Sufficiency, Classification, and the Class-Specific Feature Theorem

Steven Kay, *Fellow, IEEE*

***Abstract*—A new proof of the class-specific feature theorem is given. The proof makes use of the observed data as opposed to the set of sufficient statistics as in the original formulation. We prove the theorem for the classical case, in which the parameter vector is deterministic and known, as well as for the Bayesian case, in which the parameter vector is modeled as a random vector with known prior probability density function. The essence of the theorem is that with a suitable normalization the probability density function of the sufficient statistic for each probability density function family can be used for optimal classification. One need not have knowledge of the probability density functions of the data under each hypothesis.**

***Index Terms*—Bayes procedures, data models, information theory, pattern recognition, signal detection.**

I. Introduction

Optimal classification depends upon knowledge of the joint probability density function (pdf) of the observed data. When this is unavailable, as is always the case in practice, the pdf must be estimated. Unfortunately, it has been noted that high-dimensional pdf's, on the order of ten or higher, are notoriously difficult to estimate [7]. It is therefore advisable to transform the data to a lower dimensional feature vector. In fact, much of the research in classification is involved with the determination of a set of features which describe the data, but which has minimal dimension. Statistical hypothesis testing indicates that the optimal way to do this is to employ sufficient statistics, which reduce the data but retain all the information of the original data. The theory of sufficiency is well established when applied to a family of pdf's that are parameterized [8], i.e., each pdf in the family depends on a different value of a parameter. Less is known about sufficiency when the possible pdf's may be from several parameterized families. For example, consider for the data set $\boldsymbol{x} = [x_1 x_2 \cdots x_N]^T$ the family of pdf's

$$p(\boldsymbol{x};\, \mu) = \frac{1}{(2\pi)^{N/2}} \exp\left[-\frac{1}{2}\sum_{i=1}^{N}(x_i - \mu)^2\right]$$

where each pdf is parameterized by μ, the mean of x_i. Then, a sufficient statistic is well known to be $T(\boldsymbol{x}) = \sum_{i=1}^{N} x_i$. Any decisions concerning the value of μ can be based on knowledge of the sufficient statistic $T(\boldsymbol{x})$ only [5]. The performance of the resultant decision rule will be identical to that based on the original data set $\boldsymbol{x}$. The hypothesis test in this case might be a simple one such as deciding whether $\mu = 0$ or $\mu = 1$. A completely different situation arises when two or more pdf families may describe the observed data. Such is the usual case in classification. For example, say that $\boldsymbol{x}$ is observed, where the x_i's are independent and identically distributed (i.i.d.) according to either an exponential pdf or a log-normal pdf [1]. The univariate exponential pdf is given by

$$p_1(x;\, \theta_1) = \begin{cases} \dfrac{1}{\theta_1} \exp(-x/\theta_1), & x > 0 \\ 0, & x < 0 \end{cases}$$

while the univariate log-normal pdf is

$$p_2(x;\, \theta_2) = \begin{cases} \dfrac{1}{x\sqrt{2\pi\theta_2}} \exp\left(-\frac{1}{2\theta_2}\ln^2(x)\right), & x > 0 \\ 0, & x < 0 \end{cases}$$

where $\theta_1 > 0$ and $\theta_2 > 0$. The joint pdf's are

$$p_1(\boldsymbol{x};\theta_1) = \begin{cases} \frac{1}{\theta_1^N} \exp\left(-\frac{1}{\theta_1}\sum_{i=1}^{N} x_i\right), & x > 0 \\ 0, & x < 0 \end{cases}$$

$$p_2(\boldsymbol{x};\theta_2) = \begin{cases} \dfrac{1}{\left(\prod_{i=1}^{N} x_i\right)(2\pi\theta_2)^{N/2}} \exp\left(-\frac{1}{2\theta_2}\sum_{i=1}^{N}\ln^2(x_i)\right), & x > 0 \\ 0, & x < 0. \end{cases}$$

It then follows from the Neyman–Fisher (NF) factorization theorem [3] that each pdf admits a sufficient statistic as

$$T_1(\boldsymbol{x}) = \sum_{i=1}^{N} x_i$$

$$T_2(\boldsymbol{x}) = \sum_{i=1}^{N} \ln^2(x_i).$$

Hence, $T_1(\boldsymbol{x})$ can be used to make decisions about θ_1 and $T_2(\boldsymbol{x})$ can be used to make decisions about θ_2. The question now arises as to whether the *joint statistic* $\boldsymbol{T}(\boldsymbol{x}) = [T_1(\boldsymbol{x}) \quad T_2(\boldsymbol{x})]^T$ can be used to make an optimal decision if $\boldsymbol{x}$ was sampled from the exponential or the log-normal pdf. The answer is, unfortunately, *no*. As an example, if θ_1 and θ_2 are known, so that the hypothesis test is simple, a likelihood ratio test would involve the additional statistic $T_3(\boldsymbol{x}) = \prod_{i=1}^{N} x_i$. Thus $T_1(\boldsymbol{x})$ and $T_2(\boldsymbol{x})$, which we term the *single family sufficient statistics* is not a sufficient statistic for the *set of pdf families* given by $\{p_1(\boldsymbol{x};\, \theta_1), p_2(\boldsymbol{x};\theta_2)\}$. We will formalize this result in the next section.

Even though the single family sufficient statistic is not sufficient, it is possible to base a decision rule on the pdf's of the single family sufficient statistics, when the pdf's are *suitably normalized*. In Section III, the linear model family is used to illustrate the normalization necessary for the classifier to be based on the single family sufficient statistic. This is the essence of the class-specific feature theorem, as originally formulated by [2], and which we describe more fully in Section IV. Before doing so we examine sufficient statistics for the multiple pdf family in the next section.

II. Sufficient Statistics for Multiple PDF Families

We assume that there are two pdf families of interest, although the general case of M families follows easily. Let the families be described by $p_1(\boldsymbol{x};\, \boldsymbol{\theta}_1)$ and $p_2(\boldsymbol{x};\boldsymbol{\theta}_2)$, where the parameters now are in general *vector* parameters. The dimensionalities of the parameter vectors need not be the same. Also, assume that the families admit single family sufficient statistics $\boldsymbol{T}_1$ and $\boldsymbol{T}_2$, respectively. (We will henceforth drop

Manuscript received February 1, 1999; revised December 8, 1999. This work was supported by the Naval Undersea Warfare Center under Contract N66604-98-C-2376.

The author is with the Department of Electrical and Computer Engineering, University of Rhode Island, Kingston, RI 02881 USA (e-mail: kay@ele.uri.edu).

Communicated by S. Kulkarni, Associate Editor for Nonparametric Estimation, Classification, and Neural Networks.

Publisher Item Identifier S 0018-9448(00)04656-3.

the dependence of $\boldsymbol{T}$ on $\boldsymbol{x}$.) Then, to describe the set of possible pdf's we introduce an additional parameter γ, which takes on the values $0 \le \gamma \le 1$. Then, we can define the pdf for the *composite pdf family* [1]

$$p(\boldsymbol{x};\, \gamma,\, \boldsymbol{\theta}_1,\, \boldsymbol{\theta}_2) = \frac{p_1^{\gamma}(\boldsymbol{x};\, \boldsymbol{\theta}_1) p_2^{1-\gamma}(\boldsymbol{x};\, \boldsymbol{\theta}_2)}{\int p_1^{\gamma}(\boldsymbol{x};\, \boldsymbol{\theta}_1) p_2^{1-\gamma}(\boldsymbol{x};\, \boldsymbol{\theta}_2)\, d\boldsymbol{x}} \tag{1}$$

where $\gamma = 1$ yields p_1 and $\gamma = 0$ yields p_2. We are now in a position to find the sufficient statistic, if it exists, for the composite pdf family. From the NF factorization theorem we have that

$$p_1(\boldsymbol{x};\, \boldsymbol{\theta}_1) = g_1(\boldsymbol{T}_1,\, \boldsymbol{\theta}_1) h_1(\boldsymbol{x})$$
$$p_2(\boldsymbol{x};\, \boldsymbol{\theta}_2) = g_2(\boldsymbol{T}_2,\, \boldsymbol{\theta}_2) h_2(\boldsymbol{x})$$

and using this in (1) produces

$$p(\boldsymbol{x};\, \gamma,\, \boldsymbol{\theta}_1,\, \boldsymbol{\theta}_2) = \frac{g_1^{\gamma}(\boldsymbol{T}_1,\, \boldsymbol{\theta}_1) g_2^{1-\gamma}(\boldsymbol{T}_2,\, \boldsymbol{\theta}_2) (h_1(\boldsymbol{x})/h_2(\boldsymbol{x}))^{\gamma} h_2(\boldsymbol{x})}{\int g_1^{\gamma}(\boldsymbol{T}_1,\, \boldsymbol{\theta}_1) g_2^{1-\gamma}(\boldsymbol{T}_2,\, \boldsymbol{\theta}_2) (h_1(\boldsymbol{x})/h_2(\boldsymbol{x}))^{\gamma} h_2(\boldsymbol{x})\, d\boldsymbol{x}}.$$

Note that if $h_1(\boldsymbol{x})/h_2(\boldsymbol{x})$ does not depend on $\boldsymbol{x}$, then $\boldsymbol{T} = [\boldsymbol{T}_1 \;\; \boldsymbol{T}_2]^T$ will be a sufficient statistic for $\gamma, \boldsymbol{\theta}_1, \boldsymbol{\theta}_2$. For if this is true, then by the NF factorization theorem

$$p(\boldsymbol{x};\, \gamma,\, \boldsymbol{\theta}_1,\, \boldsymbol{\theta}_2) = \underbrace{\frac{g_1^{\gamma}(\boldsymbol{T}_1,\, \boldsymbol{\theta}_1) g_2^{1-\gamma}(\boldsymbol{T}_2,\, \boldsymbol{\theta}_2)}{\int g_1^{\gamma}(\boldsymbol{T}_1,\, \boldsymbol{\theta}_1) g_2^{1-\gamma}(\boldsymbol{T}_2,\, \boldsymbol{\theta}_2) h_2(\boldsymbol{x})\, d\boldsymbol{x}}}_{g(T_1, T_2, \gamma, \boldsymbol{\theta}_1, \boldsymbol{\theta}_2)} \cdot \underbrace{h_2(\boldsymbol{x})}_{h(\boldsymbol{x})}.$$

In the exponential versus log-normal example

$$p_1(\boldsymbol{x};\, \theta_1) = \underbrace{\frac{1}{\theta_1^N} \exp\left(-\frac{1}{\theta_1} \sum_{i=1}^{N} x_i\right)}_{g_1(T_1, \theta_1)} \cdot \underbrace{I_{(0,\,\infty)}(\boldsymbol{x})}_{h_1(\boldsymbol{x})}$$

$$p_2(\boldsymbol{x};\, \theta_2) = \underbrace{\frac{1}{(2\pi\theta_2)^{N/2}} \exp\left(-\frac{1}{2\theta_2} \sum_{i=1}^{N} \ln^2(x_i)\right)}_{g_2(T_2, \theta_2)} \cdot \underbrace{\frac{I_{(0,\,\infty)}(\boldsymbol{x})}{\prod_{i=1}^{N} x_i}}_{h_2(\boldsymbol{x})}$$

where $I_{(0,\,\infty)}(\boldsymbol{x}) = 1$ if all $x_i > 0$ and is zero otherwise. Thus

$$h_1(\boldsymbol{x})/h_2(\boldsymbol{x}) = 1/\prod_{i=1}^{N} x_i, \qquad \text{for all } x_i > 0$$

which clearly depends on $\boldsymbol{x}$. Note that in this case the sufficient statistic is

$$\left[\sum_{i=1}^{N} x_i \quad \sum_{i=1}^{N} \ln^2(x_i) \quad \prod_{i=1}^{N} x_i\right]^T.$$

In general, the sufficient statistic will be $[\boldsymbol{T}_1 \;\; \boldsymbol{T}_2 \;\; h_1(\boldsymbol{x})/h_2(\boldsymbol{x})]^T$.

III. Optimal Decision Rules

As shown previously, the set of single pdf family sufficient statistics is not in general a sufficient statistic in the multiple pdf family case. However, an optimal decision rule can still be implemented if the pdf's of the single family sufficient statistics are known. This requires a normalization, which in essence, accounts for the different $h(\boldsymbol{x})$ for different pdf families. To illustrate the result we resort to the classical Gaussian linear model using two pdf families with different dimensionality parameter vectors. The linear model is defined as [3]

$$\boldsymbol{x} = \boldsymbol{H}\boldsymbol{\theta} + \boldsymbol{w}$$

where $\boldsymbol{H}$ is an $N \times p$ known observation matrix, $\boldsymbol{\theta}$ is a $p \times 1$ parameter vector, and $\boldsymbol{w}$ is a Gaussian random vector with mean $\boldsymbol{0}$ and covariance matrix $\sigma^2 \boldsymbol{I}$, with σ^2 known. The pdf can be factored as

$$\begin{aligned} p(\boldsymbol{x};\, \boldsymbol{\theta}) &= \frac{1}{(2\pi\sigma^2)^{N/2}} \exp\left[-\frac{1}{2\sigma^2} (\boldsymbol{x} - \boldsymbol{H}\boldsymbol{\theta})^T (\boldsymbol{x} - \boldsymbol{H}\boldsymbol{\theta})\right] \\ &= \underbrace{\exp\left[-\frac{1}{2\sigma^2} (\boldsymbol{T} - \boldsymbol{\theta})^T \boldsymbol{H}^T \boldsymbol{H} (\boldsymbol{T} - \boldsymbol{\theta})\right]}_{g(\boldsymbol{T}, \boldsymbol{\theta})} \\ &\quad \cdot \underbrace{\frac{1}{(2\pi\sigma^2)^{N/2}} \exp\left[-\frac{1}{2\sigma^2} (\boldsymbol{x} - \boldsymbol{H}\boldsymbol{T})^T (\boldsymbol{x} - \boldsymbol{H}\boldsymbol{T})\right]}_{h(\boldsymbol{x})} \end{aligned} \tag{2}$$

where $\boldsymbol{T} = (\boldsymbol{H}^T\boldsymbol{H})^{-1}\boldsymbol{H}^T\boldsymbol{x}$ is a sufficient statistic for $\boldsymbol{\theta}$ and is recognized as the minimum variance unbiased estimator. Note that $h(\boldsymbol{x})$ depends on the dimensionality of $\boldsymbol{\theta}$, as well as $\boldsymbol{H}$. The pdf of the sufficient statistic is easily shown to be $\mathcal{N}(\boldsymbol{\theta},\, \sigma^2(\boldsymbol{H}^T\boldsymbol{H})^{-1})$, where $\mathcal{N}(\boldsymbol{\mu},\, \boldsymbol{C})$ denotes a multivariate Gaussian pdf with mean vector $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{C}$. Hence, we may rewrite (2) as shown in (3) at the bottom of this page. Now consider a hypothesis test within the same linear model family. An example might be to test $\boldsymbol{\theta} = \boldsymbol{\theta}_1$ versus $\boldsymbol{\theta} = \boldsymbol{\theta}_0$. The solution, which is known to be a likelihood ratio test, decides in favor of $\boldsymbol{\theta}_1$ if $L(\boldsymbol{x})$, the likelihood ratio, exceeds a threshold. But from (3) this becomes

$$\begin{aligned} L(\boldsymbol{x}) &= \frac{p(\boldsymbol{x};\, \boldsymbol{\theta}_1)}{p(\boldsymbol{x};\, \boldsymbol{\theta}_0)} \\ &= \frac{p(\boldsymbol{T};\, \boldsymbol{\theta}_1) h'(\boldsymbol{x})}{p(\boldsymbol{T};\, \boldsymbol{\theta}_0) h'(\boldsymbol{x})} \\ &= \frac{p(\boldsymbol{T};\, \boldsymbol{\theta}_1)}{p(\boldsymbol{T};\, \boldsymbol{\theta}_0)}. \end{aligned}$$

Clearly, we would obtain the same result if we had started with the sufficient statistic instead of the data and formed the likelihood ratio based on it. Now consider the problem of testing whether the data was sampled from one of two different linear models or

$$\mathcal{H}_1: \quad \boldsymbol{x} = \boldsymbol{H}_1\boldsymbol{\alpha} + \boldsymbol{w}_1$$
$$\mathcal{H}_2: \quad \boldsymbol{x} = \boldsymbol{H}_2\boldsymbol{\beta} + \boldsymbol{w}_2$$

where the parameters $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ have different dimensionalities (and, of course, $\boldsymbol{H}_1, \boldsymbol{H}_2$ are different). An example, might be whether the data

$$p(\boldsymbol{x};\, \boldsymbol{\theta}) = \frac{1}{(2\pi)^{p/2} \det^{\frac{1}{2}}(\sigma^2(\boldsymbol{H}^T\boldsymbol{H})^{-1})} \cdot \exp\left[-\frac{1}{2} (\boldsymbol{T} - \boldsymbol{\theta})^T (\boldsymbol{H}^T\boldsymbol{H}/\sigma^2)(\boldsymbol{T} - \boldsymbol{\theta})\right]$$
$$\cdot \underbrace{\frac{1}{(2\pi\sigma^2)^{(N-p)/2} \det^{\frac{1}{2}}(\boldsymbol{H}^T\boldsymbol{H})} \exp\left[-\frac{1}{2\sigma^2} (\boldsymbol{x} - \boldsymbol{H}\boldsymbol{T})^T (\boldsymbol{x} - \boldsymbol{H}\boldsymbol{T})\right]}_{h'(\boldsymbol{x})} = p(\boldsymbol{T};\, \boldsymbol{\theta}) h'(\boldsymbol{x}) \tag{3}$$

originated from a DC level in white Gaussian noise or a straight line in white Gaussian noise. Then, the likelihood ratio is from (3)

$$L(\boldsymbol{x}) = \frac{p_1(\boldsymbol{x};\,\boldsymbol{\alpha})}{p_2(\boldsymbol{x};\,\boldsymbol{\beta})} \tag{4}$$

$$= \frac{p(\boldsymbol{T}_\alpha;\,\boldsymbol{\alpha})}{p(\boldsymbol{T}_\beta;\,\boldsymbol{\beta})} \cdot \frac{h_1'(\boldsymbol{x})}{h_2'(\boldsymbol{x})}. \tag{5}$$

It is seen that because of the different models $h_1'(\boldsymbol{x}) \neq h_2'(\boldsymbol{x})$ and this difference must be taken into account by the LRT. To do so note from (4) and (5) that

$$\frac{h_1'(\boldsymbol{x})}{h_2'(\boldsymbol{x})} = \frac{p_1(\boldsymbol{x};\,\boldsymbol{\alpha})}{p_2(\boldsymbol{x};\,\boldsymbol{\beta})} \cdot \frac{p(\boldsymbol{T}_\beta;\,\boldsymbol{\beta})}{p(\boldsymbol{T}_\alpha;\,\boldsymbol{\alpha})}. \tag{6}$$

Since the left-hand side depends only on $\boldsymbol{x}$, the right-hand side cannot depend on $\boldsymbol{\alpha}$, $\boldsymbol{\beta}$. Hence, if we can find an $\boldsymbol{\alpha}$, say $\boldsymbol{\alpha}^*$ and a $\boldsymbol{\beta}$, say $\boldsymbol{\beta}^*$ so that

$$p_1(\boldsymbol{x};\,\boldsymbol{\alpha}^*) = p_2(\boldsymbol{x};\,\boldsymbol{\beta}^*) \tag{7}$$

then from (6) we have that

$$\frac{h_1'(\boldsymbol{x})}{h_2'(\boldsymbol{x})} = \frac{p(\boldsymbol{T}_\beta;\,\boldsymbol{\beta}^*)}{p(\boldsymbol{T}_\alpha;\,\boldsymbol{\alpha}^*)}.$$

This can be used in (5) to yield finally

$$L(\boldsymbol{x}) = \frac{p(\boldsymbol{T}_\alpha;\,\boldsymbol{\alpha})/p(\boldsymbol{T}_\alpha;\,\boldsymbol{\alpha}^*)}{p(\boldsymbol{T}_\beta;\,\boldsymbol{\beta})/p(\boldsymbol{T}_\beta;\,\boldsymbol{\beta}^*)}. \tag{8}$$

Note that *the likelihood ratio depends only on the pdf's of the single family sufficient statistics*. Thus in practice we need only estimate the pdf of $\boldsymbol{T}$ as opposed to that of $\boldsymbol{x}$. This approach is called the *class-specific model* [2].

The class-specific model depends upon (7) being satisfied. An obvious choice for the linear model is to choose $\boldsymbol{\alpha} = \boldsymbol{0}$ and $\boldsymbol{\beta} = \boldsymbol{0}$, which is the noise-only condition. In this case we have that

$$p_1(\boldsymbol{x};\,\boldsymbol{\alpha}^* = \boldsymbol{0}) = p_2(\boldsymbol{x};\,\boldsymbol{\beta}^* = \boldsymbol{0}) = \frac{1}{(2\pi\sigma^2)^{\frac{N}{2}}} \exp\left(-\frac{1}{2\sigma^2}\,\boldsymbol{x}^T\boldsymbol{x}\right).$$

To verify that the likelihood ratio computed from either the data as $p_1(\boldsymbol{x};\,\boldsymbol{\alpha})/p_2(\boldsymbol{x};\,\boldsymbol{\beta})$ or (8) using $\boldsymbol{\alpha}^* = \boldsymbol{\beta}^* = \boldsymbol{0}$ are identical first recall that $\boldsymbol{T}_\alpha \sim \mathcal{N}(\boldsymbol{\alpha},\,\sigma^2(\boldsymbol{H}_1^T\boldsymbol{H}_1)^{-1})$ and $\boldsymbol{T}_\beta \sim \mathcal{N}(\boldsymbol{\beta},\,\sigma^2(\boldsymbol{H}_2^T\boldsymbol{H}_2)^{-1})$. Then

$$\frac{p(\boldsymbol{T}_\alpha;\,\boldsymbol{\alpha})}{p(\boldsymbol{T}_\alpha;\,\boldsymbol{\alpha}=\boldsymbol{0})} = \frac{\frac{\exp\left[-\frac{1}{2}(\boldsymbol{T}_\alpha-\boldsymbol{\alpha})^T\left(\frac{\boldsymbol{H}_1^T\boldsymbol{H}_1}{\sigma^2}\right)(\boldsymbol{T}_\alpha-\boldsymbol{\alpha})\right]}{(2\pi)^{p_1/2}\,\det^{\frac{1}{2}}(\sigma^2(\boldsymbol{H}_1^T\boldsymbol{H}_1)^{-1})}}{\frac{\exp\left[-\frac{1}{2}\boldsymbol{T}_\alpha^T\left(\frac{\boldsymbol{H}_1^T\boldsymbol{H}_1}{\sigma^2}\right)\boldsymbol{T}_\alpha\right]}{(2\pi)^{p_1/2}\,\det^{\frac{1}{2}}(\sigma^2(\boldsymbol{H}_1^T\boldsymbol{H}_1)^{-1})}}$$

$$= \exp\left[-\frac{1}{2}\left(-2\boldsymbol{\alpha}^T\left(\frac{\boldsymbol{H}_1^T\boldsymbol{H}_1}{\sigma^2}\right)\boldsymbol{T}_\alpha + \boldsymbol{\alpha}^T\left(\frac{\boldsymbol{H}_1^T\boldsymbol{H}_1}{\sigma^2}\right)\boldsymbol{\alpha}\right)\right]$$

and, similarly, for $p(\boldsymbol{T}_\beta;\,\boldsymbol{\beta})/p(\boldsymbol{T}_\beta;\,\boldsymbol{\beta}=\boldsymbol{0})$. The likelihood ratio becomes from (8)

$$L(\boldsymbol{x}) = \frac{\exp\left[-\frac{1}{2}\left(-2\boldsymbol{\alpha}^T\left(\frac{\boldsymbol{H}_1^T\boldsymbol{H}_1}{\sigma^2}\right)\boldsymbol{T}_\alpha + \boldsymbol{\alpha}^T\left(\frac{\boldsymbol{H}_1^T\boldsymbol{H}_1}{\sigma^2}\right)\boldsymbol{\alpha}\right)\right]}{\exp\left[-\frac{1}{2}\left(-2\boldsymbol{\beta}^T\left(\frac{\boldsymbol{H}_2^T\boldsymbol{H}_2}{\sigma^2}\right)\boldsymbol{T}_\beta + \boldsymbol{\beta}^T\left(\frac{\boldsymbol{H}_2^T\boldsymbol{H}_2}{\sigma^2}\right)\boldsymbol{\beta}\right)\right]}. \tag{9}$$

But $\boldsymbol{T}_\alpha = (\boldsymbol{H}_1^T\boldsymbol{H}_1)^{-1}\boldsymbol{H}_1^T\boldsymbol{x}$ so that

$$-2\boldsymbol{\alpha}^T\left(\frac{\boldsymbol{H}_1^T\boldsymbol{H}_1}{\sigma^2}\right)\boldsymbol{T}_\alpha + \boldsymbol{\alpha}^T\left(\frac{\boldsymbol{H}_1^T\boldsymbol{H}_1}{\sigma^2}\right)\boldsymbol{\alpha} = -2\boldsymbol{\alpha}^T\left(\frac{\boldsymbol{H}_1^T\boldsymbol{H}_1}{\sigma^2}\right)\left(\boldsymbol{H}_1^T\boldsymbol{H}_1\right)^{-1}\boldsymbol{H}_1^T\boldsymbol{x} + \frac{\boldsymbol{\alpha}^T\boldsymbol{H}_1^T\boldsymbol{H}_1\boldsymbol{\alpha}}{\sigma^2}$$

$$= \frac{1}{\sigma^2}\left(-2\boldsymbol{\alpha}^T\boldsymbol{H}_1^T\boldsymbol{x} + \boldsymbol{\alpha}^T\boldsymbol{H}_1^T\boldsymbol{H}_1\boldsymbol{\alpha}\right)$$

and similarly for the denominator term in (9). Hence

$$L(\boldsymbol{x}) = \frac{\exp\left[-\frac{1}{2\sigma^2}\left(-2\boldsymbol{\alpha}^T\boldsymbol{H}_1^T\boldsymbol{x} + \boldsymbol{\alpha}^T\boldsymbol{H}_1^T\boldsymbol{H}_1\boldsymbol{\alpha}\right)\right]}{\exp\left[-\frac{1}{2\sigma^2}\left(-2\boldsymbol{\beta}^T\boldsymbol{H}_1^T\boldsymbol{x} + \boldsymbol{\beta}^T\boldsymbol{H}_2^T\boldsymbol{H}_2\boldsymbol{\beta}\right)\right]}$$

which is identical to

$$L(\boldsymbol{x}) = \frac{p_1(\boldsymbol{x};\,\boldsymbol{\alpha})}{p_2(\boldsymbol{x};\,\boldsymbol{\beta})} = \frac{\frac{1}{(2\pi\sigma^2)^{\frac{N}{2}}}\exp\left[-\frac{1}{2\sigma^2}(\boldsymbol{x}-\boldsymbol{H}_1\boldsymbol{\alpha})^T(\boldsymbol{x}-\boldsymbol{H}_1\boldsymbol{\alpha})\right]}{\frac{1}{(2\pi\sigma^2)^{\frac{N}{2}}}\exp\left[-\frac{1}{2\sigma^2}(\boldsymbol{x}-\boldsymbol{H}_2\boldsymbol{\beta})^T(\boldsymbol{x}-\boldsymbol{H}_2\boldsymbol{\beta})\right]}.$$

IV. Class-Specific Feature Theorem

We now state the class-specific feature theorem. Two cases are of interest. The classical case assumes the parameter is deterministic and known while the Bayesian case assumes the parameter is random but with a known prior pdf. We now utilize a more descriptive notation for the pdf's to avoid confusion between the classical and Bayesian cases. For the classical case, the pdf for the data $\boldsymbol{x}$, which is parameterized by the vector $\boldsymbol{\theta}$, is denoted by $p_X(\boldsymbol{x};\,\boldsymbol{\theta})$. In the Bayesian case, the parameter vector is assumed random with known prior pdf $p_\Theta(\boldsymbol{\theta})$ and the conditional pdf is denoted by $p_{X|\Theta}(\boldsymbol{x}|\boldsymbol{\theta})$. In general, $\boldsymbol{x}$ is $N \times 1$ and $\boldsymbol{\theta}$ is $p \times 1$. The following lemma is used to prove the main theorem and is of interest in its own right. It states that not only does a sufficient statistic provide all the information necessary for a decision problem [5], but in the case of a simple versus simple hypothesis the likelihood ratios are identical. A similar result is implicit in the proof of the equivalence of the Kullback–Leibler information discrimination measure based on the data and a sufficient statistic [6].

Lemma 4.1 (Equivalence of Likelihood Ratio Based on Data and Sufficient Statistics): Let $\boldsymbol{T}(\boldsymbol{x})$ be a sufficient statistic for the pdf family $p_X(\boldsymbol{x};\,\boldsymbol{\theta})$ and let $p_T(\boldsymbol{t};\,\boldsymbol{\theta})$ be the pdf of $\boldsymbol{T}$. Then for any two values of $\boldsymbol{\theta}$ we have that

$$\frac{p_X(\boldsymbol{x};\,\boldsymbol{\theta}_1)}{p_X(\boldsymbol{x};\,\boldsymbol{\theta}_0)} = \frac{p_T(\boldsymbol{t};\,\boldsymbol{\theta}_1)}{p_T(\boldsymbol{t};\,\boldsymbol{\theta}_0)} \tag{10}$$

where $\boldsymbol{t} = \boldsymbol{T}(\boldsymbol{x})$. Note that the equivalence of the likelihood ratios is at the point $\boldsymbol{x}$ for the data likelihood and at the point $\boldsymbol{t} = \boldsymbol{T}(\boldsymbol{x})$ for the sufficient statistic likelihood ratio.

The proof is based on the Neyman–Fisher factorization theorem and is omitted. We now state the main theorem, starting with the classical case. Consider the problem of choosing among the hypotheses $\{\mathcal{H}_1, \mathcal{H}_2, \cdots, \mathcal{H}_M\}$, where the ith hypothesis has the prior probability $P(\mathcal{H}_i)$. If $\mathcal{H}_i$ occurs, the data has the conditional pdf $p_{X|\mathcal{H}_i}(\boldsymbol{x};\,\boldsymbol{\theta}_i)$. The pdf families under each hypothesis need not be the same nor the dimensionalities of the $\boldsymbol{\theta}_i$'s. The decision rule that minimizes the probability of error is well known to be the maximum *a posteriori* (MAP) decision rule [4]. That is, we choose the hypothesis for which the *a posteriori* probability or $\xi_i = P(\mathcal{H}_i|\boldsymbol{x};\,\boldsymbol{\theta}_i)$ is maximum. We note that this procedure may be effected by a

sequence of binary decisions as follows. Compare ξ_1 to ξ_2 and choose the larger. If, say, ξ_1 is larger, then compare ξ_1 to ξ_3 and choose the maximum. We repeat the procedure until all ξ_i's have been examined. The surviving ξ will be the maximum. We will term this somewhat obvious approach *binary decision making*.

Theorem 4.1 (Class-Specific Feature—Classical Case): Assume that for each hypothesis the pdf family $p_{X|\mathcal{H}_i}(\boldsymbol{x};\,\boldsymbol{\theta}_i)$ admits a sufficient statistic $\boldsymbol{T}_i$ for $\boldsymbol{\theta}_i$. Each $\boldsymbol{\theta}_i$ is assumed known. If we can find a set of values of the $\boldsymbol{\theta}$'s so that for each binary decision between $\mathcal{H}_i$ and $\mathcal{H}_j$ we have

$$p_{X|\mathcal{H}_i}(\boldsymbol{x};\,\boldsymbol{\theta}_i^*) = p_{X|\mathcal{H}_j}(\boldsymbol{x};\,\boldsymbol{\theta}_j^*) \tag{11}$$

then the MAP rule for choosing among $\{\mathcal{H}_1, \mathcal{H}_2, \cdots, \mathcal{H}_M\}$ is to make binary decisions (a total of $M-1$) between $\mathcal{H}_i$ and $\mathcal{H}_j$ and to decide $\mathcal{H}_i$ if

$$\frac{p_{T_i|\mathcal{H}_i}(\boldsymbol{t}_i;\,\boldsymbol{\theta}_i)}{p_{T_i|\mathcal{H}_i}(\boldsymbol{t}_i;\,\boldsymbol{\theta}_i^*)}\,P(\mathcal{H}_i) > \frac{p_{T_j|\mathcal{H}_j}(\boldsymbol{t}_j;\,\boldsymbol{\theta}_j)}{p_{T_j|\mathcal{H}_j}(\boldsymbol{t}_j;\,\boldsymbol{\theta}_j^*)}\,P(\mathcal{H}_j). \tag{12}$$

Proof: Consider the binary decision between $\mathcal{H}_i$ and $\mathcal{H}_j$. Then, the MAP rule chooses $\mathcal{H}_i$ if

$$p_{X|\mathcal{H}_i}(\boldsymbol{x};\,\boldsymbol{\theta}_i)P(\mathcal{H}_i) > p_{X|\mathcal{H}_j}(\boldsymbol{x};\,\boldsymbol{\theta}_j)P(\mathcal{H}_j).$$

Now using (10) we have that

$$p_{X|\mathcal{H}_i}(\boldsymbol{x};\,\boldsymbol{\theta}_i) = p_{X|\mathcal{H}_i}(\boldsymbol{x};\,\boldsymbol{\theta}_i^*)\,\frac{p_{T|\mathcal{H}_i}(\boldsymbol{t}_i;\,\boldsymbol{\theta}_i)}{p_{T|\mathcal{H}_i}(\boldsymbol{t}_i;\,\boldsymbol{\theta}_i^*)}$$

and thus we choose $\mathcal{H}_i$ if

$$p_{X|\mathcal{H}_i}(\boldsymbol{x};\,\boldsymbol{\theta}_i^*)\,\frac{p_{T|\mathcal{H}_i}(\boldsymbol{t}_i;\,\boldsymbol{\theta}_i)}{p_{T|\mathcal{H}_i}(\boldsymbol{t}_i;\,\boldsymbol{\theta}_i^*)}\,P(\mathcal{H}_i) > p_{X|\mathcal{H}_j}(\boldsymbol{x};\,\boldsymbol{\theta}_j^*)\,\frac{p_{T|\mathcal{H}_i}(\boldsymbol{t}_j;\,\boldsymbol{\theta}_j)}{p_{T|\mathcal{H}_i}(\boldsymbol{t}_j;\,\boldsymbol{\theta}_j^*)}\,P(\mathcal{H}_j)$$

or, finally, using (11), we have the result.

It is sufficient but not necessary for the condition of (11) to hold that

$$p_{X|\mathcal{H}_1}(\boldsymbol{x};\,\boldsymbol{\theta}_1^*) = p_{X|\mathcal{H}_2}(\boldsymbol{x};\,\boldsymbol{\theta}_2^*) = \cdots = p_{X|\mathcal{H}_M}(\boldsymbol{x};\,\boldsymbol{\theta}_M^*).$$

In practice, it appears that this sufficient condition is most easily satisfied. For example, in the linear model example of Section III we could have considered the problem of

$$\begin{aligned}\mathcal{H}_1:&\quad \boldsymbol{x} = \boldsymbol{H}_1\boldsymbol{\alpha} + \boldsymbol{w}_1\\ \mathcal{H}_2:&\quad \boldsymbol{x} = \boldsymbol{H}_2\boldsymbol{\beta} + \boldsymbol{w}_2\\ \mathcal{H}_3:&\quad \boldsymbol{x} = \boldsymbol{H}_3\boldsymbol{\gamma} + \boldsymbol{w}_3.\end{aligned}$$

Then, the sufficient condition is satisfied if we choose $\boldsymbol{\alpha}^*, \boldsymbol{\beta}^*, \boldsymbol{\gamma}^*$ equal to the zero vector of commensurate dimension.

For the Bayesian case we have the following corresponding theorem.

Theorem 4.2 (Class-Specific Feature—Bayesian Case): Assume that for each hypothesis the conditional pdf family $p_{X|\Theta_i,\,\mathcal{H}_i}(\boldsymbol{x}|\boldsymbol{\theta}_i)$ admits a sufficient statistic $\boldsymbol{T}_i$ for $\boldsymbol{\theta}_i$. The prior pdf for $\boldsymbol{\theta}_i$ is $p(\boldsymbol{\theta}_i)$ and is assumed known. If we can find a set of values of the $\boldsymbol{\theta}$'s so that for each binary decision between $\mathcal{H}_i$ and $\mathcal{H}_j$ we have

$$p_{X|\Theta_i,\,\mathcal{H}_i}(\boldsymbol{x}|\boldsymbol{\theta}_i^*) = p_{X|\Theta_j,\,\mathcal{H}_j}(\boldsymbol{x}|\boldsymbol{\theta}_j^*) \tag{13}$$

then the MAP rule for choosing among $\{\mathcal{H}_1, \mathcal{H}_2, \cdots, \mathcal{H}_M\}$ is to make binary decisions (a total of $M-1$) between $\mathcal{H}_i$ and $\mathcal{H}_j$ and to decide $\mathcal{H}_i$ if

$$\frac{p_{T_i|\mathcal{H}_i}(\boldsymbol{t}_i)}{p_{T_i|\Theta_i,\,\mathcal{H}_i}(\boldsymbol{t}_i|\boldsymbol{\theta}_i^*)}\,P(\mathcal{H}_i) > \frac{p_{T_j|\mathcal{H}_j}(\boldsymbol{t}_j)}{p_{T_j|\Theta_j,\,\mathcal{H}_j}(\boldsymbol{t}_j|\boldsymbol{\theta}_j^*)}\,P(\mathcal{H}_j). \tag{14}$$

Proof: Consider the binary decision between $\mathcal{H}_i$ and $\mathcal{H}_j$. Then, the MAP rule chooses $\mathcal{H}_i$ if

$$p_{X|\mathcal{H}_i}(\boldsymbol{x})P(\mathcal{H}_i) > p_{X|\mathcal{H}_j}(\boldsymbol{x})P(\mathcal{H}_j) \tag{15}$$

where

$$p_{X|\mathcal{H}_i}(\boldsymbol{x}) = \int p_{X|\Theta_i,\,\mathcal{H}_i}(\boldsymbol{x}|\boldsymbol{\theta}_i)p(\boldsymbol{\theta}_i)\,d\boldsymbol{\theta}_i.$$

But from the Bayesian equivalent of (10)

$$\begin{aligned}p_{X|\mathcal{H}_i}(\boldsymbol{x}) &= \int \frac{p_{T|\Theta_i,\,\mathcal{H}_i}(\boldsymbol{t}_i|\boldsymbol{\theta}_i)}{p_{T|\Theta_i,\,\mathcal{H}_i}(\boldsymbol{t}_i|\boldsymbol{\theta}_i^*)}\,p_{X|\Theta_i,\,\mathcal{H}_i}(\boldsymbol{x}|\boldsymbol{\theta}_i^*)p(\boldsymbol{\theta}_i)\,d\boldsymbol{\theta}_i\\ &= \int p_{T|\Theta_i,\,\mathcal{H}_i}(\boldsymbol{t}_i|\boldsymbol{\theta}_i)p(\boldsymbol{\theta}_i)\,d\boldsymbol{\theta}_i\,\frac{p_{X|\Theta_i,\,\mathcal{H}_i}(\boldsymbol{x}|\boldsymbol{\theta}_i^*)}{p_{T|\Theta_i,\,\mathcal{H}_i}(\boldsymbol{t}_i|\boldsymbol{\theta}_i^*)}\\ &= p_{T|\mathcal{H}_i}(\boldsymbol{t}_i)\,\frac{p_{X|\Theta_i,\,\mathcal{H}_i}(\boldsymbol{x}|\boldsymbol{\theta}_i^*)}{p_{T|\Theta_i,\,\mathcal{H}_i}(\boldsymbol{t}_i|\boldsymbol{\theta}_i^*)}.\end{aligned}$$

Using (15) we have

$$p_{T|\mathcal{H}_i}(\boldsymbol{t}_i)\,\frac{p_{X|\Theta_i,\,\mathcal{H}_i}(\boldsymbol{x}|\boldsymbol{\theta}_i^*)}{p_{T|\Theta_i,\,\mathcal{H}_i}(\boldsymbol{t}_i|\boldsymbol{\theta}_i^*)}\,P(\mathcal{H}_i) > p_{T|\mathcal{H}_j}(\boldsymbol{t}_j)\,\frac{p_{X|\Theta_j,\,\mathcal{H}_j}(\boldsymbol{x}|\boldsymbol{\theta}_j^*)}{p_{T|\Theta_j,\,\mathcal{H}_j}(\boldsymbol{t}_j|\boldsymbol{\theta}_j^*)}\,P(\mathcal{H}_j)$$

and, finally, using (13), we have the desired result.

It is sufficient but not necessary for the condition of (13) to hold that

$$p_{X|\Theta_1,\,\mathcal{H}_1}(\boldsymbol{x}|\boldsymbol{\theta}_1^*) = p_{X|\Theta_2,\,\mathcal{H}_2}(\boldsymbol{x}|\boldsymbol{\theta}_2^*) = \cdots = p_{X|\Theta_M,\,\mathcal{H}_M}(\boldsymbol{x}|\boldsymbol{\theta}_M^*).$$

V. Discussion and Conclusions

A more rigorous proof and justification for the class-specific feature theorem has been given. The importance of the result is that in classification problems one need not "lump" all the features for all the classes together. Doing so requires the determination of a large dimensionality pdf, which in practice is usually impossible. Alternatively, we can restrict attention to the sufficient statistic for each class separately, when appropriately normalized. Then, it is only the pdf of the sufficient statistic for each class that is required. However, the normalization condition cannot always be satisfied. In the exponential versus log-normal pdf classification problem this condition does not appear to be satisfied. For signal in noise problems, such as was illustrated using the linear model in Section III, the normalization condition is easily satisfied. We need only choose the parameters to result in the noise-only case.

In summary, when applicable, the class-specific feature approach results in a much simpler pdf estimation problem and a subsequent reduction in the complexity of the classification problem. It thus yields more accurate classification and/or requires less training data [2].

Acknowledgment

The author wishes to thank P. Baggenstoss for useful discussions relating to his work.

REFERENCES

[1] A. C. Atkinson, "A method for discriminating between models," *J. Roy. Statist. Soc.*, vol. 32, pp. 323–353, 1970.

[2] P. Baggenstoss, "Class-specific feature sets in classification," *IEEE Trans. Signal Processing*, vol. 47, pp. 3428–3432, Dec. 1999.

[3] S. Kay, *Fundamentals of Statistical Signal Processing, Vol. I: Estimation Theory*. Englewood Cliffs, NJ: Prentice-Hall, 1993, Appendix 5A.

[4] ——, *Fundamentals of Statistical Signal Processing, Vol. 2: Detection Theory*. Upper Saddle River, NJ: Prentice-Hall, 1998.

[5] M. Kendall and A. Stuart, *The Advanced Theory of Statistics, Vol. II*. New York: Macmillan, 1977.

[6] S. Kullback, *Information Theory and Statistics*. New York: Dover, 1968, pp. 19–20.

[7] D. W. Scott, *Multivariate Density Estimation*. New York: Wiley, 1992.

[8] S. Zacks, *The Theory of Statistical Inference*. New York: Wiley, 1971.

Asymptotic Performance Analysis of Bayesian Target Recognition

Ulf Grenander, Anuj Srivastava, *Member, IEEE*, and Michael I. Miller

***Abstract*—This correspondence investigates the asymptotic performance of Bayesian target recognition algorithms using deformable-template representations. Rigid computer-aided design (CAD) models represent the underlying targets; low-dimensional matrix Lie-groups (rotation and translation) extend them to particular instances. Remote sensors observing the targets are modeled as projective transformations, converting three-dimensional scenes into random images. Bayesian target recognition corresponds to hypothesis selection in the presence of nuisance parameters; its performance is quantified as the Bayes' error. Analytical expressions for this error probability in small noise situations are derived, yielding asymptotic error rates for exponential error probability decay.**

***Index Terms*—Bayesian ATR, deformable templates, Laplace's asymptotics, nuisance integration.**

I. INTRODUCTION

A variety of civilian and military applications require recognizing targets of interest, either stationary or moving, situated in unknown surroundings, using standard remote sensors such as cameras and radars. The data collected by sensors are analyzed by computer algorithms for *detection* and *recognition* of the targets in the observed scene. This data collection and the algorithmic inference together form an automated target recognition (ATR) system. An inherent part of any recognition system description are its performance specifications. In view of the diverse recognition algorithms proposed (please refer to the special issue of IEEE TRANSACTIONS ON IMAGE PROCESSING [19]) and the advancing sensor technology, resulting in new modalities and better performance, the need for methods of evaluating recognition systems is becoming more acute. Several target recognition performance studies have been presented in the literature. Lower and upper bounds, on the performance in localization and recognition of flexible shapes, are presented in [16]. Agarwal *et al.* [1] provide an overview of the target recognition performance under three popular paradigms: Bayesian, neural networks, and rule-based approaches. In [12], the authors evaluate the performance of an optimum receiver designed for a noise-free target in terms of unknown rotation and scale parameters. A variety of other researchers have reported analyses of target recognition performance, but restricted to specific systems or algorithms [8], [18]. In [7], the author provides a broad analysis of several target recognition algorithms in terms of their performance.

The goal of this correspondence is to present a quantitative analysis on the asymptotic performance of the recognition systems, not limited to any particular model. We shall use a Bayesian pattern-theoretic framework as in [17], [26], and [27] in which the recognition problem is treated in a unified way, so that, for example, multitarget and multisensor situations do not require separate treatment, nor do obscuration, structured clutter, and tactical aims; they are all instances of the same general recognition problem. In fact, detection, tracking, and recognition are all treated as partial solutions of a single Bayesian problem. Given precise models of the shapes and reflectance characteristics of the targets of interest, and the physics of remote sensors, the observed variability is reduced to low-dimensional attributes such as pose, motion, illumination, temperature, in addition to the target labels. The variability in these targets-attributes is modeled by group actions (rotation and translation) on the rigid templates and, hence, target inference reduces to optimization over the these groups ($\mathbb{R}^n$ and $SO(n)$). For modeling image formation via remote sensing, we will assume statistical models motivated by the physics of sensor operation. The resulting images are random realizations with means given by the projections of the three-dimensional scenes.

To recognize a target, estimation of the associated target attributes, such as pose, motion, lighting, and thermal profile, becomes essential. Target recognition is performed through Bayesian hypothesis testing; for a given observation the likelihood ratios are compared to the ratio of priors and a hypothesis is selected. In a binary case, for an observed image I^D, the Bayesian hypothesis testing problem is

$$p(I^D|H_1)/p(I^D|H_0) \underset{H_0}{\overset{H_1}{\gtrless}} P(H_0)/P(H_1) \equiv \nu.$$

In the presence of nuisance parameters, such as pose and location, $p(I^D|H_i),\ i = 0,\ 1$ is defined via the integral

$$p(I^D|H_i) = \int_S p(I^D|s,\ H_i)p(s|H_i)\gamma(ds)$$

where s is a nuisance parameter. In most practical situations, the integrand is too complicated to be computed analytically. One common solution is to formulate a generalized likelihood-ratio test or *pseudo*-likelihood test [29] according to the rule

$$p(I^D|H_1,\ s_1^*)/p(I^D|H_0,\ s_0^*) \underset{<}{\overset{>}{}} \nu$$

for some ν and

$$s_i^* = \arg\max_{s\in S} p(I^D|H_i,\ s).$$

s_i^*'s are the maximum-likelihood estimates (MLE's), of the unknown target parameter s, under the two hypotheses. A similar ratio, of the maximized posterior densities under the two hypotheses, is proposed

Manuscript received October 16, 1998; revised February 4, 2000. This work was supported in part by Grants ARO DAAH04-95-1-0494, ARO-MURI DAAH04-96-1-0445, ARO DAA-G55-98-1-0102, ARO DAAD19-99-0267, and NSF-EIA 9871196.

U. Grenander is with the Division of Applied Mathematics, Brown University, Providence, RI 02912 USA (e-mail: ulf_grenander@Brown.edu).

A. Srivastava is with the Department of Statistics, Florida State University, Tallahassee, FL 32306 USA (e-mail: anuj@stat.fsu.edu).

M. I. Miller is with the Department of Electrical and Computer Engineering, The Johns Hopkins University, Baltimore, MD 21218 USA (e-mail: mim@cis.jhu.edu).

Communicated by P. Moulin, Associate Editor for Nonparametric Estimation, Classification, and Neural Networks.

Publisher Item Identifier S 0018-9448(00)05014-8.