Nonparametric Assessment of Contamination in Multivariate Data Using Minimum Volume Sets and FDR

Clayton Scott* and Eric Kolaczyk†

Working Paper

April 26, 2007

Abstract

Large, multivariate datasets from high-throughput instrumentation have become ubiquitous throughout the sciences. Frequently, it is of great 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 minimum volume 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 minimum volume set methods with false discovery rate principles, in the context of contaminated data, is new. Moreover, 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 cell-based assays in bioinformatics, and the detection of anomalous traffic patterns in Internet measurement studies.

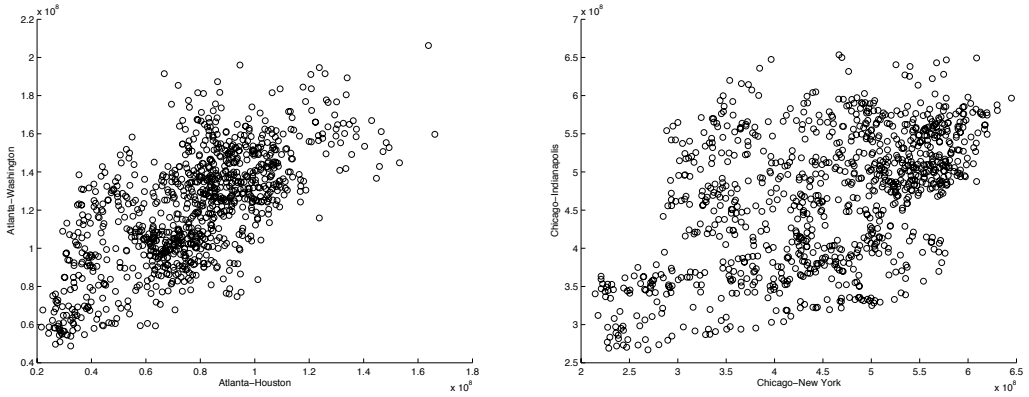
Keywords: Minimum volume sets, false discovery rate, nonparametric multivariate outlier detection, multiple level set estimation, monotone density estimation, ROC smoothing

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

*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-dot-edu



(a) (b)

Figure 1: Scatterplot of volume levels for traffic passing through select pairs of links at (a) Atlanta and (b) Chicago, in the Abilene network of Fig. 6 (a), over consecutive 10 minute intervals.

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 Section 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 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. For example, in the absence of any detailed information on the nature of the contamination, it is natural to assume that μ is simply a uniform distribution (e.g., [1, 2]). All numerical work herein was done with this assumption, although our overall framework and all stated analytical results hold for arbitrary μ. We also assume the supports of P and μ are bounded, with the former contained by the latter.

We then 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 minimum volume (MV) 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.

MV 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.,[3, 4]) and in the study of the modality of a distribution (e.g., [5, 6]). 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 [7] and references therein), and for the construction of classifiers (using uncontaminated observations from P) for predicting anomalies [1, 8, 9, 2, 10]. But our combination of MV 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., [11, 12, 13, 14, 15]), 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 MV sets and known μ.

The assumption of known μ can be justified on a number of grounds. In many situations, the assumption of uniform μ is intuitive and natural, and has recently been shown to optimize the worst-case detection rate among all choices for the unknown contamination distribution [16]. Furthermore, all of the above cited works that apply MV sets to anomaly prediction implicitly assume μ is a known distribution on anomalies [1, 8, 9, 2, 10]. That is, these predictors are only optimal when in fact the anomalies truly follow the volume-defining measure μ. Finally, ranking with respect to MV sets and uniform μ coincides with the ranking determined by the so-called likelihood data-depth [17, 7]. The connection to data depth is discussed further in the concluding section.

The rest of this paper is organized as follows. In Section 2 we introduce the basic elements of our proposed methodology. Estimation of the corresponding MV sets and FDR probabilities is addressed in Section 3, and details of our implementation are discussed in Section 4. Some numerical illustrations are presented in Section 5, using both synthetic data and data from two different real-world contexts – the pre-processing of cell-based assays in bioinformatics and the detection of anomalies in computer network traffic data (as described above). Finally, we close with a brief discussion in Section 6. All proofs of formal results stated in the text have been

collected in the appendix.

2 Methodology

Recall the model in Eqn. (1). Define $G_{P,\beta}$ to be the set with minimal μ-measure containing at least $\beta \in [0,1]$ probability mass under P i.e.,

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

This is the MV set under P, where volume is assessed with respect to μ. It is easily seen that MV sets coincide with density level sets of P, provided the density exists with respect to μ. 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. Although we refer to MV and density level sets interchangeably, we favor the former here because our method features the masses and volumes that MV sets make explicit.

As stated previously, our goal is to produce an annotated ranking of the observations. We consider the task of ranking first. For $i = 1, \ldots, n$, let

$$\beta_i := \inf_{0 \leq \beta \leq 1} \{\beta : X_i \in G_{P,\beta}\}. \tag{3}$$

That is, letting β vary freely between 0 and 1, we assign to X_i a set G_{P,β_i} that, among all MV sets, just barely includes 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 so-called 'multiple testing problem' in statistics. 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.

This problem has received a great deal of attention in the statistical literature over the past decade, since the seminal paper of Benjamini and Hochberg [11]. 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., [12, 13, 14, 15]). From among these various contributions, we choose to adopt the so-called positive FDR[1] statistic of Storey [12] as a natural one for our problem. In our context, this statistic is written as a probability

$$\mathrm{pFDR}(G) = \Pr\{X \sim P \,|\, X \notin G\} \quad , \tag{4}$$

where G denotes an arbitrary set and $X \sim Q$. 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 [12] 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 $X \sim \mu$. 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, when P is a continuous distribution and μ uniform. For another instance of a condition similar to the one assumed here, see Proposition 1 of [13].

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

3 Estimation

3.1 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

[1]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.

mixture distribution Q defined in (1), and our assumed knowledge of the contaminating distribution μ. In analogy to Eqn. (2), for $0 \leq \tilde{\beta} \leq 1$ define the MV 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 relating the MV sets under P to those under Q.

Proposition 2. *If* $0 \leq \beta \leq 1$ *and*

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

then $G_{Q,\tilde{\beta}} = G_{P,\beta}$. *Conversely, if* $0 \leq \tilde{\beta} \leq 1$ *and*

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

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

This result links Q-MV sets to P-MV sets in an explicit fashion. In particular, it implies that the ordering given by the P-MV sets coincides with that of the Q-MV sets. Intuitively, the smallest P-MV set containing X_i is also the smallest Q-MV set containing X_i. More formally, define

$$\tilde{\beta}_i \equiv \inf_{0 \leq \beta \leq 1} \{\tilde{\beta} : X_i \in G_{Q,\tilde{\beta}}\}. \tag{7}$$

in analogy to the β_i defined for P in Eqn. (3). Then $G_{P,\beta_i} = G_{Q,\tilde{\beta}_i}$ and $\beta_{(n)}, \ldots, \beta_{(1)}$ and $\tilde{\beta}_{(n)}, \ldots, \tilde{\beta}_{(1)}$ define the same rank ordering of the X_i's.

Furthermore, using Bayes' rule in conjunction with the expressions in (4) and (5), our proposed annotations γ_i may be expressed in the form

$$\begin{aligned} \gamma_i &= \pi\mu(G^c_{P,\beta_i})/Q(G^c_{P,\beta_i}) \\ &= \pi\mu(G^c_{Q,\tilde{\beta}_i})/Q(G^c_{Q,\tilde{\beta}_i}). \end{aligned} \tag{8}$$

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

3.2 Estimating the Components of γ_i

Here we describe general strategies for estimating each of the components of γ_i in (8). Specific strategies adopted in our implementation are offered in the next section.

3.2.1 Estimation of $G_{Q,\tilde{\beta}_i}$

For convenience, write $G_i = G_{Q,\tilde{\beta}_i}$, the smallest Q-MV set containing X_i. Suppose $\{\hat{G}_\lambda\}_{\lambda\in\Lambda}$ is a family of set estimates such that (a) each $\hat{G}_\lambda$ estimates some Q-MV set, and (b) Λ is such that the range of MV 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 $\hat{\lambda}_i := \arg\min\{\mu(\hat{G}_\lambda) : \lambda \in \Lambda, X_i \in \hat{G}_\lambda\}$.

We now briefly discuss two examples of such families $\{\hat{G}_\lambda\}_{\lambda\in\Lambda}$. The first requires solving the intermediate task of density estimation, while the second operates on the principle of direct set estimation.[2]

First, suppose a nonparametric estimate $\hat{f}(x)$ of the density f of Q is computed, such as a kernel density estimate. Then the sets $\hat{G}_\lambda = \{x : \hat{f}(x) \geq \lambda\}$ estimate the level sets of f, which coincide with the MV sets of Q. This approach has the advantage that the estimated sets are guaranteed to be nested. This implies that the smallest such set containing a given X_i can be computed rapidly via a bisection search on λ.

The second example is based on the one-class support vector machine (OCSVM) with Gaussian kernel [8]. Here $\hat{G}_\lambda$ is the OCSVM with regularization parameter λ. It has been shown [18] that for each λ, $\hat{G}_\lambda$ is a consistent estimator of the λ level set of Q. As λ varies through its range, all MV sets of Q are accounted for. For more on this approach, see [19], where the algorithm of Hastie et al. [20] is used to efficiently compute the entire family $\{\hat{G}_\lambda\}_{\lambda\in\Lambda}$.

Other methods for direct set estimation readily follow from classification algorithms having the ability to control the tradeoff between false positives and false negatives [21, 22]. If λ is a parameter that controls such a tradeoff, then $\hat{G}_\lambda$ may be identified with the classifier that discriminates $X_1, \ldots, X_n$ from an artificially generated sample from μ. [1, 2].

3.2.2 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.

3.2.3 Estimation of π

The estimation of π is facilitated by a transformation of variables. Specifically, define $Y_i = \mu(G_i)$, where recall that $G_i = G_{Q,\tilde{\beta}_i}$. Writing $G_i = G(X_i)$ now to emphasize the dependence on X_i, we consider $Y = \mu(G(X))$ 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 Y in a simple manner.

Proposition 3. *Let*

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

and

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

Assume $D(t)$ to be differentiable in t. Additionally, assume $D'(t) \to 0$ as $t \to 1$. Then the density of Y is $\tilde{D}'(t) = \pi + (1-\pi)D'(t)$, and therefore $\pi = \tilde{D}'(1-)$.

Thus $\tilde{D}(t)$ is the cumulative distribution function of Y. The assumption $D'(t) \to 0$ as $t \to 1$ holds provided it is not possible to write $P = (1-\theta)P_0 + \theta\mu$ for some distribution P_0 and for $\theta > 0$. Otherwise, P has a uniform component and it is impossible to resolve π accurately. Proof of the proposition employs arguments that, similar to those of Proposition 1, rely on ROC curves

[2]Note that estimating every level set of a density is equivalent to estimating the density itself, so there is no clear advantage of one approach over another.

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 $Y_1, \ldots, Y_n$. Note, however, that we do not in fact have access to the Y_i, given a lack of knowledge of Q. We propose therefore to estimate each Y_i by the value $\hat{Y}_i := \mu(\hat{G}_i)$ once the estimates $\hat{G}_i$ are computed and to proceed accordingly.

In the event that $D(t)$ (and hence $\tilde{D}(t)$) is concave in addition to being differentiable, 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 in [23]. Practical estimators have also been developed in recent work on multiple testing [24, 12] where they are used to estimate the proportion of true null hypothesis. There the p-values of a test play a role similar to our Y_i; under a null hypothesis, p-values are uniform, just as our Y_i's are uniform under $X \sim \mu$.

4 Implementation

In our experiments we estimate MV sets via the level 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. Estimation of a particular G_i involves a bisection search (up to a certain accuracy) on λ until X_i is just barely enclosed. Since each X_i is just barely contained in $\hat{G}_i$ (having X_i be on the boundary of $\hat{Q}(\hat{G}_i)$ would require infinite precision), we will estimate $\hat{\tilde{\beta}}_i = (i - 1/2)/n$, $i = 1, \ldots, n$. Note that there will be overlap in the iterations of the bisection search needed for different X_i which allows for significant computational savings. The issue of automatic bandwidth selection is addressed below.

4.1 Computation of $\mu(G_{Q,\tilde{\beta}_i})$

In practice, it is not necessary to specify the precise support of μ in advance. It suffices to specify μ on some bounding set (such as a hypercube or hypersphere) containing all of the data. After estimating the MV sets G_i, μ can be truncated to the region $\hat{G}_{(n)}$, which is an estimate of the support $G_{Q,1}$ of μ.

Our implementation produces $\hat{G}_i$ by thresholding a kernel density estimate, and these sets do not in general have simple forms.

Exact computation of $\mu(\hat{G}_i)$ is not possible in our implementation, even if μ is a uniform measure, because the sets $\hat{G}_i$, obtained by thresholding a kernel density estimate, are not sufficiently simple. Therefore we adopt a Monte Carlo approach based on simulation from μ. Sampling from the truncated μ is accomplished by simple rejection sampling from the original μ. In our experiments we take μ to be uniform and estimate μ using samples with size on the order of 10000, drawn uniformly from a hypercube containing the data.

4.2 Empirical ROC smoothing

The estimated points $(\hat{Y}_i, \hat{\tilde{\beta}}_i)$ 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

$(1-\hat{\tilde{\beta}}_i)/(1-\hat{Y}_i)$ 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 $\hat{Y}_i$ on $\hat{\tilde{\beta}}_i$ or $1-\hat{\tilde{\beta}}_i$ on $1-\hat{Y}_i$. The latter approach, which we employ in our experiments, corresponds to smoothing an empirical version of $\tilde{C}(s) := \inf\{1-\mu(G_{Q,\tilde{\beta}}) : 1-\tilde{\beta} \le s\}$, which is simply the reflection of $\tilde{D}(t)$ about the anti-diagonal of the unit square. 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 $0 \le \hat{\gamma}_i \le 1$, which they should being probabilities. Without smoothing, this might not be the case.

In our experiments we consider four methods of smoothing the empirical ROC under monotonicity, concavity, and endpoint constraints. Other options are possible, such as a natural cubic smoothing spline [25]. See [24] for a survey of results on the related problem of estimating the proportion of true null hypotheses in a traditional (p-value based) multiple testing problem. Recall that the empirical ROC is also an empirical CDF of the variable Y.

1. **Least Concave Majorant**: The derivative of the least concave majorant (LCM) of the empirical CDF was shown by Grenander [26] to be a nonparametric MLE of a nonincreasing density.

2. **Linear Spline**: The least squares linear smoothing spline with M fixed pieces, subject to monotonicity, concavity, and endpoint constraints, can be easily formulated as 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 this sized problem.

3. **Lomax Curve**: A specialized case of the Lomax family of distribution functions [27, 28] can be parametrized as
$$\tilde{D}(t) = \left[1 + \theta(t^{-\nu} - 1)\right]^{-\frac{1}{\nu}}.$$
Here $\theta > 0$ controls the area under the curve and $\nu > 0$ is a skewness parameter. We estimate these parameters by minimizing the total squared error using Matlab's `fminsearch` command for nonlinear optimization.

   Other common parametric ROC models, such as the binormal and bilogistic models, are not suitable for our purposes because their derivative at 1 is always 0.

4. **Convex Decreasing Density NPMLE**: The fourth estimate is obtained from a NPMLE for a convex decreasing density. In a recent empirical study [24], this is recommended as the most accurate among several estimators for the proportion of true null hypotheses in a multiple testing scenario. This finding does not necessarily have direct bearing on our problem, since the proportion of true nulls is typically large in a traditional multiple testing problem, while our π is typically very close to 0.

   The convexity assumption was motivated in [24] by the observation that the Grenander estimate tends to tail off toward zero at the right endpoint of its support, thus underestimating

the right endpoint. Equivalently, we note, convexity of the density estimate implies that the concavity of the estimated ROC cannot increase sharply near 1.

The resulting density estimate is a linear spline. We use the R implementation of [24], available as part of Bioconductor [29].

An alternative (which we did not explore) to enforcing a concave ROC would be to enforce the slope of the line connecting $(\tilde{D}(t), t)$ and $(1, 1)$ to be nonincreasing. The linear spline approach can be easily adapted to this setting.

4.3 Model selection

Nonparametric MV set estimators typically have tuning parameters that must be set. We adopt the *minimum integrated volume* (MIV) criterion proposed in [19]. Consider for example the bandwidth σ of a kernel density estimator, and let $\hat{G}_{Q,\tilde{\beta},\sigma}$ denote the estimate of $G_{Q,\tilde{\beta}}$ that results from a kernel density estimate with bandwidth σ. It is desirable that for each $\tilde{\beta}$, $\hat{G}_{Q,\tilde{\beta},\sigma}$ has as small a volume as possible while maintaining a mass near $\tilde{\beta}$. Therefore, assuming the actual mass of each set is near its estimated value, the quality of a given σ for estimating *all* MV sets may be assessed by the integrated volume

$$\mathrm{IV}(\sigma) = \int_0^1 \mu(\hat{G}_{Q,\tilde{\beta},\sigma})\, d\tilde{\beta},$$

with smaller values corresponding to better σ. Recalling that $\hat{\tilde{\beta}}_i = (i - 1/2)/n$, this quantity is estimated as

$$\frac{1}{n}\sum_{i=1}^{n} \mu(\hat{G}_{Q,\hat{\tilde{\beta}}_i,\sigma}) = \frac{1}{n}\sum_{i=1}^{n} \mu(\hat{G}_{i,\sigma}),$$

where the summand may be computed using the methods described above. As a side note, the integrated volume is one minus the area under the ROC curve $\tilde{D}(t)$. Therefore the MIV is a form of area under the ROC (AUC) criterion.

Finally, it is essential to note that during model selection, $\hat{Q}(\cdot)$ should not be the resubstitution estimate, otherwise overfitting will occur. Cross-validation or bootstrap estimates may be used instead. In our implementation we take the leave-one-out cross-validation estimate of the Q mass of the λ level set:

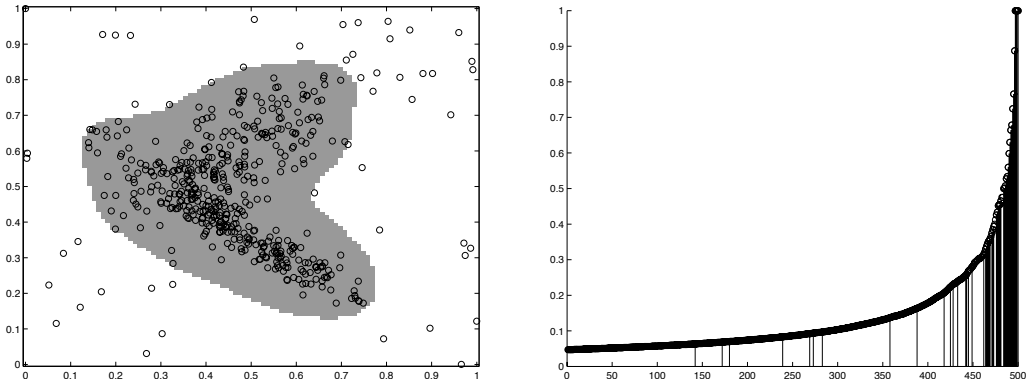
$$\hat{Q}(\hat{G}_\lambda) = \frac{|\{i : \frac{1}{n-1}\Sigma_{j\neq i} K_\sigma(X_j - X_i) \geq \lambda\}|}{n}.$$

5 Experiments

We apply *MN-SCAnn* to a synthetic data problem, the network traffic data from Section 1, and flow cytometry data.

5.1 Synthetic data

Here the typical distribution P is a two dimensional, two component Gaussian mixture and the anomalies are uniform on the unit square. A sample of size 500 consisting of $\pi = 10\%$ anomalies is shown in Fig. 2 (a). Also shown is the level set containing 90% of the data. Fig. 2 (b) plots the



(a) (b)

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

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 $\hat{\gamma}_i \geq 0.2$. The presence of a few anomalies with lower annotations is expected given the overlap between the supports of μ and P.

To assess the performance of the four methods for estimating π based on the four methods for ROC smoothing outlined in Section 4.2, we generated 100 samples of size $n = 200$ for each value of $\pi \in \{0.01, 0.05, 0.1, 0.15\}$. Figure 3 summarizes the resulting estimates of π by way of boxplots. There is no method that consistently outperforms the others. The linear spline method is more accurate for larger values of π, while the Lomax method is more accurate for smaller π.

Fig. 4 shows typical ROC fits of the four methods on the same data. The Lomax method, which is the only parametric method of the four, appears to be the least accurate, although its fit is still reasonable. Fig. 5 shows the estimated densities corresponding to the same ROC fits shown in Fig. 4.

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. 6 (a) 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 [30], a project devoted to development of the 'next-generation' Internet, Abilene 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, at a given point in time, 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., [31, 32]).

Our methodology, which makes no assumptions on the distribution P of normal network traffic,

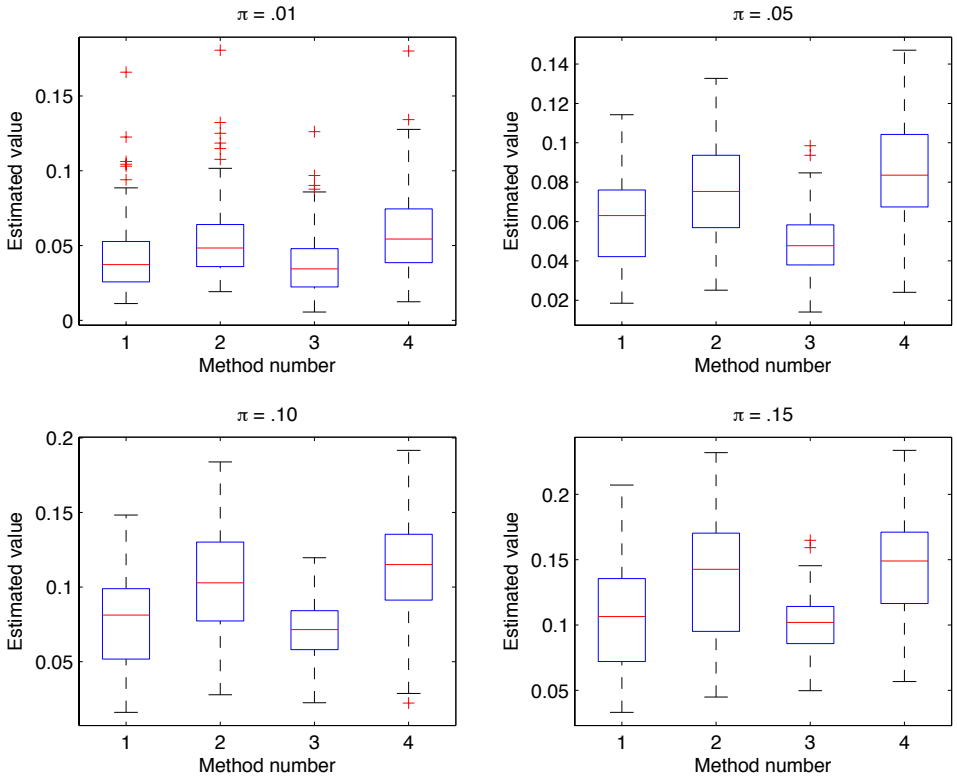


Figure 3: Boxplots summarizing four methods for estimating π, based on 100 instances of samples of size $n = 200$.

is natural for this task[3]. Fig. 6 (b) shows the largest 208 (out of 1008 total data points) annotations $\hat{\gamma}_i$ for the data at the Atlanta router shown in Fig. 1(a). The vertical stems are anomalies that were detected using a global method (having access to all data on all links in the network) and serve as ground truth anomalies for our purposes [31, 32]. There are 11 anomalies total. We see that 8 out of 11 occur past the 'knee' of the curve at roughly 0.2, and three are in the top six.

5.3 Flow cytometry

A flow cytometry instrument is capable of measuring certain optical properties of biological cells, including size, granularity, and various fluorescence properties [33]. Given a population of cells, flow cytometry data can be used to characterize the different cell types present. Fig. 7 is a scatterplot derived from two particular optical features known as sideward light scatter and CD45. 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. This dataset very clearly illustrates the need for nonparametric methods of contamination assessment.

Unfortunately, cell populations are often contaminated by air bubbles, cell debris, and various

[3]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.

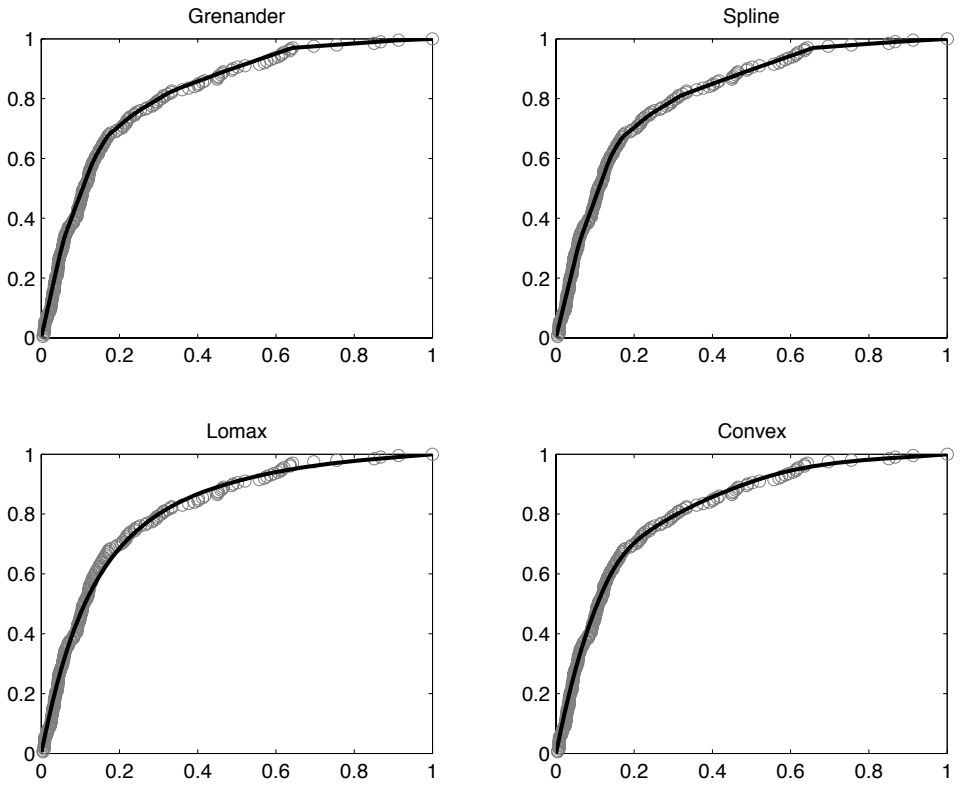


Figure 4: Typical ROC fits by the four methods of ROC smoothing.

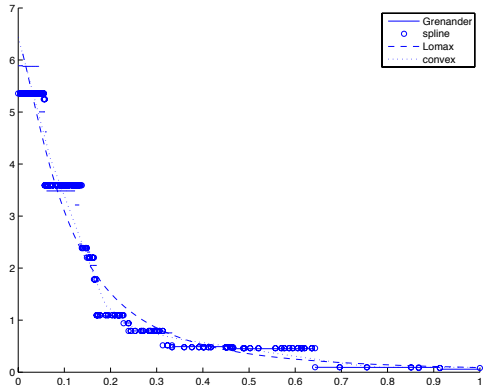


Figure 5: Estimates of the density of Y corresponding to the four ROC fits shown in Fig. 4.

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 points.

We ran *MN-SCAnn* on a subsample of size[4] 5,000 with results shown in Fig 8. The estimated

[4]Using the full sample caused memory problems in our implementation, which is not optimized for memory

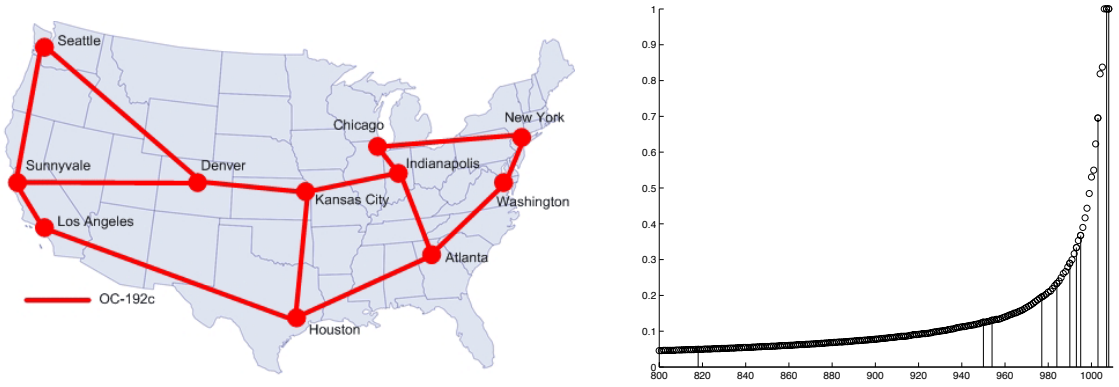


Figure 6: (a) Schematic depiction of the Abilene network (b) Sorted annotations $\hat{\gamma}_{(i)}$ for the network traffic data. Vertical stems indicate anomalies detected by a method having access to global network traffic.

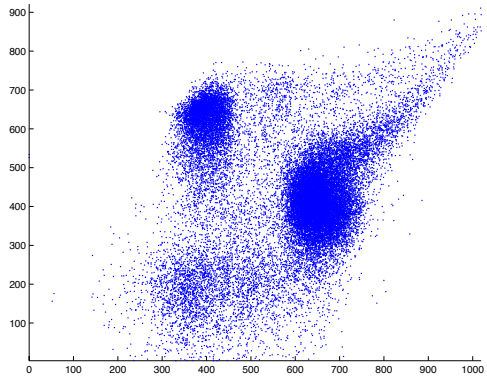


Figure 7: Full sample of approximately 30,000 data points from a flow cytometry experiment.

fraction of outliers is 0.01. The full sample in Fig 7 has about 30,000 points, and thus our method interprets about 300 of these to be outliers. Although no ground truth is available for this data, the contours and estimated proportion of outliers are consistent with our intuition based on visual inspection. 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.

6 Discussion

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 now offer some connections to and place *MN-SCAnn* in the context of this work.

The notion of data depth has flourished recently as an approach to "descriptive statistics,

efficiency.

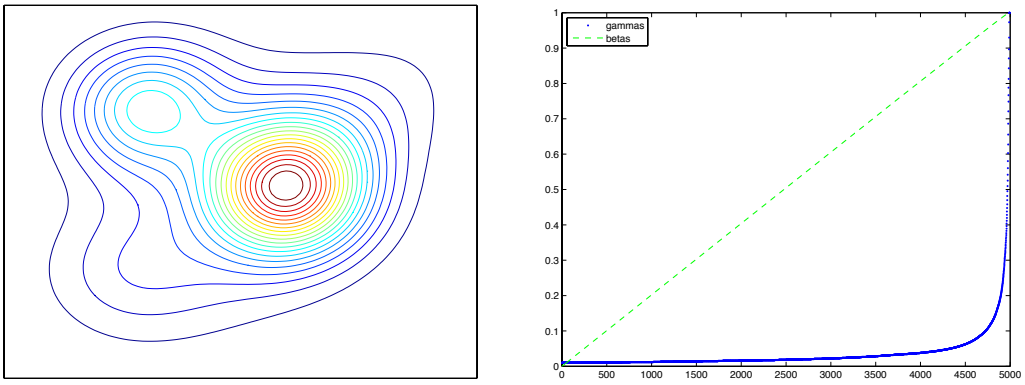


Figure 8: Results for flow cytometry data based on a subsample of size 5,000. (a) Contours of density of Q, reflecting the ranking induced by the annotations $\hat{\beta}_{(i)}$. (b) Sorted annotation values $\hat{\gamma}_{(i)}$. Also shown are the values $\hat{\beta}_{(i)}$, which are conceptually similar to "unadjusted p-values" and clearly convey no information about the composition of the sample.

graphics, and inference" in multivariate, nonparametric settings [7, 34]. 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 [7, 34]). Wang and Serfling [35] discuss robust estimation of certain classes of depth functions, while Dang and Serfling [36, 37] 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 demonstrated the robustness of certain kernel density estimate based estimates of centrality for symmetric distributions [17]. We have shown in Proposition 1 that the ranking of *MN-SCAnn* with a uniform reference 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 [34]. 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 incorporated a method for outlier rejection into a kernel Fisher discriminant approach to level set estimation (one-class classification) [38]. 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.

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 [39], 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 [39] 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. Eric Kolaczyk was supported in part by NSF grant CCR-0325701 and ONR award N00014-06-1-0096.

Appendix

The proofs of Propositions 1 and 3 rely on certain ROCs (or CDFs) which we discuss here in more detail.

Consider the optimal test for the null hypothesis $X \sim P$ against the alternative $X \sim \mu$. By definition of $G_{P,\beta}$, the critical region $G^c_{P,\beta}$ is the most powerful test of size $P(G^c_{P,\beta}) = 1 - P(G_{P,\beta}) = 1 - \beta$, with power equal to $\mu(G^c_{P,\beta}) = 1 - \mu(G_{P,\beta})$. Thus, $\{(1-\beta, 1-\mu(G_{P,\beta})) : 0 \leq \beta \leq 1\}$ traces out the ROC of the optimal test. In functional form, the ROC is given by

$$C(s) := 1 - \mu(G_{P,1-s}).$$

In a similar way, we can associate

$$\tilde{C}(s) = \inf\{1 - \mu(G_{Q,\tilde{\beta}}) : 1 - \tilde{\beta} \leq s\}$$

with the optimal test for $X \sim Q$ versus $X \sim \mu$.

The estimation of π is facilitated by consideration of what might be called the *dual* ROC to the *primal* ROC discussed in Section 2. In particular, we now view μ as the null distribution and P as the alternative. While this is the opposite of the scenario considered throughout the paper, it will be a useful analytical device. By definition of $G_{P,\beta}$, the critical region $G_{P,\beta}$ gives the most powerful test of size $\mu(G_{P,\beta})$ with power equal to $P(G_{P,\beta}) = \beta$. Thus, $\{(\mu(G_{P,\beta}), \beta) : 0 \leq \beta \leq 1\}$ traces out the ROC of the optimal test. In functional form, the ROC is given by

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

Note that the dual ROC can be obtained by reflecting the "primal" ROC $C(s)$ about the anti-diagonal of the unit square.

Similarly, the dual ROC corresponding to the optimal test of the null $X \sim \mu$ versus the alternative $X \sim Q$ (again, this test is viewed as purely an analytical device) is given by

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

and is traced out by the curve $\{(\mu(G_{Q,\tilde{\beta}}), \tilde{\beta}) : 0 \leq \tilde{\beta} \leq 1\}$. As before, this curve may be obtained by reflecting $\tilde{C}(s)$ about the anti-diagonal of the unit square.

Proof of Proposition 1.

Consider testing $H_0 : X \sim P$ versus $H_1 : X \sim \mu$. By definition of $G_{P,\beta}$, the critical region $G^c_{P,\beta}$ is the most powerful test of size $P(G^c_{P,\beta}) = 1 - P(G_{P,\beta}) = 1 - \beta$, with power equal to $\mu(G^c_{P,\beta}) = 1 - \mu(G_{P,\beta})$. Thus, $\{(1-\beta, 1-\mu(G_{P,\beta})) : 0 \le \beta \le 1\}$ traces out the receiver operating characteristic (ROC) curve of the optimal test, which is given by $C(s) := 1 - \mu(G_{P,1-s})$.

Now for any pair of indices i, i', we wish to show $\beta_i \le \beta_{i'}$ iff $\gamma_i \le \gamma_{i'}$. Note that

$$\gamma_i = 1 - \mathrm{pFDR}(G_{P,\beta_i}) = \frac{\pi\mu(G^c_{P,\beta_i})}{Q(G^c_{P,\beta_i})} = \left[1 + \frac{1-\pi}{\pi}\,\frac{P(G^c_{P,\beta_i})}{\mu(G^c_{P,\beta_i})}\right]^{-1} . \tag{9}$$

So $\gamma_i \le \gamma_{i'}$ iff $(1-\mu(G_{P,\beta_{i'}}))/(1-P(G_{P,\beta_{i'}})) \ge (1-\mu(G_{P,\beta_i}))/(1-P(G_{P,\beta_i}))$. But these ratios of μ to P are simply the slopes of lines through the origin to the points on the $\beta(t)$ ROC curve corresponding to $\beta_{i'}$ and β_i, and these slopes are decreasing in decreasing β, by assumption. □

Proof of Proposition 2:

To establish the first statement, that $G_{P,\beta}$ is the Q-MV set at level $\tilde{\beta}$, we must establish (a) $Q(G_{P,\beta}) \ge \tilde{\beta}$ and (b) if $Q(G) \ge \tilde{\beta}$, then $\mu(G) \ge \mu(G_{P,\beta})$. To establish (a), observe

$$\begin{aligned} Q(G_{P,\beta}) &= \pi\mu(G_{P,\beta}) + (1-\pi)P(G_{P,\beta}) \\ &= \pi\mu(G_{P,\beta}) + (1-\pi)\beta \\ &= \tilde{\beta}. \end{aligned}$$

To establish (b), assume it does not hold. That is, assume there exists G such that $Q(G) \ge \tilde{\beta}$ and $\mu(G) < \mu(G_{P,\beta})$. Then

$$\begin{aligned} P(G) &= \frac{Q(G) - \pi\mu(G)}{1-\pi} \\ &\ge \frac{\tilde{\beta} - \pi\mu(G)}{1-\pi} \\ &\ge \frac{\tilde{\beta} - \pi\mu(G_{P,\beta})}{1-\pi} \\ &= \beta, \end{aligned}$$

which contradicts the definition of $G_{P,\beta}$ as the P-MV set at level β.

The second statement follows in a similar manner. □

Proof of Proposition 3:

$$\begin{aligned} \Pr\{Y \le t | X \sim Q\} &= \Pr\{\mu(G(X)) \le t | X \sim Q\} \\ &= Q(\mu(G(X)) \le t) \\ &= Q(X \in G_{Q,\tilde{D}(t)}) \\ &= \tilde{D}(t). \end{aligned}$$

Thus $\tilde{D}(t) = \pi\Pr\{Y \le t | X \sim \mu\} + (1-\pi)\Pr\{Y \le t | X \sim P\}$. Now

$$\begin{aligned} \Pr\{Y \le t | X \sim \mu\} &= \Pr\{\mu(G(X)) \le t | X \sim \mu\} \\ &= \mu(\mu(G(X)) \le t) \\ &= \mu(X \in G_{Q,\tilde{D}(t)}) \\ &= t. \end{aligned}$$

Similarly,

$$\begin{aligned}\Pr\{Y \leq t | X \sim P\} &= \Pr\{\mu(G(X)) \leq t | X \sim P\} \\ &= P(\mu(G(X)) \leq t) \\ &= P(X \in G_{Q,\tilde{D}(t)}) \\ &= P(X \in G_{P,D(t)}) \\ &= D(t).\end{aligned}$$

The result follows by differentiating $\tilde{D}(t)$. $\square$

References

[1] J. Theiler and D. M. Cai, "Resampling approach for anomaly detection in multispectral images," in *Proc. SPIE*, vol. 5093, 2003, pp. 230–240.

[2] I. Steinwart, D. Hush, and C. Scovel, "A classification framework for anomaly detection," *J. Machine Learning Research*, vol. 6, pp. 211–232, 2005.

[3] T. Sager, "Estimation of a multivariate mode." *Ann. Statist.*, vol. 6, pp. 802 – 812, 1978.

[4] ——, "An interative method for estimating a multivariate mode and isopleth," *Journal of the American Statistical Association*, vol. 74, pp. 329–339, 1979.

[5] J. Hartigan, "Estimation of a convex density contour in two dimensions," *J. Amer. Statist. Assoc.*, vol. 82, no. 397, pp. 267–270, 1987.

[6] D. Müller and G. Sawitzki, "Excess mass estimates and tests of multmodality," *Journal of the American Statistical Association*, vol. 86, pp. 738–746, 1992.

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

[8] B. Schölkopf, J. Platt, J. Shawe-Taylor, A. Smola, and R. Williamson, "Estimating the support of a high-dimensional distribution," *Neural Computation*, vol. 13, no. 7, pp. 1443–1472, 2001.

[9] G. R. G. Lanckriet, L. E. Ghaoui, and M. I. Jordan, "Robust novelty detection with single-class mpm," in *Advances in Neural Information Processing Systems 15*, S. T. S. Becker and K. Obermayer, Eds. Cambridge, MA: MIT Press, 2003, pp. 905–912.

[10] C. Scott and R. Nowak, "Learning minimum volume sets," *J. Machine Learning Res.*, vol. 7, pp. 665–704, 2006.

[11] Y. Benjamini and Y. Hochberg, "Controlling the false discovery rate: a practical and powerful approach to multiple testing," *J. R. Statist. Soc B*, vol. 57, no. 1, pp. 289–300, 1995.

[12] J. Storey, "A direct approach to false discovery rates," *Journal of the Royal Statistical Society, Series B*, vol. 64, pp. 479–498, 2002.

[13] ——, "The positive false discovery rate: A Bayesian interpretation of the q-value," *Annals of Statistics*, vol. 31:6, pp. 2013–2035, 2003.

[14] C. Genovese and L. Wasserman, "Operating characteristics and extensions of the false discovery rate procedure," *Journal of the Royal Statistical Society, Series B*, vol. 64, pp. 499–517, 2002.

[15] B. Efron, R. Tibshirani, J. Storey, and V. Tusher, "Empirical bayes analysis of a microarray experiment," *Journal of the American Statistical Association*, vol. 96, pp. 1151–1160, 2001.

[16] R. El-Yaniv and M. Nisenson, "Optimal single-class classification strategies," in *Advances in Neural Information Processing Systems 19*, B. Schölkopf, J. Platt, and T. Hoffman, Eds. Cambridge, MA: MIT Press, 2007.

[17] R. Fraiman and J. Meloche, "Multivariate L-estimation," *Test*, vol. 8, no. 2, pp. 255–317, 1999.

[18] R. Vert and J.-P. Vert, "Consistency and convergence rates of one-class SVM and related algorithms," *J. Machine Learning Research*, pp. 817–854, 2006.

[19] G. Lee and C. Scott, "The one class support vector machine solution path," *Proc. Int. Conf. Acoust. Speech. Sig. Proc.*, 2007.

[20] T. Hastie, S. Rosset, R. Tibshirani, and J. Zhu, "The entire regularization path for the support vector machine," *J. Machine Learning Research*, pp. 1391–1415, 2004.

[21] C. Elkan, "The foundations of cost-sensitive learning," in *Proceedings of the 17th International Joint Conference on Artificial Intelligence*, Seattle, Washington, USA, 2001, pp. 973–978.

[22] F. R. Bach, D. Heckerman, and E. Horvitz, "Considering cost asymmetry in learning classifiers," *J. Machine Learning Research*, pp. 1713–1741, 2006.

[23] V. Kulikov and H. Lopuhaä, "The behavior of the NPMLE of a decreasing density near the boundaries of the support," *Ann. Stat.*, vol. 2, 2006, in press.

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

[25] B. Turlach, "Shape constrained smoothing using smoothing splines," *Computational Statistics*, vol. 20, pp. 81–103, 2005.

[26] U. Grenander, "On the theory of mortality measurement: part ii," *Skand. Akt.*, vol. 39, pp. 125–153, 1956.

[27] G. Campbell and M. V. Ratnaparkhi, "An application of Lomax distributions in receiver operating characteristic (ROC) curve analysis," *Commun. Statist. – Theory Meth.*, vol. 22, pp. 1681–1697, 1993.

[28] C. Lloyd, "Regression models for convex ROC curves," *Biometrics*, vol. 56, pp. 862–867, 2000.

[29] "http://www.bioconductor.org."

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

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

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

[33] W. Huber and F. Hahne, "Cell-based assays," in *Bioinformatics and Computational Biology Solutions Using R and Bioconductor*, R. Gentleman, V. Carey, W. Huber, R. Irizarry, and S. Dudoit, Eds. Springer, 2005.

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

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

[36] X. Dang and R. Serfling, "Nonparametric depth-based multivariate outlier identifiers, and robustness properties," Preprint, 2006.

[37] X. Dang, "Nonparametric multivariate outlier detection methods, with applications," Ph.D. dissertation, The University of Texas at Dallas, November 2005.

[38] V. Roth, "Kernel Fisher discriminants for outlier detection," *Neural Computation*, vol. 18, pp. 942–960, 2006.

[39] J. Storey, J. Taylor, and D. Siegmund, "Strong control, conservative point estimation, and simultaneous conservative consistency of false discovery rates: A unified approach," *Journal of the Royal Statistical Society, Series B*, vol. 66, pp. 187–205, 2004.