Joint Distributions for Two Useful Classes of Statistics, with Applications to Classification and Hypothesis Testing

Albert. H. Nuttall and Paul M. Baggenstoss *
Naval Undersea Warfare Center
Newport RI, 02841
p.m.baggenstoss@ieee.org (EMAIL)

January 10, 2002

Abstract

In this paper, we analyze the statistics of two general classes of statistics. The first class is "M quadratic and linear forms of correlated Gaussian random variables". Examples include both cyclic and non-cyclic autocorrelation function (ACF) estimates of a correlated Gaussian process or the magnitude-squared of the output samples of a filtered Gaussian process. The second class consists of a subset of order statistics together with a remainder term. An example is the largest $M-1$ bins of a discrete Fourier transform (DFT) or discrete wavelet transform (DWT), together with the sum of the remaining energies, forming an M-dimensional statistic. Both classes of statistics are useful in classification and detection of signals. In this paper, we solve for the joint probability density functions (PDFs) of both classes. Using the PDF projection method, these results can be used to transform the feature PDFs into the corresponding high-dimensional PDFs of the raw input data.

1 Introduction and Motivation

The so-called M-ary classification problem is that of assigning a multidimensional sample of data $\mathbf{x} \in \mathcal{R}^N$ to one of M classes. The statistical hypothesis that class j is true is denoted by H_j, $1 \leq j \leq M$. The statistical characterization of $\mathbf{x}$ under each of the M hypotheses is described completely by the joint PDFs, written $p(\mathbf{x}|H_j)$, $1 \leq j \leq M$. Classical theory applied to the problem results in the so-called Bayes classifier, which simplifies to the Neyman-Pearson rule for equi-probable prior probabilities, namely,

$$j^* = \arg\max_j \; p(\mathbf{x}|H_j). \tag{1}$$

Because this classifier attains the minimum probability of error of all possible classifiers, it is the basis of most classifier designs. Unfortunately, it does not provide simple solutions to the dimensionality problem that arises when the joint PDFs are unknown and must be estimated. The most common solution is to reduce the dimension of the data, by extraction of a small number of information-bearing features $\mathbf{z} = T(\mathbf{x})$, and then re-casting the classification problem in terms of $\mathbf{z}$:

$$j^* = \arg\max_j \; p(\mathbf{z}|H_j). \tag{2}$$

*This work was supported by the Office of Naval Research.

To be optimal, the feature-based classifier (2) requires that

$$\frac{p(\mathbf{x}|H_j)}{p(\mathbf{x}|H_k)} = \frac{p(\mathbf{z}|H_j)}{p(\mathbf{z}|H_k)} \tag{3}$$

for any two classes j, k. Thus, the feature space $\mathbf{z}$ must optimally separate any pair of classes. This requirement is only achieved in the simplest of problems. In trying to achieve (3), we encounter a fundamental tradeoff - whether to discard features in an attempt to reduce the dimension to something manageable - or to include them and suffer the problems associated with estimating a joint PDF with high dimensionality. Unfortunately, there may be no acceptable solution where there is both adequate information content in $\mathbf{z}$ and low enough dimension for robust PDF estimation. Virtually all methods which attempt to find decision boundaries on a high-dimensional space are subject to this tradeoff or "curse" of dimensionality. For this reason, many researchers have explored the possibility of using class-specific features.

The basic idea in using class-specific features is to extract M class-specific feature sets, $\mathbf{z}_j = T_j(\mathbf{x}),\ \ 1 \leq j \leq M$, where the dimension of each feature set is small, and then to arrive at a decision rule based only upon functions of the lower-dimensional features. Unfortunately, the classifier modeled on the Neyman-Pearson rule,

$$j^* = \arg\max_j \ p(\mathbf{z}_j|H_j), \tag{4}$$

is invalid because comparisons of densities on different feature spaces are meaningless. A number of approaches have emerged in recent years to arrive at meaningful decision rules [1] [2] [3] [4] [5] [6]. All these methods are based on strong assumptions, so are not general solutions. The class-specific method [7], however, is an optimal classifier based on a sufficiency assumption much milder than the standard feature-based classifier and is otherwise completely general. Provided a suitable reference hypothesis $H_{0,j}$ can be found for each class such that $p(\mathbf{x}|H_{0,j})$ and $p(\mathbf{z}_j|H_{0,j})$ are both known, the projected PDF of $\mathbf{x}$ may be constructed:

$$\hat{p}(\mathbf{x}|H_j) \triangleq \frac{p(\mathbf{x}|H_{0,j})}{p(\mathbf{z}_j|H_{0,j})}\, p(\mathbf{z}_j|H_j). \tag{5}$$

According to the PDF projection theorem [7], [8], the function $\hat{p}(\mathbf{x}|H_j)$ is guaranteed to be a PDF, and therefore is a PDF approximation to $p(\mathbf{x}|H_j)$. Accordingly, the class-specific classifier

$$j^* = \arg\max_j \ \frac{p(\mathbf{x}|H_{0,j})}{p(\mathbf{z}_j|H_{0,j})}\, p(\mathbf{z}_j|H_j) \tag{6}$$

can be constructed. The accuracy of the approximation depends only upon the statistical sufficiency of the class-specific feature set $\mathbf{z}_j$ in separating H_j from $H_{0,j}$. More precicely, the class-specific classifier requires for optimality that

$$\frac{p(\mathbf{x}|H_j)}{p(\mathbf{x}|H_{0,j})} = \frac{p(\mathbf{z}_j|H_j)}{p(\mathbf{z}_j|H_{0,j})} \tag{7}$$

for each class j. Thus, the feature space $\mathbf{z}_j$ must optimally separate a given class H_j from the hand-picked reference hypothesis $H_{0,j}$. This sufficiency requirement is far more manageable than (3) and is helped by the fact that while $\{H_j\}$ are *given* and cannot be chosen, the class-dependent reference hypotheses $\{H_{0,j}\}$ can be individually chosen.

The success of the method depends on being able to calculate the "correction terms"

$$Q(\mathbf{x}, T_j, H_{0,j}) \triangleq \frac{p(\mathbf{x}|H_{0,j})}{p(\mathbf{z}_j|H_{0,j})},$$

which we call "Q" functions. The numerator is often quite easy to write down, but the denominator is difficult to derive. That is the subject of this paper - the calculation of these denominator terms under a special reference hypothesis - the white Gaussian noise (WGN) hypothessis. A wide variety of solutions are already available where the standard normal distribution, or WGN hypothesis, is used for the reference hypothesis [9]. These include cyclic cepstrum and autocorrelation estimates from independent Gaussian noise samples. Through one-to-one transformation, these results can be applied as well to linear prediction (LPC) and reflection coefficients. In this paper, we add solutions for two new classes of statistics to those already available.

2 M Quadratic and Linear Forms of Correlated Random Variables

The second-order statistics of correlated Gaussian random variables (RVs) constitute an important set of statistics. Examples include the ACF estimates of a correlated Gaussian process or the magnitude-squared output samples of a linear filter. Applications exist in SONAR and RADAR detection and estimation problems. Examples include both cyclic and non-cyclic ACF estimates of *dependent* Gaussian noise samples.

2.1 Form of the statistics

The general form of the statistics of interest is

$$\mathbf{z} = T(\mathbf{x}) = [z_1 \; z_2 \ldots z_M]',$$

where $\{z_m\}$ are M quadratic forms,

$$z_m = \mathbf{x}'\mathbf{P}_m\mathbf{x} + \mathbf{p}_m'\mathbf{x} + q_m, \quad 1 \leq m \leq M, \tag{8}$$

and $\mathbf{x}$ is the N-by-1 real input data vector

$$\mathbf{x} = [x_1 \; x_2 \ldots x_N]',$$

$\{\mathbf{P}_m\}$ are M real symmetric[1] N-by-N matrices, $\{\mathbf{p}_m\}$ are N-by-1 vectors, and $\{q_m\}$ are scalars.

The challenge is to determine the joint PDF of $\mathbf{z}$ under a specified Gaussian hypothesis, that is,

$$p(\mathbf{z}; \mathbf{R}_x, \boldsymbol{\mu}_x, \{\mathbf{P}_m\}, \{\mathbf{p}_m\}, \{q_m\}),$$

where $N \times 1$ mean vector

$$\mathbf{E}(\mathbf{x}) = \boldsymbol{\mu}_x$$

and $N \times N$ covariance matrix

$$\mathbf{E}((\mathbf{x} - \boldsymbol{\mu})(\mathbf{x} - \boldsymbol{\mu})') = \mathbf{R}_x.$$

Example 1 Autocorrelation Function. *An example of a class of statistics which are of the form (8) are ACF estimates*

$$r_t = \frac{1}{N} \sum_{i=t+1}^{N} x_i \, x_{i-t}, \quad 0 \leq t \leq N-1. \tag{9}$$

[1] There is no loss of generality in assuming $\{\mathbf{P}_m\}$ are symmetric, because any anti-symmetric component of $\{\mathbf{P}_m\}$ will cancel out in the quadratic form.

Suppose we are interested only in a selected set of ACF samples at delays $t_1, t_2 \ldots t_M$. The problem is to obtain the joint PDF of the feature vector

$$\mathbf{z} = [r_{t_1} \; r_{t_2} \ldots r_{t_M}]',$$

denoted by

$$p(r_{t_1}, r_{t_2} \ldots r_{t_M}; N, \boldsymbol{\mu}_x, \mathbf{R}_x).$$

The elements of $\mathbf{z}$ *can be written as quadratic forms*

$$z_m = \mathbf{x}'\mathbf{P}_{t_m}\mathbf{x}, \quad 1 \le m \le M,$$

where

$$\mathbf{P}_0 = \tfrac{1}{N}\begin{bmatrix} 1 & 0 & 0 & 0 & \cdots \\ 0 & 1 & 0 & 0 & \cdots \\ 0 & 0 & 1 & 0 & \cdots \\ 0 & 0 & 0 & 1 & \cdots \\ \vdots & \vdots & \vdots & \vdots & \end{bmatrix},$$

$$\mathbf{P}_1 = \tfrac{1}{2N}\begin{bmatrix} 0 & 1 & 0 & 0 & \cdots \\ 1 & 0 & 1 & 0 & \cdots \\ 0 & 1 & 0 & 1 & \cdots \\ 0 & 0 & 1 & 0 & \cdots \\ \vdots & \vdots & \vdots & \vdots & \vdots \end{bmatrix},$$

$$\mathbf{P}_2 = \tfrac{1}{2N}\begin{bmatrix} 0 & 0 & 1 & 0 & \cdots \\ 0 & 0 & 0 & 1 & \cdots \\ 1 & 0 & 0 & 0 & \cdots \\ 0 & 1 & 0 & 0 & \cdots \\ \vdots & \vdots & \vdots & \vdots & \end{bmatrix},$$

and so forth. The pattern is such that $\mathbf{P}_k$ *is nonzero only on the super- and sub-diagonals spaced k away from the main diagonal.*

If the sample mean is subtracted from $\mathbf{x}$ *prior to calculation of the ACF estimates, the quadratic forms (8) still hold, but the elements of* $\{\mathbf{P}_k\}$ *are changed. For example, the j,k-th element of* $\mathbf{P}_0$ *is now $\delta_{jk} - 1/N$ instead of δ_{jk}, where δ_{jk} is the Kronecker delta; the remaining matrices* $\{\mathbf{P}_k\}$ *are more complicated, but each element in the matrices can be evaluated by means of a single sum.*

Equation (9) involves an aperiodic correlation of data $\mathbf{x}$. *The extension to cyclic correlation estimates can also be formulated in terms of quadratic forms by wrapping each of the diagonals. For example, for $N = 6$,* $\mathbf{P}_2$ *becomes*

$$\mathbf{P}_2 = \frac{1}{2N}\begin{bmatrix} 0 & 0 & 1 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 & 0 & 1 \\ 1 & 0 & 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 & 0 & 1 \\ 1 & 0 & 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 1 & 0 & 0 \end{bmatrix}. \tag{10}$$

Cross-correlations

$$z_m = \mathbf{u}'\mathbf{P}_m\mathbf{v}, \quad 1 \le m \le M, \tag{11}$$

can also be written as (8) if we define

$$\mathbf{x} = \begin{bmatrix} \mathbf{u} \\ \mathbf{v} \end{bmatrix}.$$

Example 2 Linear Filtering. *Let* $\mathbf{y}$ *be an* M*-by-1 output vector from a linear filtering operation*

$$\mathbf{y} = \mathbf{A}\ \mathbf{x},$$

where $\mathbf{A}$ *is an* M*-by-*N *filter matrix and* $\mathbf{x}$ *is a real* N*-by-1 input vector. Let* $\mathbf{z}$ *be the* M*-by-1 vector of squared values of the elements of* $\mathbf{y}$*. For example, let*

$$\mathbf{A} = \begin{bmatrix} a_2 & a_1 & a_0 & 0 & 0 & 0 \\ 0 & a_2 & a_1 & a_0 & 0 & 0 \\ 0 & 0 & a_2 & a_1 & a_0 & 0 \\ 0 & 0 & 0 & a_2 & a_1 & a_0 \end{bmatrix}. \tag{12}$$

Then, we may write the elements of $\mathbf{z}$ *as*

$$z_m = \mathbf{x}'\ \mathbf{A}'\ \mathbf{U}_m\ \mathbf{A}\ \mathbf{x}, \quad 1 \leq m \leq M,$$

where $\mathbf{U}_m$ *is an* M*-by-*M *matrix of all zeros except a single one on the main diagonal in location* m, m*.*

2.2 Compression of Parameters

Note that there is no loss of generality in assuming that $\mathbf{R}_x = \mathbf{I}_N$ and $\boldsymbol{\mu}_x = \underline{\mathbf{0}}$, where $\mathbf{I}_N$ is the N-by-N identity matrix and $\underline{\mathbf{0}}$ is the N-by-1 vector of zeros. This is because we can write

$$\begin{aligned} & p(\mathbf{z}; \mathbf{R}_x, \boldsymbol{\mu}_x, \{\mathbf{P}_m\}, \{\mathbf{p}_m\}, \{\mathbf{q}_m\}) = \\ & \qquad p(\mathbf{z}; \mathbf{I}_n, \underline{\mathbf{0}}, \{\tilde{\mathbf{P}}_m\}, \{\tilde{\mathbf{p}}_m\}, \{\tilde{\mathbf{q}}_m\}), \end{aligned}$$

where

$$\begin{aligned} \tilde{\mathbf{P}}_m &= \mathbf{C}\ \mathbf{P}_m\ \mathbf{C}', \\ \tilde{\mathbf{p}}_m &= \mathbf{C}\ \mathbf{p}_m + 2\mathbf{C}\mathbf{P}_m\boldsymbol{\mu}_x, \\ \tilde{q}_m &= q_m + \mathbf{p}_m'\boldsymbol{\mu}_x + \boldsymbol{\mu}_x'\mathbf{P}_m\boldsymbol{\mu}_x, \end{aligned}$$

and $\mathbf{C}$ is the Cholesky decomposition of $\mathbf{R}_x$,

$$\mathbf{R}_x = \mathbf{C}'\ \mathbf{C}.$$

2.3 Saddlepoint Approximation

Since no closed-form expression for the joint PDF of $\mathbf{z}$ in (8) is known, we apply the Saddlepoint approximation [10],[9]. To obtain the SPA, we need the joint cumulant generating function (CGF) of $\mathbf{z}$, namely,

$$c_z(\boldsymbol{\lambda}) \triangleq \log g_z(\boldsymbol{\lambda}),$$

where $g_z(\boldsymbol{\lambda})$ is the joint moment-generating function (MGF) of $\mathbf{z}$. Also, we need the first and second-order partial derivatives of $c_z(\boldsymbol{\lambda})$. Once these are known, the formulas in reference [9] may be used to obtain the SPA.

It is shown in [11] that

$$c_z(\boldsymbol{\lambda}) = -\frac{1}{2}\log|\mathbf{Q}(\boldsymbol{\lambda})| + \frac{1}{2}\mathbf{t}'(\boldsymbol{\lambda})\ \mathbf{Q}^{-1}(\boldsymbol{\lambda})\mathbf{t}(\boldsymbol{\lambda}) + u(\boldsymbol{\lambda}), \tag{13}$$

where

$$\mathbf{Q}(\boldsymbol{\lambda}) = \mathbf{I}_M - 2\mathbf{D}(\boldsymbol{\lambda}),$$

with

$$\mathbf{D}(\boldsymbol{\lambda}) \triangleq \sum_{m=1}^{M} \lambda_m \mathbf{P}_m, \quad \mathbf{t}(\boldsymbol{\lambda}) \triangleq \sum_{m=1}^{M} \lambda_m \mathbf{p}_m,$$

and

$$u(\boldsymbol{\lambda}) \triangleq \sum_{m=1}^{M} \lambda_m q_m.$$

The first-order partial derivatives are

$$\tfrac{\partial}{\partial\lambda_m}\, c_z(\boldsymbol{\lambda}) \quad = \quad \mathrm{tr}\left\{\mathbf{Q}^{-1}\mathbf{P}_m\right\} + \mathbf{p}_m'\mathbf{Q}^{-1}\mathbf{t} + \mathbf{t}'\mathbf{B}_m\mathbf{Q}^{-1}\mathbf{t},$$

for $1 \le m \le M$, and the second-order partial derivatives are

$$\begin{aligned} \tfrac{\partial^2}{\partial\lambda_l\partial\lambda_m}\, c_z(\boldsymbol{\lambda}) \quad &= \quad 2\ \mathrm{tr}\left\{\mathbf{B}_l\mathbf{B}_m\right\} + \mathbf{p}_l'\mathbf{Q}^{-1}\mathbf{p}_m \\ &\quad +2\mathbf{p}_l'\mathbf{B}_m\mathbf{Q}^{-1}\mathbf{t} + 2\mathbf{p}_m'\mathbf{B}_l\mathbf{Q}^{-1}\mathbf{t} \\ &\quad +4\mathbf{t}'\mathbf{B}_l\mathbf{B}_m\mathbf{Q}^{-1}\mathbf{t} \end{aligned}$$

where

$$\mathbf{B}_m(\boldsymbol{\lambda}) \triangleq \mathbf{Q}^{-1}(\boldsymbol{\lambda})\mathbf{P}_m,$$

and we drop the $(\boldsymbol{\lambda})$ dependence from $\mathbf{t}(\boldsymbol{\lambda})$, $\mathbf{Q}^{-1}(\boldsymbol{\lambda})$, and $\mathbf{B}_m(\boldsymbol{\lambda})$, for simplicity. The third and fourth derivatives, necessary for the first-order correction term of the SPA have also been worked out [11].

The equations simplify considerably if we assume that $\{\mathbf{p}_m\}$ and $\{q_m\}$ are all zero. We then have

$$c_z(\boldsymbol{\lambda}) = \log g_z(\boldsymbol{\lambda}) = -\frac{1}{2}\log|\mathbf{Q}(\boldsymbol{\lambda})|\,. \tag{14}$$

The first order partial derivatives reduce to

$$\tfrac{\partial}{\partial\lambda_m}\, c_z(\boldsymbol{\lambda}) \quad = \quad \mathrm{tr}\left\{\mathbf{B}_m\right\}$$

and the second order partial derivatives become

$$\tfrac{\partial^2}{\partial\lambda_l\partial\lambda_m}\, c_z(\boldsymbol{\lambda}) \quad = \quad 2\ \mathrm{tr}\left\{\mathbf{B}_l\mathbf{B}_m\right\}, \quad 1 \le l, m \le M.$$

Example 3 Autocorrelation Function. *We revisit example 1 and test the SPA for cyclic ACF estimates against the SPA solution in [9]. We used the cyclic ACF (e.g. equation 10) with $N = 32$ and $M = 3$. The results are shown in Figure 1. The two methods agree very closely. The differences are so small that they can be explained by differences in the stopping point of the iteration to find the saddlepoint.*

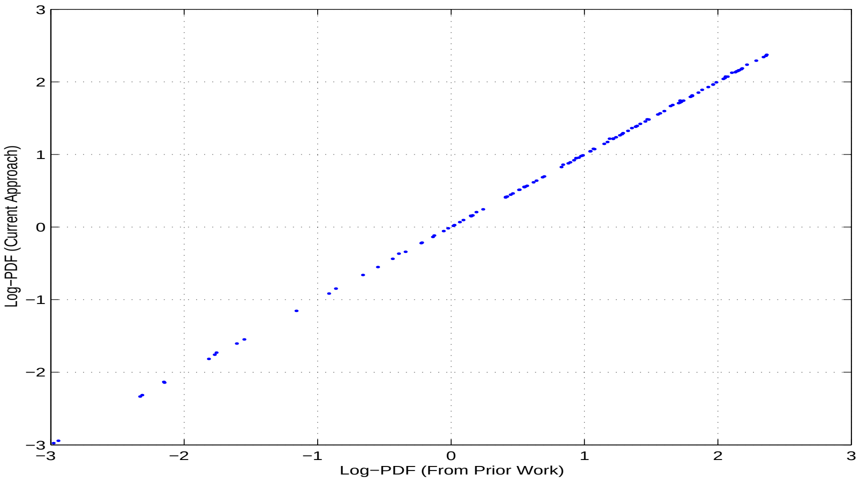

Figure 1: Comparison of Saddlepoint Approximation from previous work (x-axis) with method of this section for cyclic ACF estimates. The largest difference was .021

Example 4 Linear Filtering. *We revisit example 2 and test the SPA solution against a Gaussian mixture approximation. We used a filter length of 3 as shown in equation (12) and a feature size of* $M = 3$. *The filter coefficients were* $\mathbf{a} = [1\ 1\ 1]$. *A Gaussian mixture approximation of* $p(\mathbf{z})$ *under the WGN assumption was obtained using 5000 samples and 20 mixture components. Next, 100 new samples were generated and the log PDF from the SPA was compared with the mixture approximation.*

The results of the experiment are shown in Figure 2. On the graph, the dots represent the method of this section compared with the Gaussian mixture. There is a positive bias of about .5 associated with the SPA. This bias can be attributed to that fact that the statistics are highly non-Gaussian. There are two things to note about this bias. First, the accuracy of the SPA depends only upon the shape of the joint MGF in the vicinity of the saddlepoint, not upon its magnitude. Thus, the errors tend not to increase as we go into the tails of the PDF. This is in constrast to the central-limit theorem (CLT) approximation which tends to have extremely large errors in the tails of the PDF. Second, the error of .5 that we noted means an error of $\exp(.5) = 1.65$, *or a 65 percent error in the PDF value. While this may seems to be large, consider that this will be the error of the projected raw data PDF (using equation 5), which is quite small for a high-dimensional PDF. When working with high-dimensional PDFs, likelihood values vary over extremely wide ranges. PDF accuracy values are better thought of in terms of the error in the log-PDF per dimension.*

To test the hypothesis that this bias is associated with the shape of the MGF in the vicinity of the saddlepoint, we implemented the first-order correction term. The first-order correction term involves the third and fourth-order partial derivatives of the joint CGF and amounts to an additional term in the Taylor series expansion which gives rise to the SPA. This term has been worked out by Nuttall for the the statistics of general form (8) [11]. When the first-order correction term was added, the bias disappeared (circles on the graph). Note that the first-order correction term is often computationally impractical for large M. *Since small bias in the SPA is not significant in the context of equation (5), adding the correction term may not be worth the extra computational effort.*

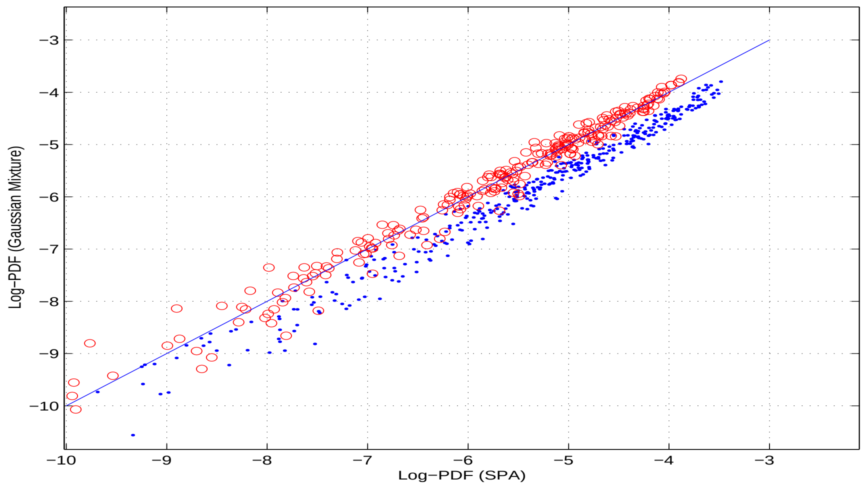

Figure 2: Comparison of Saddlepoint Approximation from reference with a Gaussian mixture approximation. Dots are without first-order correction term, circles with correction term.

3 Order Statistics with Residual Energy.

3.1 Features

Let $\mathbf{x} = [x_1 x_2 \dots x_N]$ be a set of N independent random variables (RVs) distributed under hypothesis H_0 according to the common PDF $p_0(x)$. The joint probability density function (PDF) of $\mathbf{x}$ is

$$p(\mathbf{x}|H_0) = \prod_{i=1}^{N} p_0(x_i).$$

Now let x_i be ordered in decreasing order into the set $\mathbf{y} = [y_1 y_2 \dots y_N]$ where $y_i \geq y_{i+1}$. We now choose a set of $M-1$ indexes $t_1, t_2 \dots t_{M-1}$, with $1 \leq t_1 < t_2 < \cdots t_{M-1} \leq N$ to form a selected collection of order statistics $y_{t_1}, y_{t_2} \dots y_{t_{M-1}}$. To this set, we add the residual "energy",

$$r = \sum_{j \in \mathcal{M}} h(y_j), \tag{15}$$

where the set $\mathcal{M}$ are the integers from 1 to N not including the values $t_1, t_2 \dots t_M$, and $h(x)$ is a function which is needed to insure that r has units of energy. We than form the complete feature vector of length M ($M \geq 2$):

$$\mathbf{z} = [y_{t_1}\; y_{t_2} \dots y_{t_{M-1}}\; r]'.$$

By appending the residual energy to the feature vector, we insure that $\mathbf{z}$ is a sufficient statistic for unknown scale factors applied to $\mathbf{x}$. We consider two important cases:

1. Let $\mathbf{x}$ be a set of magnitude-squared DFT bin outputs, which are exponentially distributed. Since $\mathbf{x}$ is already in units of energy, $h(x) = x$.

2. Let $\mathbf{x}$ be a set of absolute values of zero-mean Gaussian RVs. Then, $h(x) = x^2$.

3.2 Integral Solution

3.2.1 Probability Density Function

Interestingly, the joint PDF of $\mathbf{z}$ can be reduced to a single one-dimensional integral expression [12]. Define

$$c(u, \lambda) = \int_{-\infty}^{u} p_0(x) \ \exp(\lambda h(x)) \ dx \tag{16}$$

$$e(u, \lambda) = \int_{u}^{\infty} p_0(x) \ \exp(\lambda h(x)) \ dx \tag{17}$$

We find the joint PDF of $\mathbf{z}$ to be

$$\begin{aligned} p_z(\mathbf{z}) &= \tfrac{N!}{D\pi} \left\{ \textstyle\prod_{m=1}^{M-1} \ p_0(z_m) \right\} \\ & \int_0^{\infty} \mathrm{Re} \left\{ \exp\left[-(\hat{\lambda} + iy) z_M \right] I(\hat{\lambda} + iy) \right\} \ dy, \end{aligned}$$

where

$$D \triangleq (t_1 - 1)! \left\{ \prod_{m=1}^{M-2} (t_{m+1} - t_m - 1)! \right\} (N - t_{M-1})!,$$

$$\begin{aligned} I(\lambda) &\triangleq e(z_1, \lambda)^{t_1 - 1} \ c(z_{M-1}, \lambda)^{N - t_{M-1}} \\ & \left\{ \textstyle\prod_{m=1}^{M-2} \ [e(z_{m+1}, \lambda) - e(z_m, \lambda)]^{t_{m+1} - t_m - 1} \right\}. \end{aligned}$$

3.2.2 Example 1

Let x_n follow the standard exponential distribution $p_0(x) = \exp(-x)$ for $x > 0$. Let $h(x) = x$. Then, for $\mathrm{Re}(\lambda) < 1$,

$$e(u, \lambda) = \frac{1}{1 - \lambda} \ e^{(\lambda - 1)u} \quad \text{for} \ \ u > 0; \ \frac{1}{1 - \lambda} \ \text{for} \ \ u < 0.$$

$$c(u, \lambda) = \frac{1}{1 - \lambda} \left(1 - e^{(\lambda - 1)u} \right) \quad \text{for} \ \ u > 0; \ 0 \ \text{for} \ \ u < 0.$$

3.2.3 Example 2

Let $x_n = |g_n|$ where g_n follows the standard normal distribution $N(0, 1)$. Let $h(x) = x^2$. Then, for $\mathrm{Re}(\lambda) < \frac{1}{2}$,

$$\begin{aligned} e(u, \lambda) &= \int_u^{\infty} \tfrac{2}{\sqrt{2\pi}} \ \exp(-x^2/2) \ \exp(\lambda x^2) \ dx \\ &= \tfrac{2}{\sqrt{1-2\lambda}} \ \Phi(-u\sqrt{1 - 2\lambda}) \quad \text{for} \ \ u > 0; \\ & \tfrac{1}{\sqrt{1-2\lambda}} \ \text{for} \ \ u < 0. \end{aligned}$$

Also,

$$c(u, \lambda) = \tfrac{1 - 2\Phi(-u\sqrt{1-2\lambda})}{\sqrt{1-2\lambda}} \quad \text{for} \ \ u > 0; \ 0 \ \text{for} \ \ u < 0.$$

3.3 SPA Solution for Exponential RVs

We now consider the first case (exponential RVs) and apply the SPA. This will provide a means of validating the integral solution of the previous section. As we explained in Section 2.3, to obtain the SPA, we need the joint CGF and its first and second-order partial derivatives. Once these are known, the formulas in reference [9] may be used to obtain the SPA.

3.3.1 Joint MGF of y.

We start with the joint MGF of $\mathbf{y}$, which is (see [13], p. 68, eq. B-18)

$$g_0(\alpha_1, \alpha_2 \dots \alpha_N) = \frac{1}{(1-\alpha_1)\left(1-\frac{\alpha_1+\alpha_2}{2}\right)\left(1-\frac{\alpha_1+\alpha_2+\alpha_3}{3}\right)\cdots\left(1-\frac{\alpha_1+\alpha_2+\cdots+\alpha_N}{N}\right)} \tag{18}$$

for $\mathrm{Re}(\alpha_1) < 1$, $\mathrm{Re}(\alpha_1 + \alpha_2) < 2$, ... , $\mathrm{Re}(\alpha_1 + \alpha_2 + \cdots + \alpha_N) < N$. We can rewrite (18) as

$$g_0(\alpha_1, \alpha_2 \dots \alpha_N) = \frac{1}{\prod_{n=1}^{N} \phi_n(\alpha_1, \alpha_2 \dots \alpha_N)},$$

where

$$\phi_n(\alpha_1, \alpha_2 \dots \alpha_N) = 1 - \frac{1}{n}\sum_{p=1}^{n} \alpha_p, \quad 1 \le n \le N.$$

Alternatively,

$$\phi_n(\alpha_1, \alpha_2 \dots \alpha_N) = 1 - \sum_{p=1}^{N} q_{np}\, \alpha_p, \quad 1 \le n \le N,$$

where

$$q_{np} = \left\{ \begin{array}{ll} \frac{1}{n} & \text{for} \quad 1 \le p \le n \\ & \\ 0 & \text{for} \quad n < p \le N \end{array} \right\}, n = 1, 2 \dots N.$$

Define the N-by-N matrix $\mathbf{Q} = [q_{np}]$ and $\boldsymbol{\alpha} = [\alpha_1\ \alpha_2 \cdots \alpha_N]'$. Then,

$$\boldsymbol{\phi}(\boldsymbol{\alpha}) \triangleq \begin{bmatrix} \phi_1(\boldsymbol{\alpha}) \\ \phi_2(\boldsymbol{\alpha}) \\ \vdots \\ \phi_N(\boldsymbol{\alpha}) \end{bmatrix} = \underline{\mathbf{1}} - \mathbf{Q}\boldsymbol{\alpha},$$

where $\underline{\mathbf{1}}$ is an N-by-1 column vector of ones. Thus, $g_0(\boldsymbol{\alpha})$ is the reciprocal of the product of the elements of $\boldsymbol{\phi}(\boldsymbol{\alpha})$, denoted by

$$g_0(\boldsymbol{\alpha}) = \frac{1}{\mathrm{prod}(\underline{\mathbf{1}} - \mathbf{Q}\boldsymbol{\alpha})}, \tag{19}$$

where prod( ) is the product of the elements of the argument.

3.3.2 Joint MGF of z.

The joint MGF of $\mathbf{z}$ is, for $\boldsymbol{\lambda} = [\lambda_1 \lambda_2 \ldots \lambda_M]'$,

$$\begin{aligned} g_z(\boldsymbol{\lambda}) &\triangleq \mathbf{E}\left\{\exp(\boldsymbol{\lambda}'\mathbf{z})\right\} \\ &= \mathbf{E}\left\{\exp(\lambda_1 y_{t_1} + \lambda_2 y_{t_2} + \right. \\ &\qquad \left. \cdots + \lambda_{M-1} y_{t_{M-1}} + \lambda_M r)\right\}. \end{aligned} \tag{20}$$

This can be written

$$g_z(\boldsymbol{\lambda}) = g_0(\mathbf{A}\boldsymbol{\lambda})$$

where $\mathbf{A}$ is the N-by-M matrix that has 1's everywhere in the M-th column except for 0's in rows $t_1, t_2 \ldots t_{M-1}$, and $\mathbf{A}$ has 1's in row t_1, column 1; row t_2, column 2; etc. Therefore, from (19),

$$g_z(\boldsymbol{\lambda}) = \frac{1}{\text{prod}(\underline{\mathbf{1}} - \mathbf{Q}\mathbf{A}\boldsymbol{\lambda})}.$$

Note that $\mathbf{Q}$ can be a large N-by-N matrix if N is large. However, if we define

$$\mathbf{P} \triangleq \mathbf{QA},$$

$\mathbf{P}$ is a reasonable N-by-M size matrix and can be easily formed directly. The final simplified form for the joint MGF is

$$g_z(\boldsymbol{\lambda}) = \frac{1}{\text{prod}(\underline{\mathbf{1}} - \mathbf{P}\boldsymbol{\lambda})}.$$

3.3.3 Partial Derivatives of the CGF.

To obtain the SPA to the PDF of $\mathbf{z}$, we need the joint cumulant generating function (CGF) $c_z(\boldsymbol{\lambda})$ of $\mathbf{z}$ and its partial derivatives. The joint CGF is defined by

$$c_z(\boldsymbol{\lambda}) = \log(g_z(\boldsymbol{\lambda})) = -\text{sum}\left(\mathbf{log}(\underline{\mathbf{1}} - \mathbf{P}\boldsymbol{\lambda})\right)$$

where $\mathbf{log}$ is the vector log function which operates on each element of its argument and sum( ) is the vector sum, which adds up all the elements of the argument. If we define $\boldsymbol{\phi}^{-1}(\boldsymbol{\lambda})$ as the element-by-element reciprocal of $\underline{\mathbf{1}} - \mathbf{P}\boldsymbol{\lambda}$, and $\boldsymbol{\Phi}(\boldsymbol{\lambda})$ as the diagonal N-by-N matrix with elements equal to the elements of $\underline{\mathbf{1}} - \mathbf{P}\boldsymbol{\lambda}$, it is straight-forward to show that the gradient vector of $c_z(\boldsymbol{\lambda})$ is the M-by-1 vector

$$\boldsymbol{\delta}(\boldsymbol{\lambda}) \triangleq \frac{\partial}{\partial \boldsymbol{\lambda}}\, c_z(\boldsymbol{\lambda}) = \mathbf{P}'\, \boldsymbol{\phi}^{-1}(\boldsymbol{\lambda}),$$

and the M-by-M Hessian matrix of $c_z(\boldsymbol{\lambda})$ is

$$\mathbf{C}_z(\boldsymbol{\lambda}) \triangleq \frac{\partial^2}{\partial \boldsymbol{\lambda} \partial \boldsymbol{\lambda}'}\, c_z(\boldsymbol{\lambda}) = \mathbf{P}'\, \boldsymbol{\Phi}^{-2}(\boldsymbol{\lambda})\, \mathbf{P},$$

3.4 Comparison

A good check on the analysis is to compare the two methods just presented in Sections 3.3 and 3.2. We tried the case of $N = 100$, $M = 3$, $t_1 = 3$, $t_2 = 7$. Samples of input data vector $\mathbf{x}$ were generated using N independent uniformly distributed RVs (in the range 0 to 1), then scaled by a random scale factor in the range 0 to 10. Note that the PDF used for data generation is not important because we are comparing two PDF approximations for the same input feature vector. The PDF of $\mathbf{z}$ was computed using the methods of Sections 3.3 and 3.2. The results are shown in Figure 3. The two methods agreed very closely - within a maximum error of .059 in log space.

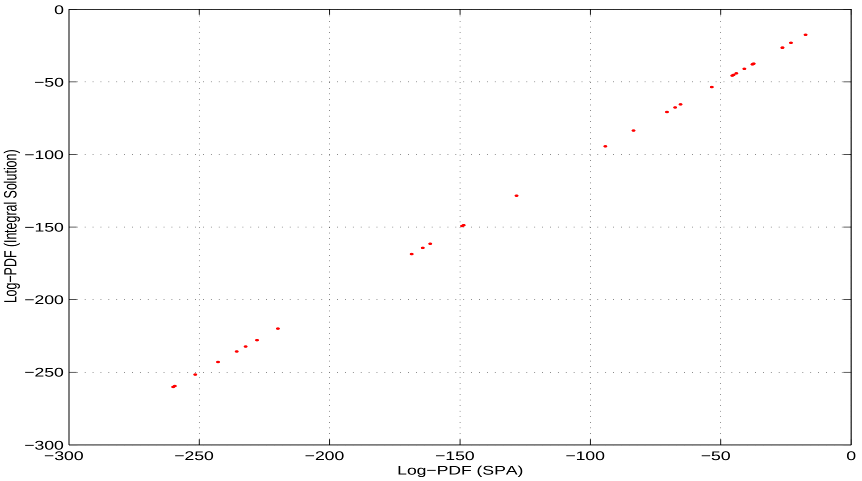

Figure 3: Comparison of Saddlepoint Approximation (Section 3.3) with integral solution (Section 3.2). Largest difference was .059

4 Conclusions

In this paper, we have derived saddlepoint approximations to the multidimensional PDFs for two general classes of features. The availability of these PDFs permits these features to be used in a class-specific classifier. The PDFs were checked using numerical simulations.

References

[1] A. Mashoshin, “Synthesis of algorithms for the classification of underwater objects from their underwater sound field,” *Acoustical Physics*, vol. 42, no. 3, pp. 347–351, 1996.

[2] S. Kumar, J. Ghosh, and M. Crawford, “A versatile framework for labeling imagery with large number of classes,” in *Proceedings of the International Joint Conference on Neural Networks*, (Washington, D.C.), 1999.

[3] S. Kumar, J. Ghosh, and M. Crawford, “A hierarchical multiclassifier system for hyperspectral data analysis,” in *Multiple Classifier Systems* (J. Kittler and F. Roli, eds.), pp. 270–279, Springer, 2000.

[4] H. Watanabe, T. Yamaguchi, and S. Katagiri, “Discriminative metric design for robust pattern recognition,” *IEEE Trans. Signal Processing*, vol. 45, no. 11, pp. 2655–2661, 1997.

[5] P. Belhumeur, J. Hespanha, and D. Kriegman, “Eigenfaces vs. fisherfaces: Recognition using class specific linear projection,” *PAMI*, vol. 19, pp. 711–720, July 1997.

[6] D. Sebald, “Support vector machines and the multiple hypothesis test problem,” *IEEE Trans. Signal Processing*, vol. 49, pp. 2865–2872, November 2001.

[7] P. M. Baggenstoss, “A theoretically optimum approach to classification using class-specific features.,” *Proceedings of ICPR, Barcelona*, 2000.

[8] P. M. Baggenstoss, "A modified Baum-Welch algorithm for hidden Markov models with multiple observation spaces.," *IEEE Trans. Speech and Audio*, pp. 411–416, May 2001.

[9] S. M. Kay, A. H. Nuttall, and P. M. Baggenstoss, "Multidimensional probability density function approximation for detection, classification and model order selection," *IEEE Trans. Signal Processing*, Oct 2001.

[10] O. E. Barndorff-Nielsen and D. R. Cox, *Asymptotic Techniques for Use in Statistics.* Chapman and Hall, 1989.

[11] A. H. Nuttall, "Saddlepoint approximation and first-order correction term to the joint probability density function of M quadratic and linear forms in K Gaussian random variables with arbitrary means and covariances," *NUWC Technical Report 11262*, December 2000.

[12] A. H. Nuttall, "An integral solution for the joint PDF of order statistics and the residual sum," *NUWC Technical Report (to be published)*, October 2001.

[13] A. H. Nuttall, "Detection performance of generalized likelihood ratio processors for random signals of unknown location, structure, extent, and strength," *NUWC Technical Report 10739*, August 1994.