Order Statistics-based “Anti-Bayesian” Parametric Classification for Asymmetric Distributions in the Exponential Family

A. Thomas, and B. John Oommen

Abstract—Although the field of parametric Pattern Recognition (PR) has been thoroughly studied for over five decades, the use of the Order Statistics (OS) of the distributions to achieve this has not been reported. The pioneering work on using OS for classification was presented in [1] for the Uniform distribution, where it was shown that optimal PR can be achieved in a counter-intuitive manner, diametrically opposed to the Bayesian paradigm, i.e., by comparing the testing sample to a few samples distant from the mean. This must be contrasted with the Bayesian paradigm in which, if we are allowed to compare the testing sample with only a single point in the feature space from each class, the optimal strategy would be to achieve this based on the (Mahalanobis) distance from the corresponding central points, for example, the means. In [2], we showed that the results could be extended for a few symmetric distributions within the exponential family. In this paper, we attempt to extend these results significantly by considering asymmetric distributions within the exponential family, for some of which even the closed form expressions of the cumulative distribution functions are not available. These distributions include the Rayleigh, Gamma and certain Beta distributions. As in [1] and [2], the new scheme, referred to as Classification by Moments of Order Statistics (CMOS), attains an accuracy very close to the optimal Bayes’ bound, as has been shown both theoretically and by rigorous experimental testing.

Keywords—Classification using Order Statistics (OS), Exponential family, Moments of OS.

I. INTRODUCTION

The basis for statistical pattern classification is that the individual classes are characterized by their distributions. These distributions have numerous indicators such as their means, variances etc., and these indices have, traditionally, played a prominent role in achieving pattern classification, and in designing the corresponding training and testing algorithms.

It is also well known that a distribution has many other characterizing indicators, for example, those related to its Order Statistics (OS). The interesting point about these indicators is that some of them are quite unrelated to the traditional moments themselves, and in spite of this, have not been used in achieving PR. The main question that we shall consider is whether these indicators/indices possess any potential in PR.

The amazing answer to this question is that OS can be used in PR, and that such classifiers operate in a completely “anti-Bayesian” manner, i.e., by only considering certain outliers of the distribution. This must be contrasted with Bayesian classifiers which attain the optimal lower bound, and that often reduces to testing the sample point using the corresponding distances/norms to the means or the “central points” of the distributions.

Earlier, in [1] and [2], we showed that we could obtain optimal results by an “anti-Bayesian” paradigm by using the OS. This was done in [1] for the Uniform distribution and in [2] for certain distributions within the exponential family. Those results, though very fascinating, were possible because the closed forms of the cumulative distributions were available.

In this paper, we attempt to extend these results significantly by considering asymmetric distributions within the exponential family, for some of which even the closed form expressions of the cumulative distribution functions are not available. Examples of these distributions are the Rayleigh, Gamma and certain Beta distributions. Again, as in [1] and [2], we show the completely counter-intuitive result that by working with a very few (sometimes as small as two) points distant from the mean, one can obtain remarkable classification accuracies, and this has been demonstrated both theoretically and by experimental verification. Interestingly enough, the novel methodology that we propose, referred to as Classification by Moments of Order Statistics (CMOS), is computationally not any more complex than working with the Bayesian paradigm itself.

Contributions of this Paper: The novel contributions of this paper are:

- We propose an “anti-Bayesian” paradigm for the classification of patterns within the parametric mode of computation, where the distance computations are not with regard to the “mean” but with regard to some samples “distant” from the mean. These points, which are sometimes as few as two, are the moments of OS of the distributions;
- We demonstrate that the proposed approach can attain
the optimal bound for symmetric distributions and near-optimal bound for non-symmetric distributions;

- To justify these claims, we submit a formal analysis and the results of various experiments which have been performed for a few distributions within the exponential family (for which even the closed form expressions of the distributions are not available), and the results are clearly conclusive.

Our results for classification using the OS are both pioneering and novel.

II. Relevant Background Areas Regarding Order Statistics

Let \( x_1, x_2, \ldots, x_n \) be a univariate random sample of size \( n \) that follows a continuous distribution function \( \Phi \), where the probability density function (pdf) is \( \psi(y) \). Let \( x_{1,n}, x_{2,n}, \ldots, x_{n,n} \) be the corresponding Order Statistics (OS). The \( r^{th} \) OS, \( x_{r,n} \), of the set is the \( r^{th} \) smallest value among the given random variables. The pdf of \( y = x_{r,n} \) is given by:

\[
f_y(y) = \frac{n!}{(r-1)!(n-r)!} (1 - \Phi(y))^{r-1} \Phi(y)^{n-r} \psi(y),
\]

where \( r = 1, 2, \ldots, n \). The reasoning for the above expression is straightforward. If the \( r^{th} \) OS appears at a location given by \( y = x_{r,n} \), it implies that the \( r-1 \) smaller elements of the set are drawn independently from a Binomial distribution with a probability \( \Phi(y) \), and the other \( n-r \) samples are drawn using the probability \( 1 - \Phi(y) \). The factorial terms result from the fact that the \( r-1 \) elements can be independently chosen from the set of \( n \) elements.

Using the distribution \( f_y(y) \), the \( k^{th} \) moment of \( x_{r,n} \), \( E[x_{r,n}^k] \), can be formulated as:

\[
E[x_{r,n}^k] = \frac{n!}{(r-1)!(n-r)!} \int_{-\infty}^{+\infty} y^k \Phi(y)^{k-1}(1 - \Phi(y))^{n-r} \psi(y) dy,
\]

provided that both sides of the equality exist [3], [4].

The fundamental theorem concerning the OS we invoke is found in many papers [4]–[6]. The theorem can be summarized as follows.

Let \( n \geq r \geq k + 1 \geq 2 \) be integers. Then, since \( \Phi \) is a nondecreasing and right-continuous function from \( \mathbb{R} \rightarrow \mathbb{R} \), \( \Phi(x_{r,n}) \) is uniform in \([0,1]\). If we now take the \( k^{th} \) moment of \( \Phi(x_{r,n}) \), it has the form [5]:

\[
E[\Phi^k(x_{r,n})] = \frac{B(r + k, n - r + 1)}{B(r, n - r + 1)} = \frac{n! (r + k - 1)!}{(n + k)! (r - 1)!},
\]

where \( B(a, b) \) denotes the Beta function, and \( B(a, b) = (a-1)/\Gamma(b) \) since its parameters are integers.

The above fundamental result can also be used for characterization purposes as explained in [5], [7]. The implications of the above are the following:

1) If \( n = 2 \), implying that only two samples are drawn from \( x \), we can deduce from Eq. (1) that:

\[
E[\Phi^1(x_{1,2})] = \frac{1}{3} \implies E[x_{1,2}] = \Phi^{-1} \left( \frac{1}{3} \right), \quad (2)
\]

\[
E[\Phi^1(x_{2,2})] = \frac{2}{3} \implies E[x_{2,2}] = \Phi^{-1} \left( \frac{2}{3} \right). \quad (3)
\]

Thus, from a computational perspective, the first moment of the first and second 2-order OS would be the values where the cumulative distribution \( \Phi \) equal \( 1/3 \) and \( 2/3 \) respectively.

2) For any \( n \geq 2 \), implying that we are considering the \( k^{th} \) OS from \( n \) samples drawn from \( x \), we can deduce from Eq. (1) that:

\[
E[\Phi^1(x_{k,n})] = \frac{k}{n + 1} \implies E[x_{k,n}] = \Phi^{-1} \left( \frac{k}{n + 1} \right), \quad (4)
\]

and

\[
E[\Phi^1(x_{n-k,n})] = \frac{n - k + 1}{n + 1} \implies E[x_{n-k,n}] = \Phi^{-1} \left( \frac{n - k + 1}{n + 1} \right). \quad (5)
\]

Although the analogous expressions can be derived for the higher order moments of these OS, for the rest of this paper we shall merely focus on the first moment of these OS, and derive the consequences of using them in classification.

III. Optimal OS-based Classification: The Generic Classifier

Let us assume that we are dealing with the 2-class problem with classes \( \omega_1 \) and \( \omega_2 \), where their class-conditional densities are \( f_1(x) \) and \( f_2(x) \) respectively (i.e., their corresponding distributions are \( F_1(x) \) and \( F_2(x) \)) respectively. Let \( \nu_1 \) and \( \nu_2 \) be the corresponding medians of the distributions. Then, classification based on \( \nu_1 \) and \( \nu_2 \) would be the strategy that classifies samples based on a single OS. We can see that for all symmetric distributions, this classification accuracy attains the Bayes’ accuracy.

This result is not too astonishing because the median is centrally located close to (if not exactly) on the mean. The result for higher order OS is actually far more intriguing because the higher order OS are not located centrally (close to the means), but rather distant from the means. Consequently, we shall show that for a large number of distributions, mostly from the exponential family [2], the classification based on these OS again attains the Bayes’ bound. These results are now extended for asymmetric exponential distributions.

1 Throughout this section, we will assume that the a priori probabilities are equal.
IV. THE RAYLEIGH DISTRIBUTION

The Rayleigh distribution is a continuous probability distribution which is often observed when the overall magnitude of a vector is related to its directional components whose applications are found in [8] and [9].

The pdf of the Rayleigh distribution, with parameter \( \sigma > 0 \) is \( \varphi(x, \sigma) = \frac{x}{\sigma^2} e^{-x^2/2\sigma^2}, x \geq 0 \) and the cumulative distribution function is \( \Phi(x) = 1 - e^{-x^2/2\sigma^2}, x \geq 0 \). The mean, the variance, and the median of the Rayleigh distribution are \( \sigma \sqrt{\frac{\pi}{2}}, \frac{4\sigma^2}{3}, \text{ and } \sigma \sqrt{\ln(4)} \), respectively.

Theoretical Analysis: Rayleigh Distribution - 2-OS

The typical PR problem involving the Rayleigh distribution would consider two classes \( \omega_1 \) and \( \omega_2 \) where the class \( \omega_2 \) is displaced by a quantity \( \theta \) and the values of \( \omega_1 \) and \( \omega_2 \) respectively. We consider the scenario when \( \sigma_1 = \sigma_2 = \sigma \). Consider the distributions: \( f(x, \sigma) = \frac{x}{\sigma^2} e^{-x^2/2\sigma^2} \) and \( f(x - \theta, \sigma) = \frac{x - \theta}{\sigma^2} e^{-\frac{(x - \theta)^2}{2\sigma^2}} \). In order to do the classification based on CMOS, we shall first derive the moments of the 2-OS for the Rayleigh distribution. By virtue of Eq. (2) and (3), the expected values of the first moments of the two OS can be obtained by determining the points where the cumulative distribution function attains the values of \( \frac{1}{2} \) and \( \frac{1}{4} \) respectively. Let \( u_1 \) be the point for the percentile \( \frac{1}{2} \) of the first distribution, and \( u_2 \) be the point for the percentile \( \frac{1}{4} \) of the second distribution. Then:

\[
\int_{0}^{u_1} \frac{x}{\sigma^2} e^{-x^2/2\sigma^2} dx = \frac{2}{3} \quad \Rightarrow \quad u_1 = \sigma \sqrt{2 \ln(3)}, \quad (6)
\]

\[
u_2 = \theta + \sigma \sqrt{2 \ln \left( \frac{3}{2} \right)} \quad (7)
\]

We now consider the efficiency of the CMOS.

Theorem 1: For the 2-class problem in which the two class conditional distributions are Rayleigh and identical, the accuracy obtained by CMOS, when using two OS, deviates from the optimal Bayes' bound as the solution of the transcendental equality

\[
\ln \left( \frac{x}{x - \theta} \right) = \frac{-\theta^2 + 2 \theta x}{2 \sigma^2} \quad (9)
\]

The discriminant is then the solution to the transcendental equation:

\[
\ln \left( \frac{x}{x - \theta} \right) = \frac{-\theta^2 + 2 \theta x}{2 \sigma^2} \quad (9)
\]

We now consider the classification with respect to the expected values of the moments of the 2-OS, \( u_1 \) and \( u_2 \), where as per Eq. (15) and (16), \( u_1 = \sigma \sqrt{2 \ln(3)} \) and \( u_2 = \theta + \sigma \sqrt{2 \ln \left( \frac{3}{2} \right)} \). The discriminant enforced by the 2-OS classifier satisfies:

\[
D(x, u_1) = D(x, u_2).
\]

The condition imposed by Eq. (10) leads to the following:

\[
\Rightarrow \quad D \left( x, \sigma \sqrt{2 \ln(3)} \right) = D \left( x, \theta + \sigma \sqrt{2 \ln \left( \frac{3}{2} \right)} \right)
\]

\[
\Rightarrow \quad 2x = \theta + \sigma \sqrt{2 \ln \left( \frac{3}{2} \right)} + \sigma \sqrt{2 \ln \left( \frac{3}{2} \right)}
\]

\[
\Rightarrow \quad x = \frac{\theta}{2} + \sigma \sqrt{2 \ln \left( \frac{3}{2} \right)} \cdot \left( \sqrt{\ln(3)} + \sqrt{\ln \left( \frac{3}{2} \right)} \right) \quad (11)
\]

The difference in the errors of the two classifiers is clearly related to differences in the corresponding discriminant functions specified by Eq. (9) and (11). Hence the theorem.

Remark: Another way of comparing the approaches is by obtaining the error difference created by the CMOS classifier when compared to the Bayesian classifier. In Figure 1, the small area marked as “Error Difference” is the difference between the probability of error formed by the CMOS classifier when compared to the Bayesian counterpart, and we can evaluate this area by using the corresponding definite integrals. As Eq. (9) is transcendental in nature, the only way to find the Bayesian classifier is to resort to a numerical strategy, for example, by using a Taylor series expansion. The uncertainty under the curve (in percentage) is depicted in Table I. From this table, we can see that the CMOS classifier is bounded by an error difference of less than 0.15%, which is truly, negligible.

<table>
<thead>
<tr>
<th>( \theta )</th>
<th>1</th>
<th>1.5</th>
<th>2</th>
<th>2.5</th>
<th>3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Max. Bounded Error(% )</td>
<td>0.15</td>
<td>0.06</td>
<td>0.05</td>
<td>0.001</td>
<td>0</td>
</tr>
</tbody>
</table>

Theorem 2: For the 2-class problem in which the two class conditional distributions are Rayleigh and identical, CMOS, the accuracy obtained by classification using two OS deviates from the classifier which discriminates based on the distance from the corresponding medians as \( \frac{\theta}{2} + \sigma \sqrt{2 \ln \left( \frac{3}{2} \right)} \) deviates from

\[
\frac{\theta}{2} + \sigma \sqrt{2 \ln \left( \frac{3}{2} \right)} \cdot \left( \sqrt{\ln(3)} + \sqrt{\ln \left( \frac{3}{2} \right)} \right)
\]
This equation simplifies to:
\[
D(x, u_1) = D(x, u_2)
\]
\[
\implies D\left(x, \sigma \sqrt{2 \ln(\theta)}\right) = D\left(x, \theta + \sigma \sqrt{2 \ln \left(\frac{3}{2}\right)}\right)
\]
\[
\implies 2\pi = \theta + \sigma \sqrt{2 \ln(\theta) + \frac{3}{2}}
\]
\[
\implies x = \frac{\theta}{2} + \sigma \sqrt{\frac{\ln(\theta) + \frac{3}{2}}{2}}. \tag{14}
\]

The difference in the errors of the two classifiers is clearly related to differences in the corresponding discriminant functions specified by Eq. (12) and (14). Hence the theorem. □

**Corollary 1:** By virtue of the almost-identical nature of the two expressions for the Rayleigh distribution, the classification using the proximity to the median is almost indistinguishable from that of the Bayesian classifier.

Proof: The proof of this corollary is straightforward and omitted here, but can be found in [9]. □

**Data Generation: Rayleigh Generation** We made use of a Uniform \((0, 1)\) random variable generator to generate data values that follow a Rayleigh distribution. The expression
\[
x = \sigma \sqrt{-2 \ln(1 - u)},
\]
where \(\sigma\) is the parameter and \(u\) is a random variate from the Uniform distribution \(U(0, 1)\), generates Rayleigh distributed values [10].

**Experimental Results: Rayleigh Distribution - 2-OS**

The CMOS classifier was rigorously tested for a number of experiments with various Rayleigh distributions having the identical parameter \(\sigma\). In every case, the 2-OS CMOS gave almost the same classification as that of the Bayesian classifier. The method was executed 50 times with the 10-fold cross validation scheme. The test results are tabulated in Table II. The results presented justify the claims of Theorems 1 and 2.

**Theoretical Analysis: Rayleigh Distribution - \(k\)-OS** We have seen from Theorem 1 that for the Rayleigh distribution, the moments of the 2-OS are sufficient for a near-optimal classification. As in the case of the other distributions, we shall now consider the scenario when we utilize other \(k\)-OS. Let \(u_1\) be the point for the percentile \(\frac{n+1-k}{n+1}\) of the first distribution, and \(u_2\) be the point for the percentile \(\frac{k}{n+1}\) of the second distribution. Then:
\[
\int_0^{u_1} \frac{x}{\sigma^2} e^{-x^2/2\sigma^2} dx = \frac{n + 1 - k}{n + 1}
\]
\[
\implies u_1 = \sigma \sqrt{2 \ln \left(\frac{n + 1}{k}\right)}, \quad \text{and} \quad \tag{15}
\]
\[
u_2 = \theta + \sigma \sqrt{2 \ln \left(\frac{n + 1}{n + 1 - k}\right)}. \tag{16}
\]

The \(k\)-OS results of CMOS follow.
TABLE II
A COMPARISON OF THE ACCURACY OF THE BAYESIAN AND THE 2-OS CMOS CLASSIFIER FOR THE RAYLEIGH DISTRIBUTION.

<table>
<thead>
<tr>
<th></th>
<th>3</th>
<th>2.5</th>
<th>2</th>
<th>1.5</th>
<th>1</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bayesian</td>
<td>99.1</td>
<td>97.35</td>
<td>94.45</td>
<td>87.75</td>
<td>78.80</td>
</tr>
<tr>
<td>CMOS</td>
<td>99.1</td>
<td>97.35</td>
<td>94.40</td>
<td>87.75</td>
<td>78.65</td>
</tr>
</tbody>
</table>

Theorem 3: For the 2-class problem in which the two class conditional distributions are Rayleigh and identical, a near-optimal Bayesian classification can be achieved by using symmetric pairs of the n-OS, i.e., the n - k OS for \( \omega_1 \) and the k OS for \( \omega_2 \) if and only if

\[
\sqrt{\ln \left( \frac{n+1}{n+1-k} \right)} < \frac{\theta}{\sigma \sqrt{2}}.
\]

The classification obtained by CMOS deviates from the optimal Bayes’ bound as the solution of the transcendental equality

\[
\ln \left( \frac{x}{x-\sigma} \right) = -\frac{\sigma^2 + 2\theta}{2\sigma^2}
\]

deviates from

\[
\frac{\theta}{2} + \frac{\sigma}{\sqrt{2}} \sqrt{\ln \left( \frac{n+1}{k} \right) + \ln \left( \frac{n+1}{n+1-k} \right)}.
\]

Proof: The proof of this theorem is omitted here, but is included in [9].

Experimental Results: Rayleigh Distribution - k-OS

The CMOS method has been rigorously tested with different possibilities of the k-OS and for various values of n, and the test results are given in Table III. For the distribution under consideration, the Bayesian approach provides an accuracy of 82.5%, and from the table, it is obvious that some of the considered k-OSs attains the optimal accuracy and the rest of the cases attain near-optimal accuracy. Also, we can see that the approach fails if the condition stated in Theorem 3 is not satisfied.

To clarify the table, consider the cases in which the 6-OS were invoked for the classification. For 6-OS, the possible symmetric OS pairs could be (1, 6), (2, 5), and (3, 4) respectively. Observe that the expected values for the first moment of the k-OS has the form \( E[x_{k,n}] = \sigma \sqrt{2 \ln \left( \frac{n+1}{k} \right)} \). For the cases where the condition

\[
\frac{\theta}{\sigma \sqrt{2}} \left( \sqrt{\ln \left( \frac{n+1}{k} \right)} - \sqrt{\ln \left( \frac{n+1}{n+1-k} \right)} \right) < \frac{\theta}{\sigma \sqrt{2}}
\]

the accuracy attained is either optimal or near-optimal, as indicated by the results in the table (denoted by Trial Nos. 5 and 6). Now, consider the results presented in the row denoted by Trial No. 7. In this case where the CMOS positions were \( \theta \sqrt{2 \ln \left( \frac{1}{k} \right)} \) and \( \theta + \sigma \sqrt{2 \ln \left( \frac{1}{k} \right)} \), the inequality of the condition imposed in Theorem 3 simplified to \( 1.002339 < 0.88388 \), which is not valid. Observe that if

\[
\sqrt{\ln \left( \frac{n+1}{k} \right)} - \sqrt{\ln \left( \frac{n+1}{n+1-k} \right)} > \frac{\theta}{\sigma \sqrt{2}},
\]

the symmetric pairs should be reversed to obtain the near-optimal Bayes’ bound. This concludes our study on the CMOS for the Rayleigh distribution.

V. THE GAMMA DISTRIBUTION

The Gamma distribution is a continuous probability distribution with two parameters - a, a shape parameter and b, a scale parameter. The pdf of the Gamma distribution is

\[
f(x; a, b) = \frac{1}{\Gamma(a) b^a} x^{a-1} e^{-x/b}, \quad a > 0, \quad b > 0,
\]

with mean ab and variance \( ab^2 \) where a and b are the parameters. Unfortunately, the cumulative distribution function does not have a closed form expression [11]-[13].

Theoretical Analysis: Gamma Distribution

The typical PR problem invoking the Gamma distribution would consider two classes \( \omega_1 \) and \( \omega_2 \) where the class \( \omega_2 \) is displaced by a quantity \( \theta \), and in the case analogous to the ones we have analyzed, the values of the scale and shape parameters are identical. We consider the scenario when \( a_1 = a_2 = a \) and \( b_1 = b_2 = b \). Thus, we consider the distributions:

\[
f(x, 2, 1) = xe^{-x} \quad \text{and} \quad f(x - \theta, 2, 1) = (x - \theta)e^{-(x-\theta)}.
\]

We first derive the moments of the 2-OS, which are the points of interest for CMOS, for the Gamma distribution. Let \( u_1 \) be the point for the percentile \( \frac{1}{3} \) of the first distribution, and \( u_2 \) be the point for the percentile \( \frac{2}{3} \) of the second distribution. Then

\[
\int_0^{u_1} xe^{-x}dx = \frac{2}{3}
\]

or \( \ln(u_1) - 2u_1 = \ln(\frac{1}{3}) \), and

\[
\ln(u_2 - \theta) - 2(u_2 - \theta) = \ln(\frac{1}{3}) - \ln(\theta).
\]

The following results hold for the Gamma distribution.

Theorem 4: For the 2-class problem in which the two class conditional distributions are Gamma and identical, the accuracy obtained by CMOS, the classification using two OS, deviates from the accuracy attained by the classifier with regard to the distance from the corresponding medians as \( 1.7391 + \frac{b}{a} \) deviates from \( 1.6783 + \frac{b}{a} \).

Proof: The proof of this theorem can be found in [9].

Data Generation: Gamma Distribution

There are a number of data generation algorithms reported for the Gamma distribution, all of which make use of the Uniform random variate \( U(0, 1) \). The data is generated using the built-in function available in MatLab, namely \( \text{gamrnd}(a, b, sz) \), where \( a \) is the shape parameter, \( b \) is the scale parameter, and \( sz \) is the size of the array. To be specific, \( \text{gamrnd}(2, 1, 10) \) will generate 100 values that follow the Gamma distribution with the shape
TABLE III
A Comparison of The Accuracy of The Bayesian (i.e., 82.5%) and The k-OS CMOS Classifier For The Rayleigh Distribution By Using The Symmetric Pairs Of The OS For Different Values Of n. The Value of σ and θ were set to be 2 and 1.5 respectively. Note that in every case, CMOS attained near-optimal accuracy whenever the conditions stated in Theorem 3 were satisfied.

<table>
<thead>
<tr>
<th>No.</th>
<th>Order(n)</th>
<th>Moments</th>
<th>OS₁</th>
<th>OS₂</th>
<th>CMOS</th>
<th>Pass/Fail</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Two</td>
<td>(k, 1/k)</td>
<td>σ√(2 ln (1/k))</td>
<td>θ + σ√(2 ln (1/k))</td>
<td>82.05</td>
<td>Passed</td>
</tr>
<tr>
<td>2</td>
<td>Four</td>
<td>(k−1/2, 1), 1 ≤ i ≤ 2</td>
<td>σ√(2 ln (1/k))</td>
<td>θ + σ√(2 ln (1/k))</td>
<td>81.8</td>
<td>Passed</td>
</tr>
<tr>
<td>3</td>
<td>Four</td>
<td>(k−1/2, 1), 1 ≤ i ≤ 2</td>
<td>σ√(2 ln (1/k))</td>
<td>θ + σ√(2 ln (1/k))</td>
<td>82.0</td>
<td>Passed</td>
</tr>
<tr>
<td>4</td>
<td>Six</td>
<td>(k−1/2, 1), 1 ≤ i ≤ 2</td>
<td>σ√(2 ln (1/k))</td>
<td>θ + σ√(2 ln (1/k))</td>
<td>18.4</td>
<td>Failed</td>
</tr>
<tr>
<td>5</td>
<td>Six</td>
<td>(k−1/2, 1), 1 ≤ i ≤ 2</td>
<td>σ√(2 ln (1/k))</td>
<td>θ + σ√(2 ln (1/k))</td>
<td>82.10</td>
<td>Passed</td>
</tr>
<tr>
<td>6</td>
<td>Six</td>
<td>(k−1/2, 1), 1 ≤ i ≤ 2</td>
<td>σ√(2 ln (1/k))</td>
<td>θ + σ√(2 ln (1/k))</td>
<td>82.15</td>
<td>Passed</td>
</tr>
<tr>
<td>7</td>
<td>Eight</td>
<td>(k−1/2, 1), 1 ≤ i ≤ 2</td>
<td>σ√(2 ln (1/k))</td>
<td>θ + σ√(2 ln (1/k))</td>
<td>18.45</td>
<td>Failed</td>
</tr>
<tr>
<td>8</td>
<td>Eight</td>
<td>(k−1/2, 1), 1 ≤ i ≤ 2</td>
<td>σ√(2 ln (1/k))</td>
<td>θ + σ√(2 ln (1/k))</td>
<td>82.05</td>
<td>Passed</td>
</tr>
<tr>
<td>9</td>
<td>Eight</td>
<td>(k−1/2, 1), 1 ≤ i ≤ 2</td>
<td>σ√(2 ln (1/k))</td>
<td>θ + σ√(2 ln (1/k))</td>
<td>82.15</td>
<td>Passed</td>
</tr>
</tbody>
</table>

Experimental Results: Gamma Distribution - 2-OS The CMOS classifier was rigorously tested for a number of experiments with Gamma distributions having the identical shape and scale parameters a₁ = a₂ = 2, and b₁ = b₂ = 1. In every case, the 2-OS CMOS gave almost the same classification as the classifier based on the central moments, namely, the mean and the median. The method was executed 50 times with the 10-fold cross validation scheme. The test results are tabulated in Table IV.

Theorem 5: For the 2-class problem in which the two class conditional distributions are Gamma and identical, a near-optimal Bayesian classification can be achieved by using certain symmetric pairs of the n-OS, i.e., the (n−k)th OS for \( \omega_1 \) (represented as \( u_1 \)) and the kth OS for \( \omega_2 \) (represented as \( u_2 \)) if and only if \( u_1 < u_2 \).

Proof: The proof of this theorem is included in [9].

Experimental Results: Gamma Distribution - k-OS The CMOS method has been rigorously tested for numerous symmetric pairs of the k-OS and for various values of n, and a subset of the test results are given in Table V. Experiments have been performed for different values of \( \theta \), and we can see that the CMOS attained near-optimal Bayes’ bound. Also, we can see that the approach fails if the condition stated in Theorem 5 is not satisfied.

Interestingly enough, if we examine the table, we can see that the Bayes’ accuracy is the highest value except for the case where \( \theta = 3.0 \), although this result must, in fact, be considered as an aberration. This concludes the study of the Gamma distribution with regard to the CMOS classification.

VI. The Beta Distribution

The Beta distribution is a family of continuous probability distributions defined in (0, 1) parameterized by two shape parameters \( \alpha \) and \( \beta \). The distribution can take different shapes based on the specific values of the parameters. If the parameters are identical, the distribution is symmetric with respect to \( \frac{1}{2} \). Further, if \( \alpha = \beta = 1 \), \( B(1,1) \) becomes \( U(0,1) \). The pdf of the Beta distribution is \( f(x; \alpha, \beta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha) \Gamma(\beta)} x^{\alpha-1}(1-x)^{\beta-1} \). The mean and the variance of the distribution are \( \frac{\alpha}{\alpha+\beta} \) and \( \frac{\alpha \beta}{(\alpha+\beta)^2 (\alpha+\beta+1)} \) respectively. We consider the case when \( \alpha = \beta > 1 \).

For this study, we mainly consider three cases:

- \( \alpha = 1, \beta > 1 \): Uniform Distribution.
- \( \alpha = \beta > 1 \): Symmetric about \( \frac{1}{2} \).
- \( \alpha > 1, \beta > 1 \): Unimodal Distribution.

Earlier, in paper [8], when we first introduced the concept of CMOS-based PR, we had analyzed the 2-OS and k-OS CMOS for the Uniform distribution, and had provided the corresponding theoretical analysis and the experimental results. We had concluded that, for the 2-class problem in which the two class conditional distributions are Uniform and identical, CMOS can, indeed, attain the optimal Bayes’ bound. So, in this paper, to avoid repetition, we skip the analysis for the Beta distribution, \( B(1,1) \), as this case reduces to the analysis for Uniform \( U(0,1) \). Thus, we reckon that the first of these cases (i.e., when \( \alpha = 1 \) and \( \beta = 1 \)) as being closed.

We now proceed to consider the Beta distribution in which \( \alpha = \beta \).

Theoretical Analysis: Beta Distribution (\( \alpha = \beta \)) Consider two classes \( \omega_1 \) and \( \omega_2 \) where the class \( \omega_2 \) is displaced by a quantity \( \theta \), and the values of the shape parameters are identical. We consider the scenario when \( \alpha_1 = \alpha_2 = \alpha, \beta_1 = \beta_2 = \beta \), and for the sake of simplicity, \( \alpha = \beta = 2 \). Then, the distributions are:

\[
\begin{align*}
\text{PDF of } \omega_1 & = f(x; 2, 2) = 6x(1-x) \\
\text{PDF of } \omega_2 & = f(x-\theta; 2, 2) = 6(x-\theta)(1-x+\theta).
\end{align*}
\]
We first derive the moments of the 2-OS, which are the points of interest for CMOS, for the Beta distribution. By virtue of Eq. (2) and (3), the expected values of the first moments of the two OS can be obtained by determining the points where the cumulative distribution function attains the values of $\frac{1}{\theta}$ and $\frac{1}{\theta^2}$ respectively. As the distribution takes different forms based on the values of the shape parameters, we have to solve each case separately, and so we can obtain numerical values for the CMOS positions. Let $u_1$ be the point for the percentile $\frac{2}{3}$ of the first distribution, and $u_2$ be the point for the percentile $\frac{1}{3}$ of the second distribution. Then:

$$\int_0^{u_1} 6x(1-x)dx = \frac{2}{3}$$

$$\Rightarrow -6u_1^3 + 9u_1^2 - 2 = 0. \quad (19)$$

Similarly, if we don’t take the displacement, $\theta$, into consideration, the form for $u_2$ leads to the equation:

$$-6u_2^3 + 9u_2^2 - 1 = 0. \quad (20)$$

We shall now prove that the CMOS, indeed, attains the Bayes’ bound.

**Theorem 6:** For the 2-class problem in which the two class conditional distributions are Beta($\alpha$, $\beta$) ($\alpha = \beta$) and identical, CMOS, the classification using two OS, attains an accuracy that is exactly identical to the optimal Bayes’ bound.

**Proof:** Without loss of generality, let the distributions of $\omega_1$ and $\omega_2$ be $B(x, 2, 2)$ and $B(x - \theta, 2, 2)$, where $\theta$ is the displacement for the second distribution. Then, to get the Bayes’ classifier, we argue that:

$$p(x|\omega_1)P(\omega_1) \leq \frac{\omega_1}{\omega_2} p(x|\omega_2)P(\omega_2)$$

$$\Rightarrow \quad 6x(1-x) \frac{\omega_1}{\omega_2} \leq \frac{6(x - \theta)(1 - (x - \theta))}{\omega_2}$$

$$\Rightarrow \quad x \leq \frac{\omega_2}{\omega_1} \frac{\theta + 1}{2}. \quad (21)$$

We now consider the classification with respect to the expected values of the moments of the 2-OS, $u_1$ and $u_2$. In order to prove our claim, we need to show that:

$$x \leq \frac{\omega_2}{\omega_1} \frac{\theta + 1}{2} \quad \Rightarrow \quad D(x, u_1) \leq \frac{\omega_1}{\omega_2} D(x, u_2). \quad (22)$$

If we examine the Eqs. (19) and (20), we can see that Eq. (20) can be obtained by substituting $1 - u_2$ for $u_1$ in Eq. (19) as:

$$-6(1 - u_2)^3 + 9(1 - u_2)^2 - 2 = 0 \Rightarrow -6u_2^3 + 9u_2^2 - 1 = 0. \quad (23)$$

From this, it is obvious that $u_2 = \theta + u_1 - 1$. Consequently,
the RHS of the claim given by Eq. (22) leads to the following:
\[
D(x, u_1) \overset{\omega_1}{\approx} D(x, u_2)
\]
\[
\implies D(x, u_1) \overset{\omega_2}{\approx} D(x, \theta + 1 - u_1)
\]
\[
\implies \frac{x}{\omega_2} = \frac{\theta + 1}{2} \tag{24}
\]

The result follows by observing that Eqs. (21) and (22) are identical comparisons. Hence the theorem.

**Experimental Results: Beta Distribution** ($\alpha = \beta - 2$-OS)
The CMOS has been rigorously tested for various Beta distributions with 2-OS with $\alpha = \beta = 2$. In the interest of brevity, a few typical results are given below. For each of the experiments, we generated 1,000 points for the classes $\omega_1$ and $\omega_2$ characterized by $B(x, 2, 2)$ and $B(x - \theta, 2, 2)$ respectively. We then invoked a classification procedure by utilizing the Bayesian and the CMOS strategies. In every case, CMOS was compared with the Bayesian classifier for different values of $\theta$, as tabulated in Table VI. The results were obtained by executing each algorithm 50 times using a 10-fold cross-validation scheme. The results given in this table justify the claim of Theorem 6.

We have seen from Theorem 6 that the moments of the 2-OS are sufficient for the classification to attain a Bayes’ bound. Now we shall examine the scenario where the $k$-OS CMOS is invoked, and thus determine the strength of the proposed method.

Let $u_1$ be the point for the percentile $\frac{n+1-k}{n+1}$ of the first distribution, and $u_2$ be the point for the percentile $\frac{k}{n+1}$ of the second distribution. Then:
\[
\int_0^{u_1} 6x(1-x)dx = \frac{n + 1 - k}{n + 1}
\]
\[
\implies -2u_1^3 + 3u_1^2 - \frac{n + 1 - k}{n + 1} = 0. \tag{25}
\]

By a similar argument, if we ignore the displacement $\theta$, the CMOS point for the $\frac{k}{n+1}$ percentile of the second distribution leads to the equation:
\[
-2u_2^3 + 3u_2^2 - \frac{k}{n + 1} = 0. \tag{26}
\]

We shall now prove that the CMOS attains the Bayes’ bound.

**Theorem 7**: For the 2-class problem in which the two class conditional distributions are Beta and identical as $B(x, \alpha, \beta)$ and $B(x - \theta, \alpha, \beta)$ where $\alpha = \beta = 2$, optimal Bayesian classification can be achieved by using symmetric pairs of the $n$-OS, i.e., the $n - k$ OS for $\omega_1$ (represented by $u_1$) and the $k$ OS for $\omega_2$ (represented by $u_2$) if and only if $u_1 < u_2$.

*Proof*: The proof of this theorem is included in [9] and omitted here in the interest of space.

**Experimental Results: Beta Distribution** ($\alpha = \beta - k$-OS)
The CMOS method has also been tested for the Beta distribution for other $k$ OS when $\alpha = \beta = 2$. In the interest of brevity, we merely cite one example where the distributions for $\omega_1$ and $\omega_2$ were characterized by $B(x, 2, 2)$ and $B(x - \theta, 2, 2)$ respectively. For each of the experiments, we generated 1,000 points for each class, and the testing samples were classified based on the selected symmetric pairs for values $k$ and $n - k$ respectively. A subset of the results are found in Table VII.

To clarify the table, consider the cases in which the $8$-OS were invoked for the classification. For 8-OS, the possible symmetric OS pairs could be $(1, 8), (2, 7), and (4, 5)$ respectively. Wherever the condition $u_1 < u_2$ is satisfied, the CMOS attained the optimal Bayes’ bound, as indicated by the results in the table (denoted by Trial Nos. 8 and 9). Now, consider the results presented in the row denoted by Trial No. 7. In this case where the CMOS positions were 0.79269 and $\theta + 0.20731$, the inequality of the condition imposed in Theorem 7 simplified to 0.79269 $< 0.78731$, which is not valid. Observe that if $u_1 > u_2$, the symmetric pairs should be reversed to obtain the optimal Bayes’ bound. This concludes the study on the symmetric Beta distribution.

**VIII. Conclusions**

In this paper, we have shown that optimal classification for symmetric distributions and near-optimal bound for asymmetric distributions can be attained by an “anti-Bayesian” approach, i.e., by working with a very few (sometimes as small as two) points distant from the mean. This scheme, referred to as CMOS, Classification by Moments of Order Statistics, operates by using these points determined by the Order Statistics of the distributions. In this paper, we have proven the claim for some distributions within the exponential family, and the theoretical results have been verified by rigorous experimental testing. Our results for classification using the OS are both pioneering and novel.

**References**


TABLE VI
A COMPARISON OF THE ACCURACY OF THE BAYESIAN AND THE 2-OS CMOS CLASSIFIER FOR THE BETA DISTRIBUTION $B(2, 2)$ FOR DIFFERENT VALUES OF $\theta$.

<table>
<thead>
<tr>
<th>$\theta$</th>
<th>0.95</th>
<th>0.90</th>
<th>0.85</th>
<th>0.80</th>
<th>0.75</th>
<th>0.70</th>
<th>0.65</th>
<th>0.60</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bayesian</td>
<td>99.845</td>
<td>99.45</td>
<td>98.185</td>
<td>96.94</td>
<td>94.95</td>
<td>92.86</td>
<td>90.31</td>
<td>88.075</td>
</tr>
<tr>
<td>CMOS</td>
<td>99.845</td>
<td>99.45</td>
<td>98.185</td>
<td>96.94</td>
<td>94.95</td>
<td>92.86</td>
<td>90.31</td>
<td>88.075</td>
</tr>
</tbody>
</table>

TABLE VII

<table>
<thead>
<tr>
<th>Trial No.</th>
<th>Order(n)</th>
<th>Moments</th>
<th>$OS_1$</th>
<th>$OS_2$</th>
<th>CMOS</th>
<th>Pass/Fail</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Two</td>
<td>$\langle \frac{2}{3}, \frac{1}{3} \rangle$</td>
<td>0.61304</td>
<td>$\theta + 0.38696$</td>
<td>87.3</td>
<td>Passed</td>
</tr>
<tr>
<td>2</td>
<td>Four</td>
<td>$\langle \frac{2}{5}, \frac{1}{5} \rangle$</td>
<td>0.71286</td>
<td>$\theta + 0.28714$</td>
<td>87.3</td>
<td>Passed</td>
</tr>
<tr>
<td>3</td>
<td>Four</td>
<td>$\langle \frac{2}{5}, \frac{2}{5} \rangle$</td>
<td>0.56707</td>
<td>$\theta + 0.43293$</td>
<td>87.3</td>
<td>Passed</td>
</tr>
<tr>
<td>4</td>
<td>Six</td>
<td>$\langle \frac{2}{3}, \frac{1}{3} \rangle$</td>
<td>0.7621</td>
<td>$\theta + 0.23790$</td>
<td>87.3</td>
<td>Passed</td>
</tr>
<tr>
<td>5</td>
<td>Six</td>
<td>$\langle \frac{2}{5}, \frac{1}{5} \rangle$</td>
<td>0.6471</td>
<td>$\theta + 0.3529$</td>
<td>87.3</td>
<td>Passed</td>
</tr>
<tr>
<td>6</td>
<td>Six</td>
<td>$\langle \frac{2}{3}, \frac{1}{3} \rangle$</td>
<td>0.54776</td>
<td>$\theta + 0.45224$</td>
<td>87.3</td>
<td>Passed</td>
</tr>
<tr>
<td>7</td>
<td>Eight</td>
<td>$\langle \frac{2}{5}, \frac{1}{5} \rangle$</td>
<td>0.79269</td>
<td>$\theta + 0.20731$</td>
<td>12.7</td>
<td>Failed</td>
</tr>
<tr>
<td>8</td>
<td>Eight</td>
<td>$\langle \frac{2}{5}, \frac{2}{5} \rangle$</td>
<td>0.69508</td>
<td>$\theta + 0.30492$</td>
<td>87.3</td>
<td>Passed</td>
</tr>
<tr>
<td>9</td>
<td>Eight</td>
<td>$\langle \frac{2}{3}, \frac{1}{3} \rangle$</td>
<td>0.55711</td>
<td>$\theta + 0.46269$</td>
<td>87.3</td>
<td>Passed</td>
</tr>
</tbody>
</table>


