Asymptotically Optimal Approximation of Multidimensional PDFs by Lower Dimensional PDFs

Steven Kay*
Dept. of Electrical and Computer Engineering
University of Rhode Island
Kingston, RI 02881
401-874-5804 (voice) 401-782-6422 (fax)
kay@ele.uri.edu

EDICS: SSP - SSAN

May 11, 2006

Abstract

Probability density functions of high dimensionality are impractical to estimate from real data. For accurate estimation the dimensionality of the PDF can be at most 5–10. In order to reduce the dimensionality a sufficient statistic is usually employed. When none is available, there is no universal agreement on how to proceed. We show how to construct a high dimension probability density function based on the probability density function of a low dimensional statistic that is closest to the true one in the sense of divergence. The latter criterion asymptotically minimizes the probability of error in a decision rule. An application to feature selection for classification is described.

1 Introduction

In many problems of practical interest it is necessary to estimate a multidimensional probability density function (PDF). One important application is to pattern recognition or classification [1]. Typically, when faced with training data from several known classes, a multidimensional PDF is estimated. As the dimensionality increases, however, we require more training data for good PDF estimates. The requirement for training data cannot always be met. For example, in space-time adaptive radar this additional training

*This work was supported by the Office of Naval Research under contract N66604-02-14748.

data is obtained from adjacent range cells. The net result is a sharp decrease in detection performance when the range cells are inhomogeneous [12]. In some cases it is possible to reduce the dimensionality of the PDF by appealing to the theory of sufficient statistics [2]. Then, only the PDF of the sufficient statistic need be estimated.

Recently, a new approach along these lines has been proposed for PDFs that do not admit sufficient statistics [3]. Based on the *PDF projection theorem*, a multidimensional PDF may be constructed from a lower dimensional PDF. For PDFs that admit sufficient statistics this method produces the true PDF, a result that is well known [4]. For PDFs that do not admit sufficient statistics, which is usually the case in practice, the constructed or projected PDF is only one of an infinite number of possible PDFs. From the projection theorem there is no indication as to how good the projected PDF is as an approximation to the true one. In this paper, *we prove that the projected PDF as described in [3] is in fact an optimal one.* The criterion of optimality is the distance of the projected PDF from the true one. It is proven that this "distance" is minimized in the Kullback-Liebler (K-L) or information divergence sense. It can be shown that asymptotically, i.e., as the data record length becomes large, minimization of the K-L distance is equivalent to minimization of the probability or error in any statistical decision problem [8,13] – hence our choice of the word "optimal". As a concrete application of this theorem we show how to compare potential features for use in classification.

Section 2 reviews the projection theorem while Section 3 gives the main orthogonal projection theorem of the paper. Some examples of the orthogonal projection are described in Sections 4 and 5. Finally, an application of the theory to feature selection for classification is given in Section 6.

2 Review of the Projection Theorem

We first describe the construction of a PDF based on the PDF of its sufficient statistic (a slight generalization of that presented in [4]) and then, briefly summarize the projection theorem as described in [3]. Consider a family of N-dimensional PDFs indexed by a parameter vector $\boldsymbol{\theta}$, where $\boldsymbol{\theta}$ is $p \times 1$ and $p < N$. The restriction of the number of parameters p to being less than N, the number of data points (and in practice usually much less than N) is to allow the unknown parameters to be estimated with reasonable accuracy. Denoting the data vector by $\mathbf{x} = [x[0]\, x[1] \ldots x[N-1]]^T$, the N-dimensional PDF is given by $p_X(\mathbf{x};\boldsymbol{\theta})$. If a minimal sufficient statistic exists for $\boldsymbol{\theta}$, then by the Neyman-Pearson factorization theorem [5], we can write the PDF as

$$p_X(\mathbf{x};\boldsymbol{\theta}) = g(\mathbf{T}(\mathbf{x}),\boldsymbol{\theta})h(\mathbf{x}) \tag{1}$$

for g a nonnegative function and $\mathbf{T}(\mathbf{x})$ a p-dimensional function of $\mathbf{x}$. This can be rewritten in normalized form as

$$p_X(\mathbf{x};\boldsymbol{\theta}) = \frac{g(\mathbf{T}(\mathbf{x}),\boldsymbol{\theta})}{\int_{S_t} g(\mathbf{u},\boldsymbol{\theta})|\mathbf{J}|d\mathbf{u}} h(\mathbf{x}) \int_{S_t} g(\mathbf{u},\boldsymbol{\theta})|\mathbf{J}|d\mathbf{u} \tag{2}$$

where

$$S_t = \{\mathbf{x} : \mathbf{T}(\mathbf{x}) = \mathbf{t} \text{ for } \mathbf{x} \in R^N\}$$

and $|\mathbf{J}|d\mathbf{u}$ is the differential volume element for S_t. If the normalization factor (the integral in (2)) is independent of $\boldsymbol{\theta}$, we have that upon denoting the integral by $I(\mathbf{t})$

$$p_X(\mathbf{x};\boldsymbol{\theta}) = p_T(\mathbf{t}(\mathbf{x});\boldsymbol{\theta})h(\mathbf{x})I(\mathbf{t}).$$

Finally the likelihood ratios based on the data and the sufficient statistic are equal [4] or

$$\frac{p_X(\mathbf{x};\boldsymbol{\theta})}{p_X(\mathbf{x};\boldsymbol{\theta}_0)} = \frac{p_T(\mathbf{t}(\mathbf{x});\boldsymbol{\theta})}{p_T(\mathbf{t}(\mathbf{x});\boldsymbol{\theta}_0)}.$$

As a result the N-dimensional PDF can be written as

$$p_X(\mathbf{x};\boldsymbol{\theta}) = \frac{p_T(\mathbf{t}(\mathbf{x});\boldsymbol{\theta})}{p_T(\mathbf{t}(\mathbf{x});\boldsymbol{\theta}_0)} p_X(\mathbf{x};\boldsymbol{\theta}_0). \tag{3}$$

The PDF $p_X(\mathbf{x};\boldsymbol{\theta}_0)$ is termed the *reference PDF* since it acts as an "origin" in the manifold of PDFs whose coordinates are given by the values of $\boldsymbol{\theta}$ [6]. More specifically, we can define a manifold $\mathcal{M}$ of log-PDFs by

$$\mathcal{M} = \{\ln p(\mathbf{x};\boldsymbol{\theta}) : \boldsymbol{\theta} \in R^p\}.$$

Choosing $\ln p(\mathbf{x};\boldsymbol{\theta}_0)$ to be the origin, any "vector" in the manifold may be accessed by adding the log-likelihood ratio or from (3)

$$\ln p_X(\mathbf{x};\boldsymbol{\theta}) = \ln p_X(\mathbf{x};\boldsymbol{\theta}_0) + \ln \frac{p_T(\mathbf{t}(\mathbf{x});\boldsymbol{\theta})}{p_T(\mathbf{t}(\mathbf{x});\boldsymbol{\theta}_0)}.$$

This important result can be restated by saying that any vector is reached by adding the log-likelihood ratio of the *sufficient statistic only*. In practice, this means that we need only determine the PDF of the sufficient statistic. Also, for (3) to be useful in practice we must be able to determine the origin or $p_X(\mathbf{x};\boldsymbol{\theta}_0)$. This is usually possible by analytical evaluation or by using a saddlepoint approximation [7].

When a minimal sufficient statistic does not exist, which is usually the case in a real-world data situation, one is tempted to use some other statistic which we denote as $\mathbf{Z}(\mathbf{x})$ in place of $\mathbf{T}(\mathbf{x})$ in (3). By doing so we will have constructed a PDF as

$$\hat{p}_X(\mathbf{x};\boldsymbol{\theta}) = \frac{p_Z(\mathbf{z}(\mathbf{x});\boldsymbol{\theta})}{p_Z(\mathbf{z}(\mathbf{x});\boldsymbol{\theta}_0)} p_X(\mathbf{x};\boldsymbol{\theta}_0) \tag{4}$$

which is clearly only an approximation to the true PDF. The constructed PDF $\hat{p}_X(\mathbf{x};\boldsymbol{\theta})$ is termed the *projected PDF* since it has been proven that $\hat{p}_X(\mathbf{x};\boldsymbol{\theta})$ is a valid PDF in that it integrates to one for all $\boldsymbol{\theta}$ [3]. In fact, one can construct an infinite number of PDFs based on $\mathbf{Z}(\mathbf{x})$ of which (4) is but one of these. The others can be found by specifying the conditional PDF $p_{X|Z}$ (defined on a surface) so that $p_X = \int p_{X|Z} p_Z dZ$. However, we prove in the next section that *the projected PDF given by (4) is the closest PDF to the true PDF using the Kullback-Liebler definition of distance.*

3 Orthogonal Projections in PDF Space

We now prove that (4) is an optimal projection and hence we will refer to it as the *orthogonal PDF projection.*

Theorem 3.1 (Orthogonal PDF projection theorem) *Consider a family of PDFs whose likelihood ratio relative to a given reference PDF $p_X(\mathbf{x};\boldsymbol{\theta}_0)$ depends only on the data via the statistic $\mathbf{Z}(\mathbf{x})$. The dimensionality of $\mathbf{Z}(\mathbf{x})$ is arbitrary. Then the family of PDFs can be written as*

$$\hat{p}_X(\mathbf{x};\boldsymbol{\theta}) = \frac{\exp[f(\mathbf{z}(\mathbf{x}))]p(\mathbf{x};\boldsymbol{\theta}_0)}{\int \exp[f(\mathbf{z}(\mathbf{x}'))]p_X(\mathbf{x}';\boldsymbol{\theta}_0)d\mathbf{x}'}$$

where $f(\mathbf{z})$ is an arbitrary function. The PDF in the set that minimizes the distance in the Kullback-Leibler sense to a given PDF $p_X(\mathbf{x};\boldsymbol{\theta}_1)$ is

$$p_X^*(\mathbf{x}) = \frac{p_Z(\mathbf{z}(\mathbf{x});\boldsymbol{\theta}_1)}{p_Z(\mathbf{z}(\mathbf{x});\boldsymbol{\theta}_0)} p_X(\mathbf{x};\boldsymbol{\theta}_0).$$

Furthermore, the minimum distance is given by $D(p_X(\mathbf{x};\boldsymbol{\theta}_1)||p_X^*(\mathbf{x})) = D_X(1||0) - D_Z(1||0)$, where D_X is the K-L distance between $p_X(\mathbf{x};\boldsymbol{\theta}_1)$ and $p_X(\mathbf{x};\boldsymbol{\theta}_0)$ and D_Z is the K-L distance between $p_Z(\mathbf{z};\boldsymbol{\theta}_1)$ and $p_Z(\mathbf{z};\boldsymbol{\theta}_0)$. Finally, this minimum distance is zero if and only if $\mathbf{Z}(\mathbf{x})$ is a sufficient statistic (not necessarily minimal) for the PDF family $p_X(\mathbf{x};\boldsymbol{\theta})$, in which case $p_X^*(\mathbf{x}) = p_X(\mathbf{x};\boldsymbol{\theta}_1)$.

Proof:

$$\begin{aligned} D(p_X(\mathbf{x};\boldsymbol{\theta}_1)||\hat{p}_X(\mathbf{x};\boldsymbol{\theta}_1)) &= \int p_X(\mathbf{x};\boldsymbol{\theta}_1) \ln \frac{p_X(\mathbf{x};\boldsymbol{\theta}_1)}{\hat{p}_X(\mathbf{x};\boldsymbol{\theta}_1)} d\mathbf{x} \\ &= \int p_X(\mathbf{x};\boldsymbol{\theta}_1) \ln \left[\frac{p_X(\mathbf{x};\boldsymbol{\theta}_1)}{p_X^*(\mathbf{x})} \frac{p_X^*(\mathbf{x})}{\hat{p}_X(\mathbf{x};\boldsymbol{\theta}_1)} \right] d\mathbf{x} \\ &= D(p_X(\mathbf{x};\boldsymbol{\theta}_1)||p_X^*(\mathbf{x})) + \underbrace{\int p_X(\mathbf{x};\boldsymbol{\theta}_1) \ln \frac{p_X^*(\mathbf{x})}{\hat{p}_X(\mathbf{x};\boldsymbol{\theta}_1)} d\mathbf{x}}_{\xi(f)}. \end{aligned}$$

Now replace $p_X^*(\mathbf{x})$ and $\hat{p}_X(\mathbf{x};\boldsymbol{\theta}_1)$ by their definitions to yield

$$\xi(f) = \int p_X(\mathbf{x};\boldsymbol{\theta}_1) \ln \left[\frac{\frac{p_Z(\mathbf{z}(\mathbf{x});\boldsymbol{\theta}_1)}{p_Z(\mathbf{z}(\mathbf{x});\boldsymbol{\theta}_0)} p_X(\mathbf{x};\boldsymbol{\theta}_0)}{\frac{\exp[f(\mathbf{z}(\mathbf{x}))]p_X(\mathbf{x};\boldsymbol{\theta}_0)}{\int \exp[f(\mathbf{z}(\mathbf{x}'))]p_X(\mathbf{x}';\boldsymbol{\theta}_0)d\mathbf{x}'}} \right] d\mathbf{x}$$

$$= \int p_X(\mathbf{x};\boldsymbol{\theta}_1) \ln \left[\frac{p_Z(\mathbf{z}(\mathbf{x});\boldsymbol{\theta}_1)}{\frac{\exp[f(\mathbf{z}(\mathbf{x}))]p_Z(\mathbf{z}(\mathbf{x});\boldsymbol{\theta}_0)}{\int \exp[f(\mathbf{z}(\mathbf{x}'))]p_X(\mathbf{x}';\boldsymbol{\theta}_0)d\mathbf{x}'}} \right] d\mathbf{x}.$$

But using a transformation of variables from $\mathbf{x}$ to $\mathbf{z}$ we have that

$$\int \exp[f(\mathbf{z}(\mathbf{x}'))]p_X(\mathbf{x}';\boldsymbol{\theta}_0)d\mathbf{x}' = \int \exp[f(\mathbf{z})]p_Z(\mathbf{z};\boldsymbol{\theta}_0)d\mathbf{z} \tag{5}$$

and then using the same transformation of variables in the main integral for $\xi(f)$ we have that

$$\begin{aligned} \xi(f) &= \int p_Z(\mathbf{z};\boldsymbol{\theta}_1) \ln \left[\frac{p_Z(\mathbf{z};\boldsymbol{\theta}_1)}{\frac{\exp[f(\mathbf{z})]p_Z(\mathbf{z};\boldsymbol{\theta}_0)}{\int \exp[f(\mathbf{z}')]p_Z(\mathbf{z}';\boldsymbol{\theta}_0)d\mathbf{z}'}} \right] d\mathbf{z} \\ &= D(p_Z(\mathbf{z};\boldsymbol{\theta}_1)||\bar{p}(\mathbf{z})) \geq 0 \end{aligned}$$

where

$$\bar{p}(\mathbf{z}) = \frac{\exp[f(\mathbf{z})]p_Z(\mathbf{z};\boldsymbol{\theta}_0)}{\int \exp[f(\mathbf{z}')]p_Z(\mathbf{z}';\boldsymbol{\theta}_0)d\mathbf{z}'}.$$

Now $\xi(f)$ will be zero if and only if $\bar{p}(\mathbf{z}) = p_Z(\mathbf{z};\boldsymbol{\theta}_1)$ in which case we have that

$$\begin{aligned} p_Z(\mathbf{z};\boldsymbol{\theta}_1) &= \frac{\exp[f(\mathbf{z})]p_Z(\mathbf{z};\boldsymbol{\theta}_0)}{\int \exp[f(\mathbf{z}')]p_Z(\mathbf{z}';\boldsymbol{\theta}_0)d\mathbf{z}'} \\ &= \frac{1}{c}\exp[f(\mathbf{z})]p_Z(\mathbf{z};\boldsymbol{\theta}_0) \end{aligned}$$

where c is the value of the integral in the denominator, which is a constant. Thus,

$$f(\mathbf{z}) = \ln c + \ln \frac{p_Z(\mathbf{z};\boldsymbol{\theta}_1)}{p_Z(\mathbf{z};\boldsymbol{\theta}_0)}.$$

Since

$$\hat{p}_X(\mathbf{x};\boldsymbol{\theta}_1) = \frac{\exp[f(\mathbf{z}(\mathbf{x}))]p_X(\mathbf{x};\boldsymbol{\theta}_0)}{\int \exp[f(\mathbf{z}(\mathbf{x}'))]p_X(\mathbf{x}';\boldsymbol{\theta}_0)d\mathbf{x}'}$$

we have upon substitution of

$$f(\mathbf{z}(\mathbf{x})) = \ln c + \ln \frac{p_Z(\mathbf{z}(\mathbf{x});\boldsymbol{\theta}_1)}{p_Z(\mathbf{z}(\mathbf{x});\boldsymbol{\theta}_0)}$$

that

$$\hat{p}_X(\mathbf{x};\boldsymbol{\theta}_1) = \frac{c\frac{p_Z(\mathbf{z}(\mathbf{x});\boldsymbol{\theta}_1)}{p_Z(\mathbf{z}(\mathbf{x});\boldsymbol{\theta}_0)}p_X(\mathbf{x};\boldsymbol{\theta}_0)}{\int \exp[f(\mathbf{z}(\mathbf{x}'))]p_X(\mathbf{x}';\boldsymbol{\theta}_0)d\mathbf{x}'}.$$

But from (5) we see that the denominator in the above expression is just c, so that finally we have

$$p_X^*(\mathbf{x}) = \hat{p}_X(\mathbf{x};\boldsymbol{\theta}_1) = \frac{p_Z(\mathbf{z}(\mathbf{x});\boldsymbol{\theta}_1)}{p_Z(\mathbf{z}(\mathbf{x});\boldsymbol{\theta}_0)}p_X(\mathbf{x};\boldsymbol{\theta}_0).$$

Hence,

$$D(p_X(\mathbf{x};\boldsymbol{\theta}_1)||\hat{p}_X(\mathbf{x};\boldsymbol{\theta}_1)) \geq D(p_X(\mathbf{x};\boldsymbol{\theta}_1)||p_X^*(\mathbf{x}))$$

with equality if and only if $\hat{p}_X(\mathbf{x};\boldsymbol{\theta}_1) = p_X^*(\mathbf{x})$.

Also, we have as the minimum K-L distance

$$\begin{aligned} D(p_X(\mathbf{x};\boldsymbol{\theta}_1)||p_X^*(\mathbf{x})) &= \int p_X(\mathbf{x};\boldsymbol{\theta}_1) \ln \left[\frac{p_X(\mathbf{x};\boldsymbol{\theta}_1)}{\frac{p_Z(\mathbf{z}(\mathbf{x});\boldsymbol{\theta}_1)}{p_Z(\mathbf{z}(\mathbf{x});\boldsymbol{\theta}_0)} p_X(\mathbf{x};\boldsymbol{\theta}_0)} \right] d\mathbf{x} \\ &= \int p_X(\mathbf{x};\boldsymbol{\theta}_1) \ln \left[\frac{p_X(\mathbf{x};\boldsymbol{\theta}_1)}{p_X(\mathbf{x};\boldsymbol{\theta}_0)} \right] d\mathbf{x} - \int p_X(\mathbf{x};\boldsymbol{\theta}_1) \ln \left[\frac{p_Z(\mathbf{z}(\mathbf{x});\boldsymbol{\theta}_1)}{p_Z(\mathbf{z}(\mathbf{x});\boldsymbol{\theta}_0)} \right] d\mathbf{x} \\ &= D_X(1||0) - \int p_Z(\mathbf{z};\boldsymbol{\theta}_1) \ln \left[\frac{p_Z(\mathbf{z};\boldsymbol{\theta}_1)}{p_Z(\mathbf{z};\boldsymbol{\theta}_0)} \right] d\mathbf{z} \\ &= D_X(1||0) - D_Z(1||0) \geq 0 \end{aligned}$$

with equality if and only if $\mathbf{Z}(\mathbf{x})$ is a sufficient statistic (standard result of Kullback [9], page 19). If $D(p_X(\mathbf{x};\boldsymbol{\theta}_1)||p_X^*(\mathbf{x})) = 0$, then it follows that (standard result of Kullback [9], page 20)

$$p_X^*(\mathbf{x}) = p_X(\mathbf{x};\boldsymbol{\theta}_1) = \frac{p_Z(\mathbf{z}(\mathbf{x});\boldsymbol{\theta}_1)}{p_Z(\mathbf{z}(\mathbf{x});\boldsymbol{\theta}_0)} p_X(\mathbf{x};\boldsymbol{\theta}_0)$$

which concludes the proof.

4 Example of an Orthogonal PDF Projection - Minimal Sufficient Statistic Exists

As a simple example of the theorem consider the linear model defined by $\mathbf{x} = \mathbf{H}\boldsymbol{\theta} + \mathbf{w}$. Here $\mathbf{H}$ is a known matrix of dimension $N \times p$ with $p < N$, $\boldsymbol{\theta}$ is a parameter vector of dimension $p \times 1$, and $\mathbf{w}$ is a $N \times 1$ noise vector distributed as $\mathcal{N}(\mathbf{0}, \sigma^2\mathbf{I})$. Consider a statistic of dimension $p \times 1$ which is a linear function of the data vector. As such it can be written as $\mathbf{Z}(\mathbf{x}) = \mathbf{A}^T\mathbf{x}$, where $\mathbf{A}$ is $N \times p$. It is well known that for the linear model, the statistic $\mathbf{H}^T\mathbf{x}$ is a minimal sufficient statistic for $\boldsymbol{\theta}$. To show that this agrees with the theorem, we construct the orthogonal projection PDF and then show that the divergence is minimized for $\mathbf{A} = \mathbf{H}$. Furthermore, the divergence in this case will be zero. The reference PDF is chosen as $p_X(\mathbf{x};\boldsymbol{\theta}_0) = p_X(\mathbf{x};\mathbf{0})$. Although the value of $\boldsymbol{\theta}$ chosen for the reference PDF is arbitrary, we usually use a value that simplifies the PDF. This is because in practice we will need to specify this high dimensional PDF (since it is a PDF on $\mathbf{x}$, which is $N \times 1$) analytically. Thus, noting that $\mathbf{Z}(\mathbf{x}) \sim \mathcal{N}(\mathbf{A}^T\mathbf{H}\boldsymbol{\theta}, \sigma^2\mathbf{A}^T\mathbf{A})$, we have from (4) that

$$\begin{aligned} \hat{p}_X(\mathbf{x};\boldsymbol{\theta}) &= \frac{p_Z(\mathbf{z}(\mathbf{x});\boldsymbol{\theta})}{p_Z(\mathbf{z}(\mathbf{x});\mathbf{0})} p_X(\mathbf{x};\mathbf{0}) \\ &= \exp\left[-\frac{1}{2}(\mathbf{z}(\mathbf{x}) - \mathbf{A}^T\mathbf{H}\boldsymbol{\theta})^T[\sigma^2(\mathbf{A}^T\mathbf{A})]^{-1}(\mathbf{z}(\mathbf{x}) - \mathbf{A}^T\mathbf{H}\boldsymbol{\theta}) + \frac{1}{2}\mathbf{z}(\mathbf{x})^T[\sigma^2(\mathbf{A}^T\mathbf{A})]^{-1}\mathbf{z}(\mathbf{x}) \right] \\ &\quad \cdot \frac{1}{(2\pi\sigma^2)^{N/2}} \exp\left[-\frac{1}{2\sigma^2}\mathbf{x}^T\mathbf{x} \right] \\ &= \frac{1}{(2\pi\sigma^2)^{N/2}} \exp\left[-\frac{1}{2\sigma^2} Q \right] \end{aligned}$$

where

$$\begin{aligned} Q &= (\mathbf{z}(\mathbf{x}) - \mathbf{A}^T\mathbf{H}\boldsymbol{\theta})^T(\mathbf{A}^T\mathbf{A})^{-1}(\mathbf{z}(\mathbf{x}) - \mathbf{A}^T\mathbf{H}\boldsymbol{\theta}) - \mathbf{z}(\mathbf{x})^T(\mathbf{A}^T\mathbf{A})^{-1}\mathbf{z}(\mathbf{x}) + \mathbf{x}^T\mathbf{x} \\ &= (\mathbf{A}^T\mathbf{x} - \mathbf{A}^T\mathbf{H}\boldsymbol{\theta})^T(\mathbf{A}^T\mathbf{A})^{-1}(\mathbf{A}^T\mathbf{x} - \mathbf{A}^T\mathbf{H}\boldsymbol{\theta}) - \mathbf{x}^T\mathbf{A}(\mathbf{A}^T\mathbf{A})^{-1}\mathbf{A}^T\mathbf{x} + \mathbf{x}^T\mathbf{x}. \end{aligned}$$

After some simplification this can be written as

$$Q = (\mathbf{x} - \mathbf{P}_A\mathbf{H}\boldsymbol{\theta})^T(\mathbf{x} - \mathbf{P}_A\mathbf{H}\boldsymbol{\theta})$$

where $\mathbf{P}_A = \mathbf{A}(\mathbf{A}^T\mathbf{A})^{-1}\mathbf{A}^T$ is the $N \times N$ orthogonal projection matrix which projects an $N \times 1$ vector onto the columns of $\mathbf{A}$. As a result we have that the orthogonal projection PDF is

$$\hat{p}_X(\mathbf{x};\boldsymbol{\theta}) = \frac{1}{(2\pi\sigma^2)^{N/2}} \exp\left[-\frac{1}{2\sigma^2}(\mathbf{x} - \mathbf{P}_A\mathbf{H}\boldsymbol{\theta})^T(\mathbf{x} - \mathbf{P}_A\mathbf{H}\boldsymbol{\theta})\right]. \tag{6}$$

The divergence between (6) and the true PDF is easily shown to be (see [9], page 189)

$$D(p_X(\mathbf{x};\boldsymbol{\theta})||\hat{p}_X(\mathbf{x};\boldsymbol{\theta})) = \frac{||\mathbf{H}\boldsymbol{\theta} - \mathbf{P}_A\mathbf{H}\boldsymbol{\theta}||^2}{2\sigma^2}$$

where $||\cdot||$ denotes the Euclidean norm. Since $D \geq 0$, the divergence is minimized for all $\boldsymbol{\theta}$ if $\mathbf{P}_A\mathbf{H} = \mathbf{H}$. This will be satisfied if $\mathbf{A} = \mathbf{H}$ or equivalently if $\mathbf{Z}(\mathbf{x}) = \mathbf{H}^T\mathbf{x}$. As expected the divergence is zero since the statistic chosen is sufficient (and in this case also minimal).

5 Example of an Orthogonal PDF Projection - No Minimal Sufficient Statistic Exists

We now give an example of an orthogonal projection for a PDF family for which no minimal sufficient statistic exists. This will illustrate the key ideas. However, this example assumes independent and identically distributed (IID) random variables, a situation in which the joint PDF is easily estimated in practice. Hence, the following is to be regarded as only a mathematically tractable and illustrative example. If $\mathbf{x}$ consists of N independent and identically distributed (IID) Laplacian random variables with unknown mean A and known variance σ^2, then the joint PDF is

$$\begin{aligned} p_X(\mathbf{x};A) &= \prod_{n=0}^{N-1} \frac{1}{\sqrt{2\sigma^2}} \exp\left[-\sqrt{\frac{2}{\sigma^2}}|x[n] - A|\right] \\ &= \left(\frac{1}{\sqrt{2\sigma^2}}\right)^N \exp\left[-\sqrt{\frac{2}{\sigma^2}}\sum_{n=0}^{N-1}|x[n] - A|\right]. \end{aligned} \tag{7}$$

In this case there is no minimal sufficient statistic for A (a statistic of dimension one), and in fact, there is not sufficient statistic of dimension less than N. Hence, the data cannot be reduced without loss. It is

reasonable then to use the maximum likelihood estimator of A as a statistic for data reduction. It can be shown that the MLE for A is just the median of the data, which we denote by M [10]. Assuming now that N is odd, it can be shown that the PDF of M is [11]

$$p_M(m;A) = \frac{N!}{\left(\frac{N-1}{2}!\right)^2}\left[F_W(m-A)(1-F_W(m-A))\right]^{(N-1)/2} p_W(m-A) \tag{8}$$

where

$$p_W(w) = \frac{1}{\sqrt{2\sigma^2}}\exp\left[-\sqrt{\frac{2}{\sigma^2}}|w|\right] \tag{9}$$

is the Laplacian PDF and $F_W(w)$ is its corresponding cumulative distribution function, both of which assume that $A = 0$. The latter can be shown to be

$$F_W(w) = u(w) - \frac{1}{2}\mathrm{sgn}(w)\exp\left[-\sqrt{\frac{2}{\sigma^2}}|w|\right] \tag{10}$$

where $u(w) = 0$ for $w < 0$, $u(0) = 1/2$, and $u(w) = 1$ for $w > 0$, and $\mathrm{sgn}(w) = 1$ for $w > 0$, $\mathrm{sgn}(w) = -1$ for $w < 0$, and $\mathrm{sgn}(0) = 0$.

Now using the orthogonal projection theorem the PDF that is closest to the true one of (7) is given by (4). This becomes upon substitution of (8), (9), and (7) with $A = 0$ as the reference PDF into (4), and replacing m by $\mathrm{med}(\mathbf{x})$

$$\begin{aligned}
\hat{p}_X(\mathbf{x};A) &= \frac{p_M(\mathrm{med}(\mathbf{x});A)}{p_M(\mathrm{med}(\mathbf{x});0)}p_X(\mathbf{x};0) \\
&= \frac{\left[F_W(\mathrm{med}(\mathbf{x})-A)(1-F_W(\mathrm{med}(\mathbf{x})-A)\right]^{(N-1)/2}\exp\left[-\sqrt{\frac{2}{\sigma^2}}|\mathrm{med}(\mathbf{x})-A|\right]}{\left[F_W(\mathrm{med}(\mathbf{x}))(1-F_W(\mathrm{med}(\mathbf{x})))\right]^{(N-1)/2}\exp\left[-\sqrt{\frac{2}{\sigma^2}}|\mathrm{med}(\mathbf{x})|\right]} \\
&\quad \cdot \left(\frac{1}{\sqrt{2\sigma^2}}\right)^N \exp\left[-\sqrt{\frac{2}{\sigma^2}}\sum_{n=0}^{N-1}|x[n]|\right]
\end{aligned}$$

In the next section we use this result to indicate how to choose between two competing statistics.

6 Application to Feature Selection

We now apply our results to the problem of comparing which of two statistics would yield a better classifier. Specifically, assume that we have two potential statistics $\mathbf{Z}_1(\mathbf{x})$ and $\mathbf{Z}_2(\mathbf{x})$, on which to base a classifier for a given class. By using the orthogonal projection theorem we construct the two PDFs $\hat{p}_{X_1}(\mathbf{x})$ and $\hat{p}_{X_2}(\mathbf{x})$. Then, we choose the statistic for which the distance between the true PDF and the projected one is minimum. It is important to note that *we can accomplish this goal without actually knowing the true PDF, which is our situation in practice.* This is done as follows. The Kullback-Liebler distance between

the true PDF $p_X(\mathbf{x})$ and the orthogonal projection PDF is $D(p_X(\mathbf{x})||\hat{p}_{X_i}(\mathbf{x}))$, which is

$$\begin{aligned} D(p_X(\mathbf{x})||\hat{p}_{X_i}(\mathbf{x})) &= \int p_X(\mathbf{x}) \ln \frac{p_X(\mathbf{x})}{\hat{p}_{X_i}(\mathbf{x})} d\mathbf{x} \\ &= \int p_X(\mathbf{x}) \ln p_X(\mathbf{x}) d\mathbf{x} - \int p_X(\mathbf{x}) \ln \hat{p}_{X_i}(\mathbf{x}) d\mathbf{x}. \end{aligned}$$

This is minimized by minimizing the second term since the first term does not depend on the approximating PDF. Thus, we should *maximize*

$$\int p_X(\mathbf{x}) \ln \hat{p}_{X_i}(\mathbf{x}) d\mathbf{x}$$

which from (4) is

$$\int p_X(\mathbf{x}) \ln \left[\frac{p_{Z_i}(\mathbf{z}_i(\mathbf{x}); \boldsymbol{\theta}_1)}{p_{Z_i}(\mathbf{z}_i(\mathbf{x}); \boldsymbol{\theta}_0)} p_X(\mathbf{x}; \boldsymbol{\theta}_0) \right] d\mathbf{x}$$

or since $p_X(\mathbf{x}; \boldsymbol{\theta}_0)$ does not depend on our choice of $\mathbf{Z}_i(\mathbf{x})$, we should choose the statistic for which

$$E\left[\ln \frac{p_{Z_i}(\mathbf{z}_i(\mathbf{x}); \boldsymbol{\theta}_1)}{p_{Z_i}(\mathbf{z}_i(\mathbf{x}); \boldsymbol{\theta}_0)} \right] \tag{11}$$

is maximum. Note that the expectation is with respect to the *true* PDF, which allows its estimation from actual data. In practice, if we have the training data $\mathbf{x}_j$ for $j = 1, 2, \ldots, L$, then we base our decision on the *estimated* average log-likelihood ratio

$$\frac{1}{L} \sum_{j=1}^{L} \ln \frac{p_{Z_i}(\mathbf{z}_i(\mathbf{x}_j); \boldsymbol{\theta}_1)}{p_{Z_i}(\mathbf{z}_i(\mathbf{x}_j); \boldsymbol{\theta}_0)}.$$

The PDFs p_{Z_i} would also be estimated from training data. Note that these are *low dimensional PDFs*.

For the example of the last section we might be interested in comparing the feature vector of the MLE, which is the median, to that of the mean. Let $p_M(\text{med}(\mathbf{x}); A)$ denote the PDF of the median and $p_{\bar{X}}(\bar{x}; A)$ denote the PDF of the mean. The PDF of the median was given by (8). The PDF of the mean can be shown to be

$$\begin{aligned} p_{\bar{X}}(\bar{x}; A) &= \frac{N\sqrt{2/\sigma^2}}{2^{2N-1}} \exp\left[-N\sqrt{\frac{2}{\sigma^2}} |\bar{x} - A| \right] \\ &\quad \cdot \sum_{k=0}^{N-1} \binom{2N-2-k}{N-1} \frac{\left(2\sqrt{2/\sigma^2} N |\bar{x} - A| \right)^k}{k!}. \end{aligned} \tag{12}$$

The required average log-likelihood ratios are from (11)

$$E\left[\ln \frac{p_M(\text{med}(\mathbf{x}); A)}{p_M(\text{med}(\mathbf{x}); 0)} \right]$$

and

$$E\left[\ln \frac{p_{\bar{X}}(\bar{x}; A)}{p_{\bar{X}}(\bar{x}; 0)} \right].$$

The importance of the reference PDFs $p_M(\text{med}(\mathbf{x}); 0)$ and $p_{\bar{X}}(\bar{x}; 0)$ is apparent. Without the reference PDFs to act as normalizers, in effect measuring the distance or discrimination from a fixed hypothesis, the comparisons using the numerator PDFs would produce severely biased results. Also, as an upper bound, we compute the average log-likelihood ratio of the original data

$$E\left[\ln \frac{p_X(\mathbf{x}; A)}{p_X(\mathbf{x}; 0)}\right].$$

Using a computer simulation we compare the average log-likelihood ratios versus data record length. We have chosen $A = 0.4$ and $\sigma^2 = 1$. In Figure 1 it is seen that the median provides a higher average log-likelihood ratio and hence would be chosen as the better statistic to use for classification. This is consistent with the known loss of statistical efficiency of the sample mean for a Laplacian PDF versus the median [10].

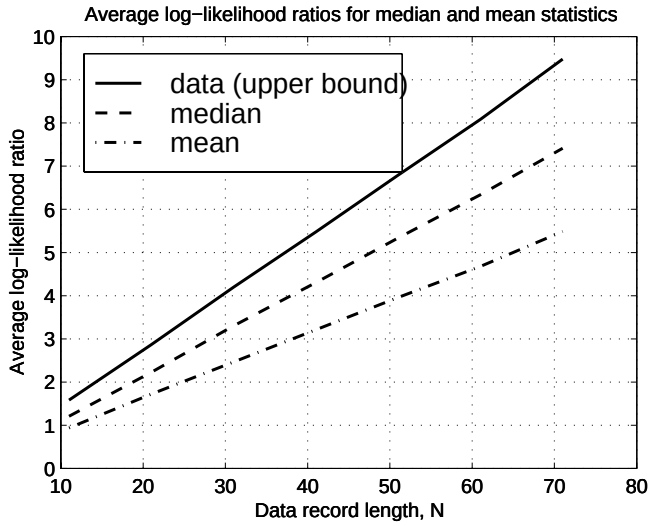

Figure 1: Comparison of average log-likelihood ratios for two statistics. The statistic with the higher average log-likelihood ratio will result in a better performing classifier.

7 Conclusions

A new method for approximating high dimensional PDFs using lower-order ones has been presented. The approximation is asymptotically optimal in that it minimizes the probability of error in a decision rule. An application to the selection of feature vectors for classification was given. In particular, a simple method is proposed to discern among multiple competing feature vectors.

References

1. R.O. Duda, P.E. Hart, D.G. Stork, *Pattern Classification*, J. Wiley, New York, 2001.

2. Lehmann, E.L., *Testing Statistical Hypotheses, 2nd Ed.*, Springer-Verlag, New York, 1986.

3. P.M. Baggenstoss, " The PDF Projection Theorem and the Class-specific Method" , *IEEE Trans. on Signal Processing*, pp. 672–685, March 2003.

4. G. Casella, R.L. Berger, *Statistical Inference*, Duxbury Press, Belmont, CA, 1990.

5. Kay, S.M, *Fundamentals of Statistical Signal Processing: Estimation Theory*, Prentice-Hall, Englewood Cliffs, NJ, 1993.

6. Amari, S., Nagaoka, H., *Methods of Information Geometry*, Oxford University Press, 1993

7. S. Kay, A. Nuttall, P.M. Baggenstoss, "Multidimensional Probability Density Function Approximations for Detection, Classification, and Model Order Selection", *IEEE Trans. on Signal Processing*, pp. 2240–2252, Oct. 2001.

8. Cover, T.M., Thomas, J.A., *Elements of Information Theory*, J. Wiley, New York, 1991.

9. Kullback, S., *Information Theory and Statistics*, Dover Pubs., New York, 1959.

10. Kay, S.M, *Fundamentals of Statistical Signal Processing: Detection Theory*, Prentice-Hall, Englewood Cliffs, NJ, 1998.

11. Sir M. Kendall, A. Stuart, *The Advanced Theory of Statistics, Vol. 1*, Macmillan Pub., New York, 1977

12. Melvin, W.L., "Space-Time Adaptive Radar Performance in Heterogenous Clutter", *IEEE Trans. on Aerospace and Electronics*, pp. 621–633, Vol. 36, April 2000.

13. Blahut, R.E., *Principles and Practice of Information Theory*, Addison-Wesley, Reading, MA, 1987.