for $\hat{G}_\lambda$ include those based on the one-class support vector machine (OCSVM) with Gaussian kernel (Schölkopf et al., 2001). Path algorithms implementing this strategy are described in Lee and Scott (2007) and Lee and Scott (2010).

4.2.3 Estimation of $Q(G_{Q,\tilde{\beta}_i})$

Given estimates $\hat{G}_i$ of the sets $G_i = G_{Q,\tilde{\beta}_i}$, we may then estimate $\tilde{\beta}_i = Q(G_i)$ and related quantities through $\hat{Q}(\hat{G}_i)$, where $\hat{Q}(\cdot)$ is the empirical measure deriving from the data. This motivates the estimate $\hat{\tilde{\beta}}_i = ((i) - 1/2)/n$, $i = 1, \ldots, n$, where (i) here refers to the rank of the i-th observation under the ordering with respect to the sets $\hat{G}_i$.

4.2.4 Estimation of π

The estimation of π is facilitated by a transformation of variables. Specifically, consider $Z = \mu(G(X))$ as a univariate random variable on the interval $[0, 1]$ resulting from transformation of the generic random variable $X \sim Q$. The following result shows π to be related to the density of Z in a simple manner.

Proposition 3. *Assume conditions* **[B]**-**[D]** *hold. Define*

$$D(t) := \inf\{\beta : \mu(G_{P,\beta}) \leq t\}$$

and

$$\tilde{D}(t) := \inf\{\tilde{\beta} : \mu(G_{Q,\tilde{\beta}}) \leq t\} \ .$$

Then the density of Z *is* $\tilde{D}'(t) = \pi + (1-\pi)D'(t)$*, and* $\pi = \tilde{D}'(1-)$*.*

Thus $\tilde{D}(t)$ is the cumulative distribution function of Z. Proof of the proposition employs arguments that, similar to those of Proposition 1, rely on ROC curves of optimal tests, only in this case in a dual sense, with P and μ switched in their roles as null and alternative. The reader is referred to the appendix for details.

The obvious strategy now is to estimate π by estimating $\tilde{D}'(1-)$ based on the values $Z_i := \mu(G(X_i)) = \mu(G_i)$, $i = 1, \ldots, n$. Note, however, that we do not in fact have access to the Z_i, given our lack of knowledge of Q. We propose therefore to replace each Z_i by the value $\hat{Z}_i := \hat{\mu}(\hat{G}_i)$ once the estimates $\hat{G}_i$ are computed and to proceed accordingly.

Because of conditions **[B]**-**[C]**, $D(t)$ (and hence $\tilde{D}(t)$) is concave. Then estimating π amounts to estimating the value of a monotone decreasing density at the right boundary of its support. A consistent estimator for this problem has been studied by Kulikov and Lopuhaä (2006). Practical estimators have also been developed in recent work on multiple testing (Langaas et al., 2005; Storey, 2002) where they are used to estimate the proportion of true null hypotheses. There the p-values of a test play a role similar to our Z_i; under a null hypothesis, p-values are uniform, just as our Z_i's are uniform under $X \sim \mu$.

The estimated points $(\hat{Z}_i, \hat{\tilde{\beta}}_i)$ form an empirical version of $\tilde{D}(t)$. By an argument similar to that which established Proposition 1, the $\hat{\gamma}_i$ and $\hat{\tilde{\beta}}_i$ determine the same ranking provided the values $(1 - \hat{\tilde{\beta}}_i)/(1 - \hat{Z}_i)$ are nonincreasing as i increases. Because of estimation error, however, this will typically not be the case.

Our experiments involve continuous data, and therefore we expect $\tilde{D}(t)$ to be concave. To ensure that the $\hat{\gamma}_i$s and $\hat{\tilde{\beta}}_i$s produce the same rankings, we propose to smooth the empirical ROC

by fitting a function that is monotone, concave, and has endpoints at $(0,0)$ and $(1,1)$. Note that one may either regress $\hat{Z}_i$ on $\hat{\tilde{\beta}}_i$ or $1-\hat{\tilde{\beta}}_i$ on $1-\hat{Z}_i$. The latter approach, which we employ in our experiments, corresponds to smoothing an empirical version of $\tilde{C}(s) := 1-\mu(G_{Q,1-s})$ which is simply the reflection of $\tilde{D}(t)$ about the anti-diagonal of the unit square (see Appendix). In either case, the rankings are preserved.

ROC smoothing has two additional benefits. First, the slope of the estimated ROC at 1 gives an estimate of π as per Proposition 3. Second, the estimates $\hat{\gamma}_i$ satisfy $0 \leq \hat{\gamma}_i \leq 1$, which they should, being probabilities. Without smoothing, this might not be the case.

To implement ROC smoothing we employ a least squares linear smoothing spline with M fixed pieces, subject to monotonicity, concavity, and endpoint constraints. This is easily seen to be the solution of a quadratic program with M constraints. In our implementation we take $M = 20$ logarithmically spaced pieces and use Matlab's `quadprog` routine, which converges reliably for problems of this size.

5 Experiments

We apply our overall framework, *MN-SCAnn*, to a synthetic data problem, the network traffic data from Sec. 1, and flow cytometry data. Unless otherwise noted, μ is taken to be the uniform distribution. Full implementation details are available in our software, available as supplemental material on the JCGS website.

5.1 Synthetic data

Here the typical distribution P is a two dimensional, two component Gaussian mixture and the anomalies are uniform on a rectangle. The components of the mixture have equal weight, and the two components are

$$\mathcal{N}\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & 0.7 \\ 0.7 & 1 \end{bmatrix}\right) \quad \text{and} \quad \mathcal{N}\left(\begin{bmatrix} 0 \\ -2 \end{bmatrix}, \begin{bmatrix} 1 & -0.95 \\ -0.95 & 1 \end{bmatrix}\right)$$

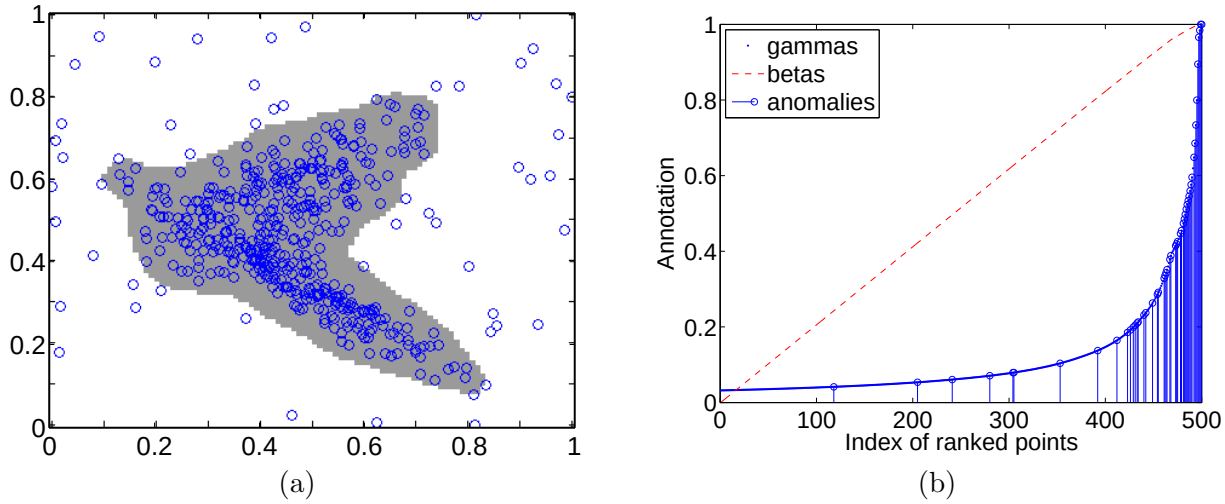
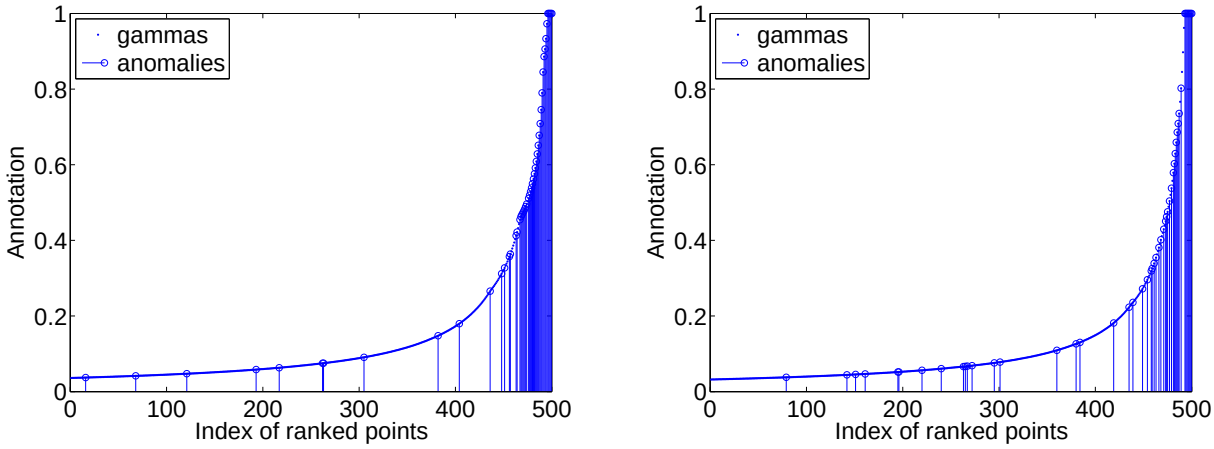


Figure 2: Results of Synthetic Data Experiment. (a) Sample of 500 points from synthetic data, together with the GQ set containing 90% of the data based on a thresholded kernel density estimate. (b) Sorted annotations $\hat{\gamma}_{(i)}$. Vertical stems indicate actual anomalies.

To facilitate processing, the data were translated and rescaled to the unit square. A sample of size 500 consisting of $\pi = 0.1$ anomalies is shown in Fig. 2 (a). Also shown is the Q-GQ set containing 90% of the data. Fig. 2 (b) plots the sorted annotation values $\hat{\gamma}_i$ for a realization of the synthetic data. The vertical stems indicate the approximately 50 observed anomalies. Anomalies constitute a strong majority of the data points with highest annotation values, with almost all having value $\hat{\gamma}_i \geq 0.1$. The presence of a few anomalies with lower annotations is expected given the overlap between the supports of μ and P.

We also investigated the impact of a non-uniform anomaly distribution and the presence of non-informative features.

To address non-uniform μ, we implemented an alternative measure that puts more probability mass toward the periphery of the space, and less near the centroid of the observed data. The details are available in the software. We reran the above Gaussian mixture experiment with this new anomaly distribution, and the resulting annotations are shown in Fig. 3 (a). The annotation values tend to rise more sharply toward the right end of the curve. That is, relative to the uniform anomaly distribution, more points are annotated as being more anomalous, as expected. The relative rankings, however, are fairly consistent: both choices of anomaly distribution lead to most



(a) (b)

Figure 3: Extensions of Gaussian mixture experiment. (a) Annotations based on a non-uniform anomaly distribution μ. (b) Annotations when five non-informative variables are appended to the data.

of the anomalous points being bunched toward the right side of the curve.

To address the issue of non-informative features, we appended five variables to the 2-d Gaussian mixture data, where the new variables are uniformly distributed, independent of whether the point is nominal or anomalous, and also independent of which Gaussian component the first two coordinates belong to (if the point is nominal). Thus the appended variables are completely non-informative. The results are shown in 3 (b). The annotation plot shows that most of the anomalies are still bunched toward the right of the plot, but not quite as much as when there were no non-informative variables. This indicates that MN-SCAnn possesses some tolerance to non-informative features. Results (not shown) from varying the number of non-informative features from 1 to 10 suggests the tolerance diminishes as this number increases.

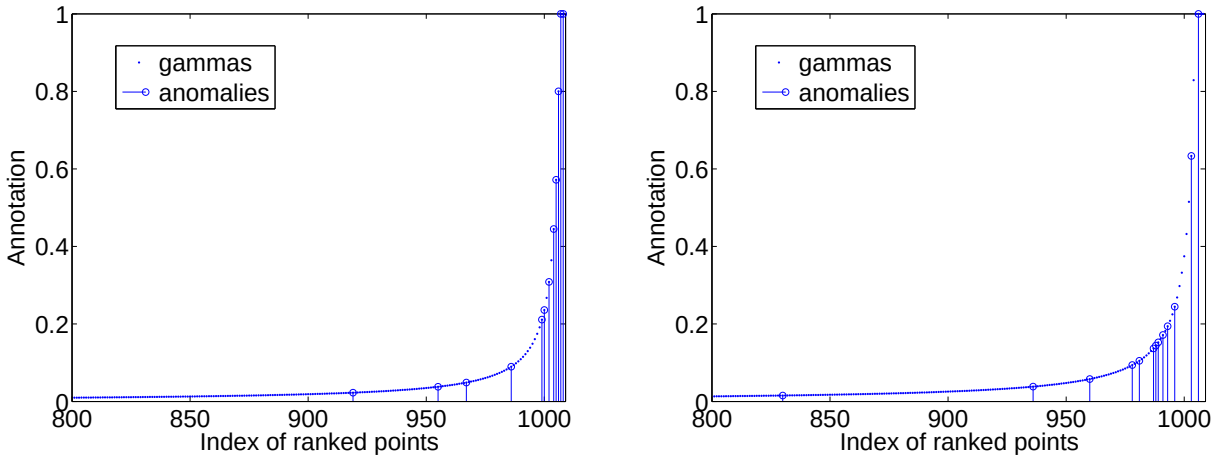
5.2 Network anomaly detection

Now return to the problem of detecting anomalous Internet traffic on a given network, described at the start of this paper. Fig. 4 shows a map of the Abilene network, the 'backbone' network serving most universities and research labs in the United States. Developed as part of the Internet2 project (www.internet2.org), a project devoted to development of the 'next-generation' Internet, Abilene

Figure 4: Schematic depiction of the Abilene network.

and Abilene data frequently serve as a testbed for development methodologies. Typically measurements on a network like Abilene are most easily available locally at network nodes (e.g., routers, regional aggregation points, etc.). So a natural way to approach the problem of anomaly detection is to seek to determine whether the traffic through a given network node is anomalous in nature or not. This problem is made challenging by many issues, particularly the facts that (a) traffic at a network node is a combination of the traffic from a number of incoming and outgoing links, and (b) traffic on fixed links has been found to have subtle combinations of various characteristics, and hence is not highly amenable to simple parametric modeling (e.g., Lakhina et al. (2004b,a)). Since traffic measurements are available on both inbound and outbound links, the dimensionality of the data X at each node ranges from 4 to 8.

Our methodology, which makes no assumptions on the distribution P of normal network traffic, is natural for this task. Here, we analyze a week's worth of data X_i, $i = 1, \ldots, 1008$, where each X_i is a vector of traffic volumes in a 10 minute window along all links connecting to the given router. (1008 is the number of 10 minute intervals in a week. Technically, the datapoints in this example are correlated, due to temporal correlations in the underlying traffic flows. However, we ignore these correlations here for the purpose of illustration.) Note that there is no pure 'truth set' available to us here. To evaluate the performance of our methodology, we compare it to that of the method due to Lakhina et al. (2004b,a), which is a standard in the literature. Their method



(a) (b)

Figure 5: Sorted annotations $\hat{\gamma}_{(i)}$ for the Abilene network traffic data at (a) the six dimensional Atlanta router, and (b) the eight dimensional Sunnyvale router. Vertical stems indicate anomalies detected by a method having access to global network traffic.

uses principle component regression to remove from the original multivariate network flow time series the main, common underlying trends. Importantly, it utilizes all of the data on all links in the network simultaneously, and therefore in that sense is a 'global' method, arguably working with more information than our 'local' method applied at each individual node. The data for this experiment are available online (www.internet2.org), and were preprocessed as described in Lakhina et al. (2004b).

Fig. 5 (b) shows the largest 208 (out of 1008 total data points, corresponding to all 10 minute intervals over a period of one week) annotations $\hat{\gamma}_i$ for the data at the six-dimensional Atlanta router. There are 12 anomalies total identified by the method of Lakhina et al., indicated by vertical stems, and we see that 8 out of these 12 occur past the 'knee'of the curve in Fig. 1(a), at roughly 0.1. Furthermore, all eight of these are in the top ten. These results strongly suggest that our methodology, despite using only information local to the router, is capable of mimicking well the performance of a global method that is standard in the literature. Similar comments apply to the eight-dimensional Sunnyvale router, shown in Fig. 5 (b). The conclusions here are further supported by a much more extensive analysis that has been done by the authors and colleagues (Chhabra et al., 2008).

Note that in both the Atlanta and Sunnyvale routers, there are also a few data points with annotations $\hat{\gamma}_i$ above 0.1 that were not declared anomalous by Lakhina et al. Since the method of Lakhina et al. is only a proxy to an unavailable truth set, it is impossible to say whether these points represent false discoveries or, alternatively, are in fact true discoveries missed by Lakhina et al.

5.3 Flow cytometry

A flow cytometry instrument is capable of measuring certain optical properties of biological cells, including size, granularity, and fluorescent tags associated to different antigens of interest (Huber and Hahne, 2005). Given a population of cells, flow cytometry data can be used to characterize the different cell types present. Fig. 6 (a) is a scatterplot showing two features, known as sideward light scatter and CD45, from a six dimensional dataset. There are multiple cell types present in the population, in this case four. Three of these cell types are associated with one of the three visible clusters, while a fourth is less apparent and overlaps the upper left cluster somewhat.

Unfortunately, cell populations are often contaminated by air bubbles, cell debris, and various other artifacts. These contaminants give rise to outliers in the flow cytometry data. It is desirable to identify these outliers and account for their prevalence so as to minimize their affect on subsequent processing. *MN-SCAnn* provides a natural way to assess the proportion of outliers present and to quantify the degree of outlyingness of individual cells.

We ran *MN-SCAnn* on this dataset as follows. The full dataset has around 30,000 cells, so for computational convenience we analyzed a random subsample of size 5,000. In addition, one of the features has an exceptionally high number of degenerate values (i.e., it takes on the max or min, giving the distribution an atomic component), so we eliminated this feature, and ran MN-SCAnn on the resulting five dimensional dataset. The results are shown in Fig. 6 (b). The estimated fraction of outliers is 0.001. The full sample in Fig. 6 (a) has about 30,000 points, and thus our method interprets about 30 of these to be outliers. For comparison, we also show the values $\hat{\beta}_{(i)}$, estimated by plugging in to Eqn. (6). Although the rank ordering is consistent with $\hat{\gamma}_{(i)}$, these values offer no information regarding the proportion and extremity of the contaminating points.

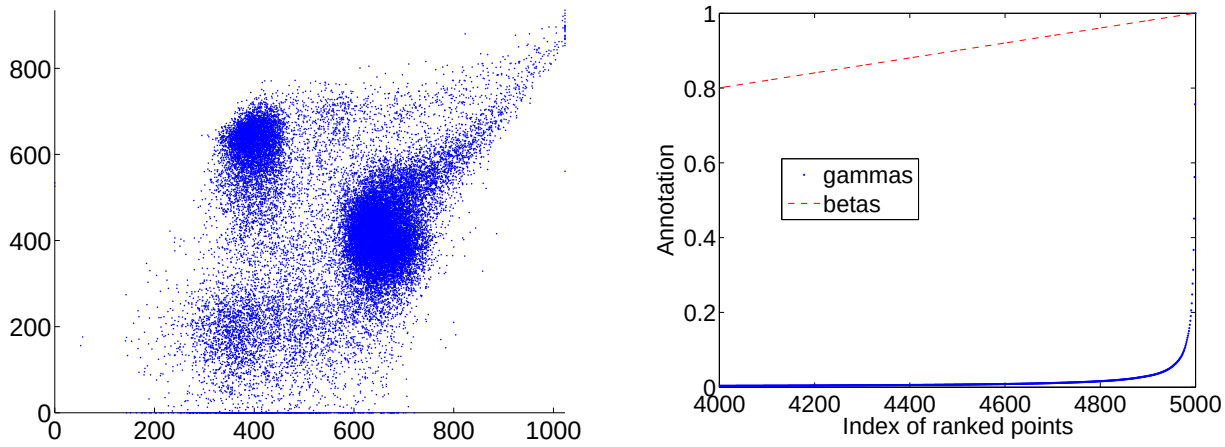


Figure 6: Flow Cytometry Data. (a) A two-dimensional projection of the five-dimensional data we analyze, showing its nonparametric nature. (b) Sorted annotation values $\hat{\gamma}_{(i)}$ for the 1000 most outlying points. Also shown are the values $\hat{\beta}_{(i)}$, which are conceptually similar to 'unadjusted p-values', and convey no information about the composition of the sample.

6 Discussion and Conclusions

A growing body of literature has begun to address the same issues that motivated *MN-SCAnn*, namely, rigorous statistical assessment of complex, multivariate data. We offer some connections to this work, and also discuss limitations of *MN-SCAnn*.

The notion of data depth has flourished recently as an approach to 'descriptive statistics, graphics, and inference' in multivariate, nonparametric settings (Liu et al., 1999; Serfling, 2006). The key ingredient in data depth analysis is a 'depth function' that defines an ordering of points in a multivariate sample with respect to degree of outlyingness. Much work on data depth focuses on estimating parameters such as generalized notions of location or kurtosis (see references in Liu et al. (1999); Serfling (2006)). Wang and Serfling (2006) discuss robust estimation of certain classes of depth functions, while Dang and Serfling (2010); Dang (2005) analyze swamping and masking breakdown points for data depth under contamination.

The so-called likelihood depth orders points with respect to the height of the density governing the sample. When the sample is contaminated, the points are ordered with respect to the density of the uncontaminated portion of the sample. Adopting the likelihood depth perspective, Fraiman and Meloche (1999) demonstrated the robustness of certain kernel density estimate based estimates

of centrality for symmetric distributions. We have shown in Proposition 1 that the ranking of *MN-SCAnn* with a uniform contamination measure coincides with that of likelihood depth. Some have argued that likelihood depth is not a valid measure of depth because multimodal densities complicate notions of center, among other reasons (Serfling, 2006). Yet the lack of a well-defined notion of center does not preclude the likelihood depth from imparting a valid ordering. The likelihood ordering is to us the most natural multivariate extension of 'extremeness' as captured by univariate p-values. Indeed, for multimodal nominal distributions, such as the flow cytometry data analyzed above, it is not clear whether it is even reasonable to insist on defining centrality. Moreover, our work demonstrates how likelihood depth can be extended beyond mere rankings to interpretable, quantitative annotations.

Other recent work has sought to combine traditional statistical approaches within a machine learning perspective. Roth (2006) incorporated a method for outlier rejection into a kernel Fisher discriminant approach to level set estimation (one-class classification). He takes advantage of the implicit Gaussian assumption in the kernel feature space to devise a quantile-quantile driven procedure for iteratively detecting and rejecting outliers.

One limitation of our approach is that it is unlikely to perform well when applied directly to high-dimensional data. For dimensions much greater that 10, our method would suffer from the limitations of kernel density estimation, and from the need to generate an extremely large random sample to estimate the anomaly distribution accurately. Therefore, to apply our method in such settings, it is recommended to first perform dimensionality reduction using some standard method, such as PCA or kernel PCA, to reduce the dimension to 10 or less.

On a final note, we point out that establishing the asymptotic behavior of our estimators $\hat{\gamma}_i$ of the quantities γ_i is an interesting open theoretical problem. For example, although the γ_i are essentially just one minus a version of Storey's q-value, and a form of consistency has been established for q-value estimates in Storey et al. (2004), the estimation strategy here is necessarily different, and so consistency would need to be shown independently. Also, given the structure of our problem and the nature of the estimators used, it would appear that arguments somewhat distinct from those in Storey et al. (2004) will be necessary.

Acknowledgment

The authors thank Mark Crovella and Parminder Chhabra for helpful discussions. They thank Parminder Chhabra for supplying the network traffic data and known anomalies, and William Finn and the UM Department of Pathology for the flow cytometry data. Clayton Scott was supported in part by NSF Award No. 0830490. Eric Kolaczyk was supported in part by NSF grant CCR-0325701 and ONR award N00014-06-1-0096.

References

Benjamini, Y. and Hochberg, Y. (1995), "Controlling the False Discovery Rate: a Practical and Powerful Approach to Multiple Testing," *J. R. Statist. Soc B*, 57, 289–300.

Chhabra, P., Scott, C., Kolaczyk, E., and Crovella, M. (2008), "Distributed Spatial Anomaly Detection," in *Proc. 27th IEEE Conf. on Computer Communications (IEEE INFOCOM)*, pp. 1705–1713.

Dang, X. (2005), "Nonparametric multivariate outlier detection methods, with applications," Ph.D. thesis, The University of Texas at Dallas.

Dang, X. and Serfling, R. (2010), "Nonparametric Depth-Based Multivariate Outlier Identifiers, and Masking Robustness Properties," *Journal of Statistical Planning and Inference*, 140, 198–213.

Efron, B., Tibshirani, R., Storey, J., and Tusher, V. (2001), "Empirical Bayes Analysis of a Microarray Experiment," *Journal of the American Statistical Association*, 96, 1151–1160.

El-Yaniv, R. and Nisenson, M. (2007), "Optimal Single-Class Classification Strategies," in *Adv. in Neural Inform. Proc. Systems 19*, eds. Schölkopf, B., Platt, J., and Hoffman, T., Cambridge, MA: MIT Press.

Fraiman, R. and Meloche, J. (1999), "Multivariate L-estimation," *Test*, 8, 255–317.

Genovese, C. and Wasserman, L. (2002), "Operating Characteristics and Extensions of the false discovery rate procedure," *Journal of the Royal Statistical Society, Series B*, 64, 499–517.

Hartigan, J. (1987), "Estimation of a Convex Density Contour in Two Dimensions," *J. Amer. Statist. Assoc.*, 82, 267–270.

Huber, W. and Hahne, F. (2005), "Cell-Based Assays," in *Bioinformatics and Computational Biology Solutions Using R and Bioconductor*, eds. Gentleman, R., Carey, V., Huber, W., Irizarry, R., and Dudoit, S., Springer.

Kulikov, V. and Lopuhaä, H. (2006), "The behavior of the NPMLE of a decreasing density near the boundaries of the support," *Ann. Stat.*, 34, 742–768.

Lakhina, A., Crovella, M., and Diot, C. (2004a), "Diagnosing Network-Wide Traffic Anomalies," in *Proc. ACM SIGCOMM*.

Lakhina, A., Papagiannaki, K., Crovella, M., Diot, C., Kolaczyk, E., and Taft, N. (2004b), "Structural Analysis of Network Traffic Flows," in *Proc. ACM SIGMETRICS/Performance*.

Lanckriet, G., Ghaoui, L. E., and Jordan, M. I. (2003), "Robust Novelty Detection with Single-Class MPM," in *Advances in Neural Information Processing Systems 15*, eds. S. Becker, S. T. and Obermayer, K., Cambridge, MA: MIT Press, pp. 905–912.

Langaas, M., Lindqvist, B. H., and Ferkingstad, E. (2005), "Estimating the proportion of true null hypotheses, with application to DNA microarray data," *Journal of the Royal Statistical Society, Series B*, 67, 555–572.

Lee, G. and Scott, C. (2007), "The one class support vector machine solution path," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing — ICASSP 2007*, Honolulu, USA, vol. 2, pp. II–521–II–524.

— (2010), "Nested support vector machines," *IEEE Trans. Signal Processing*, 58, to appear.

Liu, R., Parelius, J., and Singh, K. (1999), "Multivariate analysis by data depth: Descriptive statistics, graphics, and inference (with discussion)," *Ann. Stat.*, 27, 783–858.

Müller, D. and Sawitzki, G. (1992), "Excess mass estimates and tests of multmodality," *Journal of the American Statistical Association*, 86, 738–746.

Polonik, W. (1997), "Minimum Volume Sets and Generalized Quantile Processes," *Stochastic Processes and their Applications*, 69, 1–24.

Roth, V. (2006), "Kernel Fisher Discriminants for Outlier Detection," *Neural Computation*, 18, 942–960.

Sager, T. (1978), "Estimation of a multivariate mode." *Ann. Stat.*, 6, 802 – 812.

— (1979), "An interative method for estimating a multivariate mode and isopleth," *Journal of the American Statistical Association*, 74, 329–339.

Schölkopf, B., Platt, J., Shawe-Taylor, J., Smola, A., and Williamson, R. (2001), "Estimating the Support of a High-Dimensional Distribution," *Neural Computation*, 13, 1443–1472.

Scott, C. and Nowak, R. (2006), "Learning Minimum Volume Sets," *J. Machine Learning Res.*, 7, 665–704.

Serfling, R. (2006), "Depth functions in nonparametric multivariate inference," in *Data Depth: Robust Multivariate Analysis, Computational Geometry and Applications*, eds. Liu, R. Y., Serfling, R., and Souvaine, D. L., DIMACS Series in Discrete Mathematics and Theoretical Computer Science, American Mathematical Society, pp. 1–16.

Steinwart, I., Hush, D., and Scovel, C. (2005), "A Classification Framework for Anomaly Detection," *J. Machine Learning Research*, 6, 211–232.

Storey, J. (2002), "A Direct Approach to False Discovery Rates," *Journal of the Royal Statistical Society, Series B*, 64, 479–498.

— (2003), "The Positive False Discovery Rate: A Bayesian Interpretation of the q-value," *Annals of Statistics*, 31:6, 2013–2035.

Storey, J., Taylor, J., and Siegmund, D. (2004), "Strong Control, Conservative Point Estimation, and Simultaneous Conservative Consistency of False Discovery Rates: A Unified Approach," *Journal of the Royal Statistical Society, Series B*, 66, 187–205.

Theiler, J. and Cai, D. M. (2003), "Resampling Approach for Anomaly Detection in Multispectral Images," in *Proc. SPIE*, vol. 5093, pp. 230–240.

Walther, G. (1997), "Granulometric Smoothing," *Ann. Stat.*, 25, 2273–2299.

Wang, J. and Serfling, R. (2006), "Influence functions for a general class of depth-based quantile functions," *Journal of Multivariate Analysis*, 97, 810–826.