MAKING DECISIONS WITH SHUFFLED BITS

Stefano Marano

DIEM, University of Salerno,
Fisciano (SA), Italy
marano@unisa.it

*Peter Willett**

ECE Dept., University of Connecticut,
Storrs, CT, USA
peter.willett@uconn.edu

ABSTRACT

Unlabeled detection is an emerging paradigm for modern decentralized decision systems faced with big-data applications, and for all those applications in which data must be fused without exploiting their identity, due to the lack of provenance labels, or to uncontrolled data shuffling. Our focus here is on binary alphabets, and we ask: If our data have been shuffled in an unknown way, can a reliable decision about the underlying state of nature be made? Should the decision be made after an attempt to estimate the lost labels? And do there exist easily implementable decision rules? In answering these questions, we gain much insight: We show that two greedy algorithms previously introduced in the literature are equivalent to the GLRT, whose performance can be quite poor, and the detector known as ULR is equivalent to a simple counting rule. A new detector based on the central limit theorem is simply implementable and offers close-to-optimal performance in many scenarios of practical interest.

Index Terms— Unlabeled Detection, Shuffled Data, Unknown Permutation, Big Data, Large Sensor Networks.

1. INTRODUCTION

Let us begin with a motivating example. Suppose there are two boxes, $\mathcal{H}_1$ and $\mathcal{H}_0$, each containing n coins. According to $\mathcal{H}_1$, half the coins are unfairly weighted to show a "head" with probability 0.9 and the other half conversely show a head with probability 0.1. Under $\mathcal{H}_0$ all coins are fair. A box is selected and the coins inside that are tossed: the total number of heads, say it $k_{\mathbf{x}}$, is revealed. Can one infer which box was selected? In both cases the expected number of heads is $n/2$, so a threshold test on $k_{\mathbf{x}}$ is of little use. On the other hand, less variability in $k_{\mathbf{x}}$ would be expected with $\mathcal{H}_1$, and therefore with a small value of $|k_{\mathbf{x}} - n/2|$ seems reasonable to choose $\mathcal{H}_1$.

A different (generalized likelihood ratio test, GLRT) approach to making the decision is to consider the specific sequence of heads and tails. With box $\mathcal{H}_1$, the (maximum likelihood, ML) assignment between observed heads and the coins they came from can be made with few errors, yielding $0.9^{n-|k_{\mathbf{x}}-n/2|}\, 0.1^{|k_{\mathbf{x}}-n/2|}$ as the likelihood. With box $\mathcal{H}_0$, all assignments are equally likely but anyway irrelevant: the likelihood is 0.5^n in all cases.

The heads and tails in our boxes exemplify unlabeled measurements: the observer has access to the measurements, but has no idea which coin (fair, biased head-light or head-heavy) produced each one. Three approaches are discussed: ignore labels and make a threshold test using $k_{\mathbf{x}}$ (useless, in the example); test if $k_{\mathbf{x}}$ is close to $n/2$ (this is based on CLT, as we shall show), and the GLRT (assign the labels). In the case just given, the GLRT is the same as with CLT (central limit theorem). However, suppose instead that the probabilities of head for the coins in box $\mathcal{H}_1$ are 0.99 and 0.1. Does the aforementioned ML assignment procedure (the GLRT) still work? It turns out that the GLRT performs poorly; and in fact $k_{\mathbf{x}}$ itself seems a (more?) sensible decision statistic.

Admittedly, in these examples where there are at most two different types of coins in each box, making the optimal decision is easy, because the statistical distribution of the observed $k_{\mathbf{x}}$ is available in simple form. But with many different types of coins this is no longer true. In principle, the statistical characterization of $k_{\mathbf{x}}$ is still known, but is not tractable and no closed-form expression for the optimal decision statistic is available. The following questions arise: What are reasonable decision criteria? In what cases simply comparing $k_{\mathbf{x}}$ to a threshold level represents an acceptable criterion? Is it always reasonable to build the decision statistic by first inferring what is the unseen head/tail sequence?

1.1. Relation to Prior Work

The stated problem falls under the wide umbrella of *signal processing with unlabeled data*, an emerging paradigm with applications, among other fields, in large sensor network faced with big-data analysis. Signal processing with unlabeled data refers to signal processing tools specifically designed when the data under analysis lack of a time/space reference and can be seen as subject to an unknown permutation before becoming available for processing. Studies in this field have been pioneered by [1, 2], with roots that can be traced back to [3]. Recent contributions include [4–8]. Closer to this article, because focusing on inference problems, are the works in [9–12]. The theoretical contribution of [12] is to

*P. Willett was supported by AFOSR under contract FA9500-18-1-0463.

derive the asymptotically optimal performance for detection by unlabeled data, in the limit $n \to \infty$. The authors of [12] also propose sub-optimal detection algorithms for finite n.

Our contribution is to explore the detection problem with unlabeled data in the special case of binary alphabets. We find closed-form expressions for some of the detectors proposed in [12], providing new insights. Two greedy algorithms are shown to be equivalent to the GLRT (generalized likelihood ratio test), and a closed-form expression for this latter is offered. A low-complexity detector termed ULR (unlabeled likelihood ratio) has been proposed in [12], and here we find that the ULR decision is simply based on comparing $k_{\mathbf{x}}$ to a threshold. Finally, we develop an approximation based on CLT, which is shown to achieve close-to-optimal performance in many cases of practical interest.

2. PROBLEM FORMULATION

Let $\mathcal{X} = \{0, 1\}$ be the alphabet of the observations, and let us start from the case in which the detector observes the *labeled* vector $\mathbf{X}^n = (X_1, \ldots, X_n)$. We use capital letters for random quantities, such as $\mathbf{X}^n$, and the corresponding lowercase $\mathbf{x}^n$ for their realizations. Let $p_i = \mathbb{P}_1(X_i = 1)$ and $q_i = \mathbb{P}_0(X_i = 1)$, where $\mathbb{P}_j$ is the probability operator under hypothesis $\mathcal{H}_j$, $j = 0, 1$. We assume $0 < p_i, q_i < 1$, for all $i = 1, \ldots, n$. Denoting by r_i either p_i or q_i, with the assumptions that data are independent under both hypotheses, we have $\mathbf{X}^n \sim \prod_{i=1}^n r_i^{x_i}(1 - r_i)^{1-x_i}$, and the statistical test can be formalized as follows:

$$\begin{array}{ll} \mathcal{H}_1 : & r_i = p_i, \\ \mathcal{H}_0 : & r_i = q_i, \end{array} \qquad i = 1, 2, \ldots, n. \tag{1}$$

The optimal detection statistic, in log-likelihood form, is

$$\sum_{i=1}^{n} \left[x_i \log \frac{p_i}{q_i} + (1 - x_i) \log \frac{1 - p_i}{1 - q_i} \right]. \tag{2}$$

When we speak of "detection statistic" we mean a quantity that, compared to a threshold, leads to a decision in favor of $\mathcal{H}_1$ if the threshold is strictly crossed, and for $\mathcal{H}_0$ otherwise.

With unlabeled (i.e., *shuffled*) data we are faced with a binary hypothesis test in which we get the values taken by the n observations, but the position of these observations inside vector $\mathbf{X}^n$ is lost: there is no way to associate an observed value to its original position. This can be formalized by saying that we observe one of the $n!$ permutations of the entries of $\mathbf{X}^n$, but we do not know which. In formula: $\mathbf{X}^n \sim \prod_{i=1}^n r_i^{x_{\pi(i)}}(1-r_i)^{1-x_{\pi(i)}}$, where $\pi(i) \in \{1, \ldots, n\}$ is the new index assigned to the sample originally appearing at the i-th position, upon applying the permutation. Thus, with unlabeled data, we have to solve test (1) when π is unknown. The lack of knowledge of the actual sample positions inside vector $\mathbf{x}^n$ implies that only the *type* of vector $\mathbf{x}^n$ is available, namely, we observe the number $k_{\mathbf{x}}$ of ones appearing in $\mathbf{x}^n$.

In this situation, one possible approach is to resort to the GLRT [13], which amounts to estimating with the ML criterion the unknown permutation, separately under $\mathcal{H}_1$ and under $\mathcal{H}_0$, and then computing the (log of the) ratio of the likelihoods, with the estimated permutations in place of the unknown ones. Denoting by $\hat{\pi}_1$ and $\hat{\pi}_0$ the two ML estimates under $\mathcal{H}_1$ and $\mathcal{H}_0$, respectively, the GLRT statistic is

$$\begin{aligned} &\textstyle\sum_{i=1}^n \left[x_{\hat{\pi}_1(i)} \log p_i + (1 - x_{\hat{\pi}_1(i)}) \log(1 - p_i) \right] \\ &- \textstyle\sum_{i=1}^n \left[x_{\hat{\pi}_0(i)} \log q_i + (1 - x_{\hat{\pi}_0(i)}) \log(1 - q_i) \right] \end{aligned} \tag{3}$$

Note that $\hat{\pi}_1$ can be written as

$$\arg\max_{\pi} \sum_{i=1}^{n} \left[x_{\pi(i)} \log p_i + (1 - x_{\pi(i)}) \log(1 - p_i) \right], \tag{4}$$

and a similar expression holds for $\hat{\pi}_0$, with p_i replaced by q_i.

3. DETECTORS FOR SHUFFLED DATA

Let us consider the 2-by-n matrix

$$\begin{pmatrix} \log p_1 & \log p_2 & \log p_3 & \ldots & \log p_n \\ \log(1 - p_1) & \log(1 - p_2) & \log(1 - p_3) & \ldots & \log(1 - p_n) \end{pmatrix} \tag{5}$$

and suppose that we observe the number of ones $k_{\mathbf{x}}$ appearing in $\mathbf{x}^n$. From (4) we see that computation of $\hat{\pi}_1$ amounts to selecting $k_{\mathbf{x}}$ distinct columns over the first row of (5) and other $n - k_{\mathbf{x}}$ distinct columns over the second row, in such a way that the sum of the selected matrix entries is maximized. Likewise, computation of $\hat{\pi}_0$ is equivalent to maximizing the sum of the entries obtained by selecting $k_{\mathbf{x}}$ distinct columns over the first row and other $n - k_{\mathbf{x}}$ distinct columns over the second row of:

$$\begin{pmatrix} \log q_1 & \log q_2 & \log q_3 & \ldots & \log q_n \\ \log(1 - q_1) & \log(1 - q_2) & \log(1 - q_3) & \ldots & \log(1 - q_n) \end{pmatrix}. \tag{6}$$

The following proposition shows how these maximizations can be done.

PROPOSITION 1 *(Computing $\hat{\pi}_1, \hat{\pi}_0$) The ML estimate $\hat{\pi}_1$ is obtained by selecting the $k_{\mathbf{x}}$ distinct columns with largest entries over the first row of (5), and the remaining $n - k_{\mathbf{x}}$ distinct columns over the second row. The ML estimate $\hat{\pi}_0$ is obtained similarly, by considering matrix (6) in place of matrix (5).* □

Proof: The proof is elementary. Let $(p_{(1)}, p_{(2)}, \ldots, p_{(n)})$ be the decreasing ordered version of $(p_1, p_2, \ldots, p_n)$, namely $p_{(1)} \geq p_{(2)} \geq \cdots \geq p_{(n)}$. If $\hat{\pi}_1$ is obtained as stated in Proposition 1, the corresponding log-likelihood [see (4)] is:

$$\sum_{i=1}^{k_{\mathbf{x}}} \log p_{(i)} + \sum_{i=k_{\mathbf{x}}+1}^{n} \log(1 - p_{(i)}), \tag{7}$$

and we have to show that no other permutation yields a value larger than (7). A different permutation amounts to moving one or more indices from the first sum appearing in (7) to the second, and vice versa, so that the number of indices in the two sums remains unchanged. It is seen that any such

modification would decrease, or leave unchanged, the value of (7). Indeed, suppose that index $\ell \leq k_{\mathbf{x}}$ would be moved to the second sum, and $m > k_{\mathbf{x}}$, to the first. The value of (7) would be changed by $\log p_{(m)} - \log p_{(\ell)} + \log(1 - p_{(\ell)}) - \log(1 - p_{(m)}) \leq 0$. The same argument applies under $\mathcal{H}_0$. •

The structure of the GLRT is given by the following proposition where where void sums are defined as zero.

PROPOSITION 2 *(GLRT) The GLRT statistic is given by*

$$\mathcal{S}_{\mathrm{GLRT}} = \sum_{i=1}^{k_{\mathbf{x}}} \log \frac{p_{(i)}}{q_{(i)}} + \sum_{i=k_{\mathbf{x}}+1}^{n} \log \frac{1 - p_{(i)}}{1 - q_{(i)}}. \qquad (8)$$

□

Proof: Follows immediately from (3) and (7). •

For the case of general finite alphabets, an alternative to the GLRT has been proposed in the literature under the name of ULR (unlabeled likelihood ratio), see [12]. When observations are binary, the ULR statistic reduces to $k_{\mathbf{x}} \log \frac{\bar{p}}{\bar{q}} + (n - k_{\mathbf{x}}) \log \frac{1-\bar{p}}{1-\bar{q}}$, where $\bar{p} = \frac{1}{n} \sum_{i=1}^{n} p_i$, and $\bar{q} = \frac{1}{n} \sum_{i=1}^{n} q_i$. The ULR statistic can be equivalently cast in the form

$$\mathcal{S}_{\mathrm{ULR}} = \mathrm{sign}(\bar{p} - \bar{q})\, k_{\mathbf{x}}, \qquad (9)$$

where the signum function is defined as: $\mathrm{sign}(z) = 1$ for $z > 0$, $\mathrm{sign}(z) = -1$ for $z < 0$, and $\mathrm{sign}(0) = 0$.

In [12] two greedy approaches, named Algorithm-A and Algorithm-B, have been also proposed for labeling.

PROPOSITION 3 *(Algorithms) With binary observations, both algorithms in [12] find the exact ML estimates* $\hat{\pi}_1, \hat{\pi}_0$. □

3.1. Finite Number of Classes

In practical applications, when n is large, it makes sense to assume that the n-vectors $(p_1, \ldots, p_n)$ and $(q_1, \ldots, q_n)$ contain only a relatively smaller number of different entries. This happens, for instance, in sensor networks where a few sensors make many independent measurements each, and these data are sent to a fusion center without preserving their provenance. Data coming from the same sensor share the same statistical distribution, which, presumably, is instead different from sensor to sensor. With m sensors, we may speak of m *classes* of measurements. Accordingly, suppose that

$$p = (\underbrace{p_{c1}, \ldots, p_{c1}}_{n_1}, \underbrace{p_{c2}, \ldots, p_{c2}}_{n_2}, \ldots\ldots, \underbrace{p_{cm}, \ldots, p_{cm}}_{n_m}), \qquad (10)$$

$$q = (\underbrace{q_{c1}, \ldots, q_{c1}}_{n_1}, \underbrace{q_{c2}, \ldots, q_{c2}}_{n_2}, \ldots\ldots, \underbrace{q_{cm}, \ldots, q_{cm}}_{n_m}), \qquad (11)$$

with $\sum_{\ell=1}^{m} n_\ell = n$. Note that the subscript ci in (10)-(11) denotes the i-th class.

In the m-class case of (10)-(11), it is very simple to characterize statistically $k_{\mathbf{X}}$. Indeed:

$$k_{\mathbf{X}} = \sum_{i=1}^{n_1} X_i + \sum_{i=n_1+1}^{n_1+n_2} X_i + \cdots + \sum_{\sum_{k=1}^{m-1} n_k + 1}^{n} X_i, \qquad (12)$$

so that, under $\mathcal{H}_1$ (respectively, $\mathcal{H}_0$), the distribution of $k_{\mathbf{X}}$ is the convolution of m binomial distributions with number of trials n_ℓ and success probability $p_{c\ell}$ (resp. $q_{c\ell}$), $\ell = 1, \ldots, m$. Distributions of this type appear frequently in reliability [14], economy/finance [15–17], and healthcare [18] contexts. In general, they do not admit a closed-form expression, and various approximations have been developed, see [19].

Denoting by $\mathbb{P}_j(k_{\mathbf{X}} = k_{\mathbf{x}})$ the PMF (probability mass function) obtained by this convolution, under $\mathcal{H}_j$, $j = 0, 1$, the optimal detection statistic for unlabeled detection is (for simplicity, we do not consider randomized tests [20]):

$$\mathcal{S}_{\mathrm{OPT}} = \frac{\mathbb{P}_1(k_{\mathbf{X}} = k_{\mathbf{x}})}{\mathbb{P}_0(k_{\mathbf{X}} = k_{\mathbf{x}})}. \qquad (13)$$

The explicit form of the PMFs $\mathbb{P}_j(k_{\mathbf{X}} = k_{\mathbf{x}})$, $j = 0, 1$, is difficult to manipulate, even for m in the order of few units. However, when $\min\{n_1, \ldots, n_m\}$ is large enough, a substantial simplification is obtained by invoking the De Moivre-Laplace theorem for approximating the CDF (cumulative distribution function) of each binomial distribution with the Gaussian CDF. Under $\mathcal{H}_1$ we get (with obvious notation) $k_{\mathbf{X}} \sim \mathcal{N}\left(\mu_1, \sigma_1^2\right)$, where $\mu_1 = \sum_{\ell=1}^{m} n_\ell\, p_{c\ell} = n\bar{p}$, and $\sigma_1^2 = \sum_{\ell=1}^{m} n_\ell\, p_{c\ell}\,(1 - p_{c\ell})$. The equivalent of these parameters under $\mathcal{H}_0$, namely $\mu_0 = n\bar{q}$ and σ_0^2, are obtained by replacing $p_{c\ell}$ with $q_{c\ell}$, $\ell = 1, \ldots, m$. De Moivre-Laplace theorem is an instance of the CLT, and CLT allows us to approximate the PMF (not only the CDF) of $k_{\mathbf{X}}$ with the samples of the Gaussian density. This is because the random variable $k_{\mathbf{X}}$ is of lattice type [21]. Accordingly, an approximation of the optimal detection statistic is given by the ratio $f(k_{\mathbf{x}}; \mu_1, \sigma_1^2)/f(k_{\mathbf{x}}; \mu_0, \sigma_0^2)$ where $f(z; \mu, \sigma^2)$ represents the Gaussian PDF (probability density function) with mean μ and variance σ^2, computed at the point z. Namely, the decision statistic can be cast in the form:

$$\mathcal{S}_{\mathrm{CLT}} = \left(\frac{k_{\mathbf{x}} - n\bar{q}}{\sigma_0}\right)^2 - \left(\frac{k_{\mathbf{x}} - n\bar{p}}{\sigma_1}\right)^2. \qquad (14)$$

Clearly, if $\sigma_0 = \sigma_1$, then $\mathcal{S}_{\mathrm{CLT}}$ is equivalent to $\mathcal{S}_{\mathrm{ULR}}$. Note that, for the special case of $m = 2$, computing the exact PMF of $k_{\mathbf{X}}$ is straightforward because only one convolution is required.

4. NUMERICAL EXPERIMENTS

Let $P_m = \mathbb{P}_1(\mathcal{S} \leq \gamma)$ be the probability of a miss, and $P_f = \mathbb{P}_0(\mathcal{S} > \gamma)$ the probability of false alarm, where $\mathcal{S}$ denotes generically one of the previously introduced detection statistics. The value of P_m versus P_f can be obtained by varying the threshold level γ, and we now investigate the relationship P_m vs. P_f, starting with two, somehow extreme, examples.

Suppose we have $n = 200$ observations, and $m = 2$ classes, with $n_1 = 100$ and $n_2 = 100$. Suppose also that data under $\mathcal{H}_0$ are equiprobable and IID (independent and identically distributed). This means that under $\mathcal{H}_0$ data labeling

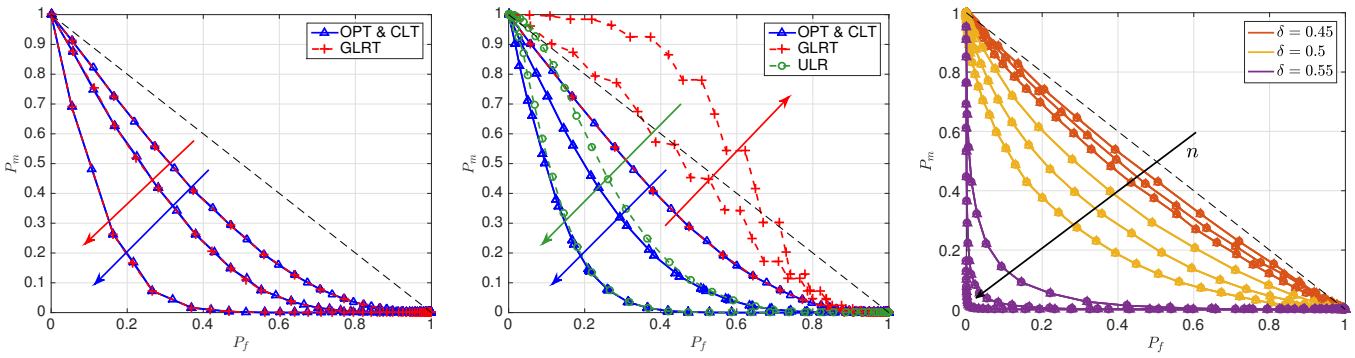

Fig. 1. Results of computer simulations by using 10^4 Monte Carlo runs for each point. Points are joined for easier visualization. *Left:* An experiment with $m = 2$ classes and $n_1 = n_2 = 100$. Detection performance with $q_{c1} = q_{c1} = .5$, and three values of p: in the direction of the arrows we have $(p_{c1} = .9, p_{c2} = .1)$, $(p_{c1} = .95, p_{c2} = .05)$, and $(p_{c1} = .99, p_{c2} = .01)$. *Middle:* The same as in *Left*, with the values of p: $(p_{c1} = .9, p_{c2} = .1)$, $(p_{c1} = .95, p_{c2} = .1)$, and $(p_{c1} = .99, p_{c2} = .1)$. *Right:* An example with $m = 10$ classes, each with n/m entries, and $n = 50, 100, 200$, following the direction of the arrow. The probabilities q_{ci}, $i = 1, \ldots, 10$, are generated uniformly at random in the interval $(.45, .55)$, and p_{ci}, $i = 1, \ldots, 10$, uniformly at random in $(\delta, \delta + 0.1)$. Note that in all cases the curves of $\mathcal{S}_{\rm CLT}$, $\mathcal{S}_{\rm GLRT}$ and $\mathcal{S}_{\rm ULR}$ are perfectly superimposed.

carries no information, and any permutation of the data results in one and the same statistical distribution for $\mathbf{X}^n$: $\mathbb{P}_0(X_i = 1) = 1/2, \forall i = 1, \ldots, n$. With the notation adopted in (10)-(11), we have $q_{c1} = q_{c2} = .5$. Under $\mathcal{H}_1$, let us consider three scenarios, $(p_{c1} = .9, p_{c2} = .1)$, $(p_{c1} = .95, p_{c2} = .05)$, and $(p_{c1} = .99, p_{c2} = .01)$. Note that in all cases $\bar{p} = \bar{q} = 0.5$.

Figure 1-left shows the curves P_m versus P_f for $\mathcal{S}_{\rm OPT}$, $\mathcal{S}_{\rm CLT}$, and $\mathcal{S}_{\rm GLRT}$, see eqs. (13), (14), (8), respectively. The performance of $\mathcal{S}_{\rm OPT}$ and $\mathcal{S}_{\rm CLT}$ coincide, meaning that the Gaussian approximation works well. It is remarkable that the curves of these detectors also coincide with that of $\mathcal{S}_{\rm GLRT}$. This can be explained by noting that in this example $\mu_0 = \mu_1 = n/2$, $\sigma_0^2 = n/4$, $\sigma_1^2 = np_{c1}(1 - p_{c1})$, and then the CLT statistic (14) is equivalent to compare $-|k_{\mathbf{x}} - n/2|$ to a threshold. Moreover, for this example, the GLRT statistic in (8) reduces to the form

$$-|k_{\mathbf{x}} - n/2| \, \log \frac{p_{c1}}{1 - p_{c1}} + n \log(2\, p_{c1}), \qquad (15)$$

revealing that $\mathcal{S}_{\rm GLRT}$ is equivalent to $\mathcal{S}_{\rm CLT}$. The curve of the $\mathcal{S}_{\rm ULR}$ is not shown in Fig. 1-left because, as we see from (9), the ULR detector no longer makes sense when $\bar{p} = \bar{q}$.

In the second experiment, we still assume $q_{c1} = q_{c2} = .5$, and we set $(p_{c1} = .9, p_{c2} = .1)$ as before, and then $(p_{c1} = .95, p_{c2} = .1)$, and $(p_{c1} = .99, p_{c2} = .1)$. In the last two cases $\bar{p}$ is slightly larger than $\bar{q}$, a small difference with a large impact. As seen in Fig. 1-middle, the performance of $\mathcal{S}_{\rm OPT}$ and $\mathcal{S}_{\rm CLT}$ still coincides. As to the ULR, that the larger is $\bar{p}$ with respect to $\bar{q}$, the more nearly optimal is the ULR. In this situation, indeed, the number of ones makes a clear difference between the hypotheses, and the ULR detector exploits that.

The behavior of $\mathcal{S}_{\rm GLRT}$ changes drastically. For $p_{c1} = .95$ and .99, we see that the GLRT performs very poorly, and for $p_{c1} = .99$ is biased, in the sense that, at least for small values of P_f, the test would work better by inverting the decisions [20, p. 47]. This can be justified by computing $\mathbb{E}_j[\mathcal{S}_{\rm GLRT}]$ under $\mathcal{H}_j$, $j = 0, 1$. With $n_1 = n_2 = n/2$, $n = 200$, $p_{c1} = .99$, $p_{c2} = .1$, $q_{c1} = q_{c2} = q_c = .5$, the computation is straightforward, and we find that $\mathbb{E}_1[\mathcal{S}_{\rm GLRT}]$ is slightly smaller than $\mathbb{E}_0[\mathcal{S}_{\rm GLRT}]$ (≈ 107.31 against ≈ 107.95).

The third experiment, described in the figure caption, shows that, even in the presence of modest differences between the vectors $(p_{c1}, \ldots, p_{cm})$ and $(q_{c1}, \ldots, q_{cm})$, all the decision rules may share the same performance, which improves by increasing this difference, and by increasing n.

5. SUMMARY

Unlabeled detection with binary data is special compared to the case of alphabets with arbitrary (finite) cardinality. The MLE of the permutations can be obtained from Algorithm-A and -B in [12] in general contexts. But in the binary case, the MLE is explicit. It requires matching according to the ordered versions of $(p_1, p_2, \ldots, p_n)$ and $(q_1, q_2, \ldots, q_n)$, as stated in Proposition 1; and the GLRT follows as shown in Proposition 2. The structure of the GLRT in (8) sheds light on its behavior, which can be surprisingly poor.

Computational simplicity has been advocated as the major motivation for the adoption of the ULR detector. As shown by simulations in [12], its performance is in many cases quite satisfying, perhaps unexpectedly. In the binary case, in the light of (9), it is easy to figure out in what cases the ULR works. In the presence of observations from m classes, the CLT approximation (14) works fine even for moderately small values of $\min\{n_1, \ldots, n_m\}$. It exploits both μ_i and σ_i. When $\sigma_0 \approx \sigma_1$, $\mathcal{S}_{\rm ULR}$ performs approximately as $\mathcal{S}_{\rm CLT}$. On the other hand, with $\mu_1 \approx \mu_0$, $\mathcal{S}_{\rm ULR}$ loses efficacy, because it is blind to the change-in-variance, and the difference with $\mathcal{S}_{\rm CLT}$ may become significant.

6. REFERENCES

[1] J. Unnikrishnan, S. Haghighatshoar, and M. Vetterli, "Unlabeled sensing with random linear measurements," *IEEE Transactions on Information Theory*, vol. 64, no. 5, pp. 3237–3253, May 2018.

[2] ——, "Unlabeled sensing: Solving a linear system with unordered measurements," in *2015 53rd Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, Sept 2015, pp. 786–793.

[3] S. Thrun and J. J. Leonard, "Simultaneous localization and mapping," in *Springer Handbook of Robotics*, B. Siciliano and O. Khatib, Eds. Heidelberg: Springer Berlin, 2008, pp. 871–889.

[4] S. Haghighatshoar and G. Caire, "Signal recovery from unlabeled samples," *IEEE Transactions on Signal Processing*, vol. 66, no. 5, pp. 1242–1257, March 2018.

[5] A. Abid, A. Poon, and J. Zou. (2017, May 4) Linear regression with shuffled labels. [Online]. Available: http://arxiv.org/abs/1705.01342

[6] A. Pananjady, M. J. Wainwright, and T. A. Courtade. (2017, April 24) Denoising linear models with permutated data. [Online]. Available: http://arxiv.org/abs/1704.07461

[7] ——, "Linear regression with an unknown permutation: Statistical and computational limits," in *2016 54th Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, Sept 2016, pp. 417–424.

[8] ——. (2016, August 9) Linear regression with an unknown permutation: statistical and computational limits. [Online]. Available: https://arxiv.org/abs/1608.02902

[9] G. Wang, J. Zhu, R. S. Blum, P. Willett, S. Marano, V. Matta, and P. Braca, "Signal amplitude estimation and detection from unlabeled binary quantized samples," *IEEE Transactions on Signal Processing*, vol. 66, no. 16, pp. 4291–4303, Aug. 2018.

[10] J. Zhu, H. Cao, C. Song, and Z. Xu, "Parameter estimation via unlabeled sensing using distributed sensors," *IEEE Communications Letters*, vol. 21, no. 10, pp. 2130–2133, Oct 2017.

[11] S. Marano, V. Matta, P. Willett, P. Braca, and R. Blum, "Hypothesis testing in the presence of Maxwell's daemon: Signal detection by unlabeled observations," in *Proc. of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2017)*, New Orleans, LA, USA, 5-9 Mar. 2017.

[12] S. Marano and P. Willett, "Algorithms and fundamental limits for unlabeled detection using types," *IEEE Transactions on Signal Processing*, Oct. 2018, under revision. Available: http://arxiv.org/abs/1810.07563.

[13] S. M. Kay, *Fundamentals of Statistical Signal Processing, Volume II: Detection Theory*. Englewood Cliffs, New Jersey: Prentice Hall, 1998.

[14] P. J. Boland and F. Proschan, "The reliability of k out of n systems," *The Annals of Probability*, vol. 11, pp. 760–764, 1983.

[15] R. K. Sah, "Comparative properties of sums of independent binomials with different parameters," *Econ. Lett.*, no. 31, pp. 27–30, 1989.

[16] D. G. and A. P. Robinson, "Confidence intervals for the weighted sum of two independent binomial proportions," *Australian & New Zealand Journ. Stat.*, no. 54, pp. 281–289, 2012.

[17] J. Van der Hoek and R. J. Elliott, *Binomial models in finance*. New York: Springer, 2006.

[18] J. Benneyan and A. Taşeli, "Exact and approximate probability distributions of evidence-based bundle composite compliance measures," *Health Care Manag.*, no. 13, pp. 193–209, 2010.

[19] K. Butler and M. A. Stephens, "The distribution of a sum of independent binomial random variables," *Methodol. Comput. Appl. Probab.*, no. 19, pp. 557–571, 2017.

[20] H. V. Poor, *An Introduction to Signal Detection and Estimation*. New York: Springer-Verlag, 1988.

[21] A. Papoulis, *Probability, Random Variables, and Stochastic Processes*, 3rd ed. New York: McGraw-Hill, 1991.