Ann. Inst. Statist. Math.
Vol. 52, No. 3, 426–437 (2000)

ASYMPTOTIC DISTRIBUTION OF ESTIMATED AFFINITY BETWEEN MULTIPARAMETER EXPONENTIAL FAMILIES*

STEVEN T. GARREN

Department of Statistics, University of Virginia, 104 Halsey Hall, Charlottesville, VA 22903, U.S.A.

(Received September 1, 1998; revised February 8, 1999)

Abstract. Let $F_1, \ldots, F_J$ be the distributions of J independent multiparameter exponential families, and $\rho_J(F_1, \ldots, F_J)$ denote the affinity between $F_1, \ldots, F_J$. We consider the problem of estimating ρ_J on the basis of independent random samples from these distributions. Subject to some mild regularity conditions, we derive the asymptotic distribution of the maximum likelihood estimator of ρ_J. Applications to hypothesis testing and discriminant analysis are discussed, and an example is provided.

Key words and phrases: Affinity, asymptotic distribution, distance measure, exponential family.

1. Introduction

Let $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$ be random variables defined on the set Ω. Assume their distribution functions are F_1 and F_2, and the probability density functions are f_1 and f_2, respectively, with respect to the common measure m. Matusita (1955, 1966) defined a distance measure between two distributions F_1 and F_2 to be

$$\xi(F_1, F_2) = \left[\int_\Omega \{\sqrt{f_1(\boldsymbol{x})} - \sqrt{f_2(\boldsymbol{x})}\}^2 dm(\boldsymbol{x})\right]^{1/2}.$$

He also defined the affinity between F_1 and F_2 to be

$$\rho_2(F_1, F_2) = \int_\Omega [f_1(\boldsymbol{x}) f_2(\boldsymbol{x})]^{1/2} dm(\boldsymbol{x}),$$

and remarked that $\xi^2(F_1, F_2) = 2[1 - \rho_2(F_1, F_2)]$. Note that $0 \leq \rho_2(F_1, F_2) \leq 1$, and $\rho_2(F_1, F_2) = 1$ if and only if $F_1 = F_2$. Furthermore, there is a one-to-one mapping between $\xi(F_1, F_2)$ and $\rho_2(F_1, F_2)$. Matusita (1967*b*, 1973) generalized the notion of affinity to an arbitrary number of distributions $F_1, \ldots, F_J$ by defining the affinity,

$$\rho_J(F_1, F_2, \ldots, F_J) = \int_\Omega [f_1(\boldsymbol{x}) f_2(\boldsymbol{x}) \cdots f_J(\boldsymbol{x})]^{1/J} dm(\boldsymbol{x}). \tag{1.1}$$

Matusita (1966) determined an analytical formula for the affinity between two multivariate normal distributions. He derived the exact distribution of the estimated affinity, using sample means and variances to estimate the unknown parameters based on independent samples from the two populations. He produced a test statistic, based on a

* Research partially supported by NIMH grant MH53259-01A2.

sample analog of the affinity, for testing equality of the unknown parameters from the two multivariate normal distributions.

Matusita (1955) used the affinity to determine confidence bands on discrete distributions. Matusita (1967*b*) also used the affinity to test whether or not multiple sets of observations have the same distribution, and he applied these findings to multivariate normal data. Matusita (1967*a*) tested whether or not a set of observations is from one of several known distributions, and again applied his findings to multivariate normal data. These results were generalized by Matusita (1971) by allowing the known distributions to be selected according to prior probabilities.

McLachlan ((1992), pp. 22–26, 220–229) provided a brief literary review of affinity-based measures. LeCam (1970) and Beran (1977) referred to the affinity between two distributions as the Hellinger distance. Discrimination among populations containing a mixture of discrete and continuous data was discussed by Krzanowski (1983, 1984, 1987).

Bar-Hen and Daudin (1998) studied the asymptotic distribution of affinity in relation to the location model (Olkin and Tate (1961); Krzanowski (1975)). This model is a mixture of r multivariate normal distributions with different mean vectors but a common covariance matrix, where the mixing probabilities are multinomial, and all parameters are unknown. The distribution of affinity for $r = 1$ was analyzed in detail by Matusita (1966). When $r \geq 2$, this location model is not in the exponential family and is not examined herein.

In this paper we examine the case of arbitrary multiparameter exponential families, and measure the affinity when the distributions are the same except for the unknown parameters. The asymptotic distribution of the estimated affinity when the parameters are not equal is shown to be univariate normal, and we derive explicit formulas for the asymptotic variance using the delta method. The estimated affinity is shown to be invariant under one-to-one transformations of the parameters or random variables. We use (1.1) when examining the affinity in the exponential family. For example, we may be interested in measuring the distance between two independent Poisson distributions with unknown means, which may be unequal in general. When discussing inferences for hypothesis testing and discriminant analysis involving two distributions in Section 5, we will set $J = 2$ in (1.1). A brief discussion regarding practical issues in Section 7 concludes the paper. Proofs of theorems are provided in the Appendix.

2. Invariance of affinity

Suppose the random variable $\boldsymbol{X}$ is transformed via a one-to-one mapping from $\boldsymbol{X}$ to $g(\boldsymbol{X})$, such that $g(\boldsymbol{X})$ is continuously differential and the Jacobian of $g(\boldsymbol{X})$ does not vanish. Then, ρ_J is the same under either $\boldsymbol{X}$ or $g(\boldsymbol{X})$ as shown below. The same is true for estimates of ρ_J. Observe that

$$f_j(\boldsymbol{x}) = f_j(g(\boldsymbol{x}))\|\partial g(\boldsymbol{x})/\partial \boldsymbol{x}\|, \qquad j = 1, \ldots, J, \tag{2.1}$$

where $|\partial g(\boldsymbol{x})/\partial \boldsymbol{x}|$ is the Jacobian. The invariance result for ρ_J is fairly obvious, since (1.1) and (2.1) imply that

$$\begin{aligned}
\rho_J(F_1, \ldots, F_J) &= \int_\Omega [f_1(\boldsymbol{x}) \cdots f_J(\boldsymbol{x})]^{1/J} dm(\boldsymbol{x}) \\
&= \int_\Omega [f_1(g(\boldsymbol{x})) \cdots f_J(g(\boldsymbol{x}))]^{1/J} \|\partial g(\boldsymbol{x})/\partial \boldsymbol{x}\| dm(\boldsymbol{x}) \\
&= \int_{g(\Omega)} [f_1(g(\boldsymbol{x})) \cdots f_J(g(\boldsymbol{x}))]^{1/J} dm(g(\boldsymbol{x})).
\end{aligned}$$

For example, if X has a normal distribution, then $\exp\{X\}$ has a log-normal distribution, and the affinity is invariant under this transformation. The same reasoning suggests that estimates of ρ_J (when parameters in $F_1, \ldots, F_J$ are unknown) also are invariant under the transformation from $\boldsymbol{X}$ to $g(\boldsymbol{X})$.

We now examine reparametrization. Since ρ_J as defined in (1.1) does not depend on which parametrization of the unknown parameters is chosen, then estimates of ρ_J also do not depend on the parametrization. Thus, when studying the exponential family, we can use natural parameters without loss of generality.

PROPOSITION 2.1. *Both the value of ρ_J and the distribution of any estimate of ρ_J are invariant under the transformation from the random variable $\boldsymbol{X}$ to $g(\boldsymbol{X})$, and under any reparametrization of the unknown parameters.*

3. Exponential families

We now provide the notation for multiparameter exponential families (cf. Bickel and Doksum (1977), pp. 71–73) with K unknown parameters, for J distributions. Denote the *natural parameter* of F_j by $\boldsymbol{\eta}_j = (\eta_{j1}, \ldots, \eta_{jK})'$, for $j = 1, \ldots, J$, and denote the *natural statistic* by $\boldsymbol{T}(\boldsymbol{x}) = (T_1(\boldsymbol{x}), \ldots, T_K(\boldsymbol{x}))'$. The multivariate distribution function F_j has probability density function

$$f(\boldsymbol{x}, \boldsymbol{\eta}_j) = \exp\left\{\sum_{k=1}^{K} \eta_{jk} T_k(\boldsymbol{x}) + q(\boldsymbol{\eta}_j) + s(\boldsymbol{x})\right\} 1_B(\boldsymbol{x}), \tag{3.1}$$

for $j = 1, \ldots, J$, where $s(\boldsymbol{x})$ and the indicator function, $1_B(\boldsymbol{x})$, do not depend on $\boldsymbol{\eta}_j$, and $q(\boldsymbol{\eta}_j)$ is some function. Hence, the distributions $F_1, \ldots, F_J$ are similar in functional form except for the unknown parameters. The affinity under model (3.1) can be shown to be

$$\begin{aligned} \rho_J(F_1, \ldots, F_J) &= \exp\left\{J^{-1} \sum_{j=1}^{J} q(\boldsymbol{\eta}_j)\right\} \int_B \exp\left\{J^{-1} \sum_{j=1}^{J} \sum_{k=1}^{K} \eta_{jk} T_k(\boldsymbol{x}) + s(\boldsymbol{x})\right\} d\boldsymbol{x} \\ &= \exp\left\{J^{-1} \sum_{j=1}^{J} q(\boldsymbol{\eta}_j) - q\left(J^{-1} \sum_{j=1}^{J} \boldsymbol{\eta}_j\right)\right\}. \end{aligned} \tag{3.2}$$

Let J^* be the number of distributions estimated. Also, let $\hat{\boldsymbol{\eta}}_j$ denote the maximum likelihood estimator (MLE) of $\boldsymbol{\eta}_j$, based on N_j independent observations $\boldsymbol{x}_{j1}, \ldots, \boldsymbol{x}_{jN_j}$ from F_j, $j = 1, \ldots, J^*$. Replacing $\boldsymbol{\eta}_j$ by $\hat{\boldsymbol{\eta}}_j$ and F_j by $\hat{F}_j$ in (3.2), we obtain the MLE of ρ_J. It follows that

$$\log \frac{\rho_J(\hat{F}_1, \ldots, \hat{F}_J)}{\rho_J(F_1, \ldots, F_J)} = J^{-1} \sum_{j=1}^{J} [q(\hat{\boldsymbol{\eta}}_j) - q(\boldsymbol{\eta}_j)] - q\left(J^{-1} \sum_{j=1}^{J} \hat{\boldsymbol{\eta}}_j\right) + q\left(J^{-1} \sum_{j=1}^{J} \boldsymbol{\eta}_j\right), \tag{3.3}$$

and the asymptotic distribution of (3.3) will be derived in Section 4, when standardized by sample size.

4. Asymptotic distribution

The asymptotic distribution of (3.3) now is stated. The gradient and Hessian of any real function $\zeta(\boldsymbol{y})$ will be denoted by $\nabla\zeta(\boldsymbol{y})$ and $\nabla^2\zeta(\boldsymbol{y})$, respectively. If $\boldsymbol{y}$ is a function of $\boldsymbol{\eta}$, then these derivatives for computing the gradient and Hessian of $\zeta(\boldsymbol{y}(\boldsymbol{\eta}))$ still are taken with respect to $\boldsymbol{y}$.

Consider the following mild regularity conditions.

(C1) $\nabla^2 q(\boldsymbol{\eta})$ is bounded in a single neighborhood containing $\boldsymbol{\eta}_j$, for all $j = 1, \ldots, J$.

(C2) There exists a continuous function $\boldsymbol{h} : \Re^K \to \Re^K$ such that for all $j = 1, \ldots, J^*$, $\hat{\boldsymbol{\eta}}_j = \boldsymbol{h}(N_j^{-1} \sum_{n=1}^{N_j} \boldsymbol{T}(\boldsymbol{x}_{jn}))$, and $\boldsymbol{\eta} = \boldsymbol{h}(-\nabla q(\boldsymbol{\eta}))$ for all $\boldsymbol{\eta}$.

(C3) There exists an N, depending on $N_1, \ldots, N_{J^*}$, such that $N_j/N \to 1$ as $N \to \infty$ for all $j = 1, \ldots, J^*$.

When $J^* = 1$ in condition (C3), let N be the sample size. Now define

$$V_J(\boldsymbol{\eta}_1, \ldots, \boldsymbol{\eta}_J) = \sum_{j=1}^{J} \left[\nabla q(\boldsymbol{\eta}_j) - \nabla q\left(J^{-1}\sum_{l=1}^{J}\boldsymbol{\eta}_l\right)\right]' \boldsymbol{A}(\boldsymbol{\eta}_j) \left[\nabla q(\boldsymbol{\eta}_j) - \nabla q\left(J^{-1}\sum_{l=1}^{J}\boldsymbol{\eta}_l\right)\right] \tag{4.1}$$

where

$$\boldsymbol{A}(\boldsymbol{\eta}) = [\nabla \boldsymbol{h}(-\nabla q(\boldsymbol{\eta}))]'[-\nabla^2 q(\boldsymbol{\eta})]\nabla \boldsymbol{h}(-\nabla q(\boldsymbol{\eta})).$$

Notice that $\boldsymbol{A}(\boldsymbol{\eta})$ is nonnegative definite since $[-\nabla^2 q(\boldsymbol{\eta})]$ is the covariance matrix of $\boldsymbol{T}(\boldsymbol{x})$. The expression V_J, as defined in (4.1), will appear in asymptotic distributions involving the estimated affinity.

THEOREM 4.1. *Assume the regularity conditions* (C1)–(C3).

(a) *Then*

$$J\sqrt{N} \log \frac{\rho_J(\hat{F}_1, \ldots, \hat{F}_J)}{\rho_J(F_1, \ldots, F_J)} \xrightarrow{d} \mathcal{N}(0, V_J(\boldsymbol{\eta}_1, \ldots, \boldsymbol{\eta}_J))$$

as $N \to \infty$.

(b) *Suppose* $F_1 = \cdots = F_J$, *and all third derivative components of* $q(\boldsymbol{\eta})$ *are bounded in a neighborhood about* $\boldsymbol{\eta}_1$. *Then,*

$$-2NJ \log \rho_J(\hat{F}_1, \ldots, \hat{F}_J) \xrightarrow{d} \sum_{k=1}^{K} \delta_k Y_k, \tag{4.2}$$

as $N \to \infty$, *where* $Y_1, \ldots, Y_K$ *are mutually independent* χ^2_{J-1} *random variables, and the* $\delta_1, \ldots, \delta_K$ *are the eigenvalues of the nonnegative definite matrix*

$$[-\nabla^2 q(\boldsymbol{\eta}_1)]^{1/2} \boldsymbol{A}(\boldsymbol{\eta}_1) [-\nabla^2 q(\boldsymbol{\eta}_1)]^{1/2}.$$

Note that the added condition on $q(\boldsymbol{\eta})$ in Theorem 4.1(b) strengthens condition (C1). Kotz *et al.* (1967) derived series expansions for the distribution function and the probability density function of the right hand side of (4.2), so this asymptotic distribution is well understood.

5. Inferences

In this section we assume that conditions (C1)–(C3) hold with $J = 2$. The notion of affinity in exponential models is applied to hypothesis testing and discriminant analysis. Assume that the exponential distributions are the same except for the parameter $\boldsymbol{\eta}$.

5.1 *Simple vs. simple hypothesis testing*

Suppose one is interested in testing the null hypothesis that $\boldsymbol{\eta} = \boldsymbol{\eta}_1$ against the alternative that $\boldsymbol{\eta} = \boldsymbol{\eta}_2$, where $\boldsymbol{\eta}$ is the unknown true value of the parameter, and $\boldsymbol{\eta}_1$ and $\boldsymbol{\eta}_2$ are fixed. Exponential distributions with parameters $\boldsymbol{\eta}$, $\boldsymbol{\eta}_1$, and $\boldsymbol{\eta}_2$ are denoted by F, F_1, and F_2, respectively. For this subsection take $J^* = 1$, such that $\hat{\boldsymbol{\eta}}$ is the MLE of $\boldsymbol{\eta}$ based on the N observations from F. The proposed test statistic is

$$Q(\hat{\boldsymbol{\eta}}) = \log \frac{\rho_2(\hat{F}, F_2)}{\rho_2(\hat{F}, F_1)} = \frac{q(\boldsymbol{\eta}_2) - q(\boldsymbol{\eta}_1)}{2} + q\left(\frac{\hat{\boldsymbol{\eta}} + \boldsymbol{\eta}_1}{2}\right) - q\left(\frac{\hat{\boldsymbol{\eta}} + \boldsymbol{\eta}_2}{2}\right), \tag{5.1}$$

and the null hypothesis is rejected when $Q(\hat{\boldsymbol{\eta}})$ is sufficiently large. Part (a) of the following theorem shows that $Q(\hat{\boldsymbol{\eta}})$ is a reasonable test statistic, and part (b) provides the asymptotic distribution of $Q(\hat{\boldsymbol{\eta}})$.

THEOREM 5.1. *Assume the regularity conditions* (C1)–(C3), *where* $J = 2$ *and* $J^* = 1$. *Let* $\boldsymbol{\eta}_1$ *and* $\boldsymbol{\eta}_2$ *be fixed.*

(a) *Then* $Q(\hat{\boldsymbol{\eta}})$ *monotonically increases as* $\hat{\boldsymbol{\eta}}$ *moves along the line* $(\boldsymbol{\eta}_2 - \boldsymbol{\eta}_1)$ *in the direction of* $(\boldsymbol{\eta}_2 - \boldsymbol{\eta}_1)$.

(b) *Under* F_1,

$$2\sqrt{N}[Q(\hat{\boldsymbol{\eta}}) - \log \rho(F_1, F_2)] \xrightarrow{d} \mathcal{N}\left(0, \left\{\nabla q(\boldsymbol{\eta}_1) - \nabla q\left(\frac{\boldsymbol{\eta}_1 + \boldsymbol{\eta}_2}{2}\right)\right\}' \boldsymbol{A}(\boldsymbol{\eta}_1) \left\{\nabla q(\boldsymbol{\eta}_1) - \nabla q\left(\frac{\boldsymbol{\eta}_1 + \boldsymbol{\eta}_2}{2}\right)\right\}\right)$$

as $N \to \infty$. *Also, under* F_2,

$$2\sqrt{N}[Q(\hat{\boldsymbol{\eta}}) + \log \rho(F_1, F_2)] \xrightarrow{d} \mathcal{N}\left(0, \left\{\nabla q(\boldsymbol{\eta}_2) - \nabla q\left(\frac{\boldsymbol{\eta}_1 + \boldsymbol{\eta}_2}{2}\right)\right\}' \boldsymbol{A}(\boldsymbol{\eta}_2) \left\{\nabla q(\boldsymbol{\eta}_2) - \nabla q\left(\frac{\boldsymbol{\eta}_1 + \boldsymbol{\eta}_2}{2}\right)\right\}\right)$$

as $N \to \infty$.

Theorem 5.1(a) suggests that if the dimension of the unknown parameter is one, then $Q(\hat{\eta})$ is a monotone function of $\hat{\eta}$, which is a monotone function of the complete, sufficient statistic $\sum_{n=1}^{N} T(x_n)$. By the Neyman-Pearson Lemma (cf. Bickel and Doksum (1977), pp. 192–194) we have proved the following result.

PROPOSITION 5.1. *For a one-dimensional exponential family with unknown parameter* η, *the test statistic* $Q(\hat{\eta})$ *produces the uniformly most powerful test for testing the null hypothesis that* $\eta = \eta_1$ *against the alternative that* $\eta = \eta_2$, *where* $\hat{\eta}$ *is the MLE of* η.

5.2 *Two-sided hypothesis testing*

Suppose one wishes to test the null hypothesis that $\eta = \eta^*$ against the alternative that $\eta \neq \eta^*$, where η is the unknown true value of the parameter. Denote by F the exponential distribution with *unknown* parameter η, and denote by F^* the exponential distribution with *known* parameter η^*. For this subsection take $J^* = 1$, such that $\hat{\eta}$ is the MLE of η based on the N observations from F. Under the null hypothesis, we will add the condition that all third derivative components of $q(\eta)$ are bounded in a neighborhood about η^*. For testing these hypotheses, (3.3) becomes

$$\log \frac{\rho_2(\hat{F}, F^*)}{\rho_2(F, F^*)} = \frac{q(\hat{\eta}) + q(\eta^*)}{2} - q\left(\frac{\hat{\eta} + \eta^*}{2}\right). \tag{5.2}$$

This test statistic (5.2) has the following desirable property.

THEOREM 5.2. *For a fixed null value η^*, the test statistic* (5.2) *monotonically decreases as $\hat{\eta}$ moves away from η^* along the line $(\hat{\eta} - \eta^*)$.*

Hypothesis tests may be performed by noting that under the null hypothesis

$$\log \rho_2(\hat{F}, F^*) = \frac{1}{8}(\hat{\eta} - \eta^*)'\{\nabla^2 q(\eta^*)\}(\hat{\eta} - \eta^*) + O_p(N^{-3/2})$$

as $N \to \infty$. In a similar spirit as Theorem 4.1(b), a Taylor series expansion implies that

$$-8N \log \rho_J(\hat{F}, F^*) \xrightarrow{d} \sum_{k=1}^{K} \delta_k Y_k \tag{5.3}$$

as $N \to \infty$, where $Y_1, \ldots, Y_K$ are independent χ_1^2 random variables, and $\delta_1, \ldots, \delta_K$ are the eigenvalues of the nonnegative definite matrix

$$\{A(\eta_1)\}^{1/2}\{-\nabla^2 q(\eta_1)\}\{A(\eta_1)\}^{1/2}.$$

Power calculations may be performed by noting that under the alternative hypothesis

$$2\sqrt{N} \log \frac{\rho_2(\hat{F}, F^*)}{\rho_2(F, F^*)} \tag{5.4}$$
$$\xrightarrow{d} \mathcal{N}\left(0, \left\{\nabla q(\eta) - \nabla q\left(\frac{\eta + \eta^*}{2}\right)\right\}' A(\eta) \left\{\nabla q(\eta) - \nabla q\left(\frac{\eta + \eta^*}{2}\right)\right\}\right)$$

as $N \to \infty$, in the same spirit as Theorem 4.1(a).

Theorem 5.2 suggests that the test statistic (5.2), although valid for two-sided tests, is not valid for one-sided tests. For example, in the univariate case suppose one is testing the null hypothesis that $\eta = \eta^*$ against the alternative that $\eta > \eta^*$. If $\hat{\eta} \ll \eta^*$, then the null hypothesis is likely to be erroneously rejected. The probability of committing a Type I error would be much larger than the nominal value.

5.3 *Discriminant analysis*

Suppose N_j observations are taken from exponential distribution F_j with unknown parameter η_j for $j = 1, \ldots, J^*$, where $J^* = 3$. Assume that either F_3 is the same as F_1, or F_3 is the same as F_2. In other words, one wishes to determine whether the N_3 observations are from F_1 or F_2. Assume that F_1 and F_2 have equal *a priori* distributions.

The classification will be based on affinity. The data will be classified as being from F_1 if $\rho_2(\hat{F}_1, \hat{F}_3) > \rho_2(\hat{F}_2, \hat{F}_3)$, and will be classified as being from F_2 if $\rho_2(\hat{F}_1, \hat{F}_3) < \rho_2(\hat{F}_2, \hat{F}_3)$.

The asymptotic probability of misclassification now is derived. Let

$$U(\hat{F}_1, \hat{F}_2, \hat{F}_3) = \log \frac{\rho_2(\hat{F}_2, \hat{F}_3)}{\rho_2(\hat{F}_1, \hat{F}_3)} = \frac{q(\hat{\eta}_2) - q(\hat{\eta}_1)}{2} + q\left(\frac{\hat{\eta}_1 + \hat{\eta}_3}{2}\right) - q\left(\frac{\hat{\eta}_2 + \hat{\eta}_3}{2}\right). \tag{5.5}$$

Thus, the N_3 observations from F_3 are classified as being from F_1 if $U < 0$, and are classified as being from F_2 if $U > 0$. A straightforward calculation shows that

$$U(\hat{F}_1, \hat{F}_2, \hat{F}_3) \xrightarrow{a.s.} \begin{cases} \log \rho_2(F_1, F_2), & \text{under } F_1 \\ -\log \rho_2(F_1, F_2), & \text{under } F_2 \end{cases} \tag{5.6}$$

as $\min\{N_1, N_2, N_3\} \to \infty$. Select N as in condition (C3). The asymptotic distribution of $U(\hat{F}_1, \hat{F}_2, \hat{F}_3)$ is stated here.

THEOREM 5.3. *Assume that the regularity conditions* (C1)–(C3) *hold, where* $J = 2$ *and* $J^* = 3$. *Also, assume that* η_1, η_2, *and* η_3 *are unknown. If* $F_3 = F_1$, *then*

$$2[U(\hat{F}_1, \hat{F}_2, \hat{F}_3) - \log \rho_2(F_1, F_2)]\sqrt{N} \xrightarrow{d} \mathcal{N}(0, V_2(\eta_1, \eta_2))$$

as $N \to \infty$. *Also, if* $F_3 = F_2$, *then*

$$2[U(\hat{F}_1, \hat{F}_2, \hat{F}_3) + \log \rho_2(F_1, F_2)]\sqrt{N} \xrightarrow{d} \mathcal{N}(0, V_2(\eta_1, \eta_2))$$

as $N \to \infty$.

When the true distribution is F_1, the probability of misclassification is

$$\begin{aligned} &P(U(\hat{F}_1, \hat{F}_2, \hat{F}_3) > 0 \mid F_1) \\ &\quad = P\left(\frac{2\{U(\hat{F}_1, \hat{F}_2, \hat{F}_3) - \log \rho_2(F_1, F_2)\}}{\sqrt{V_2(\eta_1, \eta_2)/N}} > -\frac{2 \log \rho_2(F_1, F_2)}{\sqrt{V_2(\eta_1, \eta_2)/N}}\right) \\ &\quad \simeq P\left(Z > -\frac{2 \log \rho_2(F_1, F_2)}{\sqrt{V_2(\eta_1, \eta_2)/N}}\right) = \Phi\left(\frac{2 \log \rho_2(F_1, F_2)}{\sqrt{V_2(\eta_1, \eta_2)/N}}\right), \end{aligned} \tag{5.7}$$

for large N, where Z is a standard normal random variable and $\Phi(\cdot)$ is the standard normal distribution function. Similarly, when the true distribution is F_2, the probability of misclassification also is approximately the right hand side of (5.7) for large N. Therefore, unconditional on which distribution, F_1 or F_2, is the true one, the probability of misclassification is approximately the right hand side of (5.7) for large N.

Consider an example involving univariate normal random variables. Let F_j denote a normal distribution with unknown mean η_j for $j = 1, \ldots, 3$ and known equal variances

σ^2. The natural parameter is η_j, and $q(\eta) = -\eta^2/(2\sigma^2)$. Defining $\bar{X}_j$ to be the sample mean based on the N_j observations from F_j, it follows from (5.5) that

$$4U(\hat{F}_1, \hat{F}_2, \hat{F}_3) = \left[\bar{X}_3 - \frac{1}{2}(\bar{X}_1 + \bar{X}_2)\right] \sigma^{-2}(\bar{X}_2 - \bar{X}_1). \tag{5.8}$$

Anderson ((1984), p. 210, eq. (5)) used the same criterion for determining misclassification as in (5.8), except he estimated σ^2, whereas we assume σ^2 is known.

Now suppose that $\boldsymbol{\eta}_1$ and $\boldsymbol{\eta}_2$ are known and $\boldsymbol{\eta}_3$ is unknown for arbitrary exponential distributions F_1, F_2, and F_3. Then $J^* = 1$, and a sample of size N is taken from F_3. The test statistic (5.5) becomes

$$U(F_1, F_2, \hat{F}_3) = \log \frac{\rho_2(\hat{F}_3, F_2)}{\rho_2(\hat{F}_3, F_1)} = Q(\hat{\boldsymbol{\eta}}_3), \tag{5.9}$$

which is statistic (5.1) used for testing simple vs. simple hypotheses. Hence, (5.9) already has been discussed in Section 5.1.

6. An example

As in Section 5 take $J = 2$. Consider a gamma random variable X with unknown scale parameter η and known shape parameter $c > 0$. The probability density function of X may be written

$$f(x, \eta) = \eta^c x^{c-1} e^{-\eta x}/\Gamma(c), \qquad x > 0.$$

The natural parameter is η; the natural statistic is $-X$; $q(\eta) = c \log \eta$; $h(y) = -c/y$, and $A(\eta) = \eta^2/c$. Mutually independent samples of size N_j are taken from gamma random variables X_j such that there exists an N as defined in condition (C3), for $j = 1, \ldots, J^*$. The MLE of η_j is $\hat{\eta}_j = c/\bar{X}_j$, where $\bar{X}_j$ is the sample mean from the j-th population of size N_j. It follows from (3.2) that

$$\rho_2(F_1, F_2) = \left(\frac{2\sqrt{\eta_1 \eta_2}}{\eta_1 + \eta_2}\right)^c. \tag{6.1}$$

We will apply two-sided hypothesis testing and discriminant analysis to these gamma random variables. The statistic for testing simple vs. simple hypotheses is equivalent to the Neyman-Pearson test statistic for this one-dimensional case, and therefore is not discussed further.

6.1 *Two-sided hypothesis testing*

In this subsection take $J^* = 1$, $\eta = \eta_1$, and $F = F_1$. To perform hypothesis tests and power calculations, (5.3) implies that

$$-8N \log \rho_2(\hat{F}, F^*) = -8cN \log \frac{2\sqrt{\hat{\eta}\eta^*}}{\hat{\eta} + \eta^*} \xrightarrow{d} \chi_1^2 \tag{6.2}$$

as $N \to \infty$ under the null hypothesis $\eta = \eta^*$. Under the alternative hypothesis $\eta \neq \eta^*$, (5.4) implies that

$$2\sqrt{N} \log \frac{\rho_2(\hat{F}, F^*)}{\rho_2(F, F^*)} = 2c\sqrt{N} \log \left\{ \left(\frac{\eta + \eta^*}{\hat{\eta} + \eta^*}\right) \sqrt{\frac{\hat{\eta}}{\eta}} \right\} \xrightarrow{d} \mathcal{N}\left(0, c\left\{\frac{\eta - \eta^*}{\eta + \eta^*}\right\}^2\right)$$

as $N \to \infty$. Therefore, the null hypothesis should be rejected for large values of the left hand side of (6.2), as compared to a χ_1^2-table.

6.2 *Discriminant analysis*

We now perform discriminant analysis for testing whether $F_3 = F_1$ or $F_3 = F_2$, where the distributions F_1 and F_2 have equal *a priori* probabilities and are unknown. In this subsection take $J^* = 3$ and assume that condition (C3) holds. It follows from (5.5) and (4.1) that

$$U(\hat{F}_1, \hat{F}_2, \hat{F}_3) = c \log \left\{ \left(\frac{\hat{\eta}_1 + \hat{\eta}_3}{\hat{\eta}_2 + \hat{\eta}_3} \right) \sqrt{\frac{\hat{\eta}_2}{\hat{\eta}_1}} \right\}$$

and

$$V_2(\eta_1, \eta_2) = 2c \left(\frac{\eta_1 - \eta_2}{\eta_1 + \eta_2} \right)^2 .$$

It then follows from the right hand side of (5.7) that for large N the probability of misclassification is approximately

$$\Phi \left(\left\{ \frac{\eta_1 + \eta_2}{|\eta_1 - \eta_2|} \right\} \sqrt{2cN} \log \left\{ \frac{2\sqrt{\eta_1 \eta_2}}{\eta_1 + \eta_2} \right\} \right).$$

7. Practical issues

The regularity conditions (C1)–(C3) are theoretically the only conditions which need to hold in order for the theorems herein to be valid. Note that condition (C2) requires the MLE of the natural parameter to be a function of the natural statistic. However, if this MLE is not tractable, then the asymptotic distribution of the estimated affinity cannot be expressed analytically.

For example, the MLE of the shape parameter of a gamma random variable is not tractable. On the other hand, the MLE of the scale parameter of a gamma random variable with known shape parameter *is* tractable, as mentioned in Section 6. Furthermore, the MLE of the natural parameter is tractable for the multivariate normal, log-normal, Poisson, and Bernoulli distributions, among many others, in which cases one may make practical inferences similar to those made in Section 6.

Acknowledgements

The author is very thankful to Donald Richards for useful conversations and references on this subject, and to Shyamal Peddada for a careful reading of the manuscript. The author is also thankful to an anonymous referee for many helpful suggestions.

Appendix: Proof of Theorems

LEMMA A.1. *Under the regularity conditions* (C1)–(C3),

$$(\hat{\boldsymbol{\eta}}_j - \boldsymbol{\eta}_j)\sqrt{N} \xrightarrow{d} \mathcal{N}(0, \boldsymbol{A}(\boldsymbol{\eta}_j))$$

as $N \to \infty$, $j = 1, \ldots, J^*$.

PROOF. By a Taylor series expansion,

$$\begin{aligned}\hat{\boldsymbol{\eta}}_j - \boldsymbol{\eta}_j &= \boldsymbol{h}\left(N_j^{-1}\sum_{n=1}^{N_j}\boldsymbol{T}(\boldsymbol{x}_{jn})\right) - \boldsymbol{h}(-\nabla q(\boldsymbol{\eta}_j)) \\ &= \boldsymbol{h}\left(N_j^{-1}\sum_{n=1}^{N_j}\boldsymbol{T}(\boldsymbol{x}_{jn})\right) - \boldsymbol{h}(E_{\boldsymbol{\eta}_j}\boldsymbol{T}(\boldsymbol{x}_j)) \\ &= \left[N_j^{-1}\sum_{n=1}^{N_j}\boldsymbol{T}(\boldsymbol{x}_{jn}) - E_{\boldsymbol{\eta}_j}\boldsymbol{T}(\boldsymbol{x}_j)\right]' \nabla\boldsymbol{h}(E_{\boldsymbol{\eta}_j}\boldsymbol{T}(\boldsymbol{x}_j)) + O_p(N^{-1})\end{aligned}$$

as $N \to \infty$, $j = 1, \ldots, J^*$. Since $E_{\boldsymbol{\eta}}\boldsymbol{T}(\boldsymbol{x}) = -\nabla q(\boldsymbol{\eta})$ and $\mathrm{Var}_{\boldsymbol{\eta}}\boldsymbol{T}(\boldsymbol{x}) = -\nabla^2 q(\boldsymbol{\eta})$, the result follows from the Central Limit Theorem. □

PROOF OF THEOREM 4.1. First we prove part (a). By a Taylor series expansion of (3.3),

$$\begin{aligned}&\log\frac{\rho_J(\hat{F}_1, \ldots, \hat{F}_J)}{\rho_J(F_1, \ldots, F_J)} \\ &\quad = J^{-1}\sum_{j=1}^{J}(\hat{\boldsymbol{\eta}}_j - \boldsymbol{\eta}_j)'\left[\nabla q(\boldsymbol{\eta}_j) - \nabla q\left(J^{-1}\sum_{l=1}^{J}\boldsymbol{\eta}_l\right)\right] + \sum_{j=1}^{J} O_p(\{\hat{\boldsymbol{\eta}}_j - \boldsymbol{\eta}_j\}'\{\hat{\boldsymbol{\eta}}_j - \boldsymbol{\eta}_j\})\end{aligned}$$

as $N \to \infty$. Then, part (a) follows from Lemma A.1. Now we prove part (b). Let $\boldsymbol{\eta} = \boldsymbol{\eta}_1$. By a Taylor series expansion of (3.3),

$$\begin{aligned}(\text{A.1}) \qquad -2\log\rho_J(\hat{F}_1, \ldots, \hat{F}_J) &= J^{-1}\sum_{j=1}^{J}(\hat{\boldsymbol{\eta}}_j - \boldsymbol{\eta})'\{-\nabla^2 q(\boldsymbol{\eta})\}(\hat{\boldsymbol{\eta}}_j - \boldsymbol{\eta}) \\ &\quad - \left(J^{-1}\sum_{j=1}^{J}\hat{\boldsymbol{\eta}}_j - \boldsymbol{\eta}\right)'\{-\nabla^2 q(\boldsymbol{\eta})\}\left(J^{-1}\sum_{j=1}^{J}\hat{\boldsymbol{\eta}}_j - \boldsymbol{\eta}\right) + O_p(N^{-3/2})\end{aligned}$$

as $N \to \infty$. Let $\boldsymbol{I}_K$ be the $K \times K$ identity matrix, and let $\boldsymbol{Z}_1, \ldots, \boldsymbol{Z}_J$ be mutually independent $\mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_K)$. Lemma A.1 and (A.1) imply that

$$\begin{aligned}&-2NJ\log\rho_J(\hat{F}_1, \ldots, \hat{F}_J) \\ &\quad \xrightarrow{d} \sum_{j=1}^{J}\boldsymbol{Z}_j'\{\boldsymbol{A}(\boldsymbol{\eta})\}^{1/2}\{-\nabla^2 q(\boldsymbol{\eta})\}\{\boldsymbol{A}(\boldsymbol{\eta})\}^{1/2}\boldsymbol{Z}_j \\ &\qquad - J^{-1}\left(\sum_{j=1}^{J}\boldsymbol{Z}_j\right)'\{\boldsymbol{A}(\boldsymbol{\eta})\}^{1/2}\{-\nabla^2 q(\boldsymbol{\eta})\}\{\boldsymbol{A}(\boldsymbol{\eta})\}^{1/2}\left(\sum_{j=1}^{J}\boldsymbol{Z}_j\right) \\ &\quad = \sum_{j=1}^{J}\left[\{-\nabla^2 q(\boldsymbol{\eta})\}^{1/2}\{\boldsymbol{A}(\boldsymbol{\eta})\}^{1/2}\left(\boldsymbol{Z}_j - J^{-1}\sum_{l=1}^{J}\boldsymbol{Z}_l\right)\right]' \\ &\qquad \cdot\left[\{-\nabla^2 q(\boldsymbol{\eta})\}^{1/2}\{\boldsymbol{A}(\boldsymbol{\eta})\}^{1/2}\left(\boldsymbol{Z}_j - J^{-1}\sum_{l=1}^{J}\boldsymbol{Z}_l\right)\right] \\ &\quad = \mathrm{trace}\{W\},\end{aligned}$$

as $N \to \infty$ where

$$W = \sum_{j=1}^{J} \left[\{-\nabla^2 q(\eta)\}^{1/2} \{A(\eta)\}^{1/2} \left(Z_j - J^{-1} \sum_{l=1}^{J} Z_l \right) \right] \cdot \left[\{-\nabla^2 q(\eta)\}^{1/2} \{A(\eta)\}^{1/2} \left(Z_j - J^{-1} \sum_{l=1}^{J} Z_l \right) \right]'.$$

Since

$$W \sim Wishart(\{-\nabla^2 q(\eta)\}^{1/2} A(\eta) \{-\nabla^2 q(\eta)\}^{1/2}, J-1),$$

then the characteristic function of the trace of W evaluated at t can be written

$$\varphi(t) = |I_K - 2it\{-\nabla^2 q(\eta)\}^{1/2} A(\eta) \{-\nabla^2 q(\eta)\}^{1/2}|^{-(J-1)/2} = |I_K - 2it\Lambda|^{-(J-1)/2},$$

where Λ is the diagonal matrix of eigenvalues of $\{-\nabla^2 q(\eta)\}^{1/2} A(\eta) \{-\nabla^2 q(\eta)\}^{1/2}$. Therefore, $\varphi(t) = \prod_{k=1}^{K} (1 - 2it\delta_k)^{-(J-1)/2}$, which is the characteristic function of $\sum_{k=1}^{K} \delta_k Y_k$, where $Y_1, \ldots, Y_K$ are mutually independent χ^2_{J-1} random variables, and $\delta_1, \ldots, \delta_K$ are the diagonal elements of Λ. Part (b) then follows. □

PROOF OF THEOREM 5.1. First we prove part (a). By a Taylor series expansion of (5.1),

$$2\nabla Q(\hat{\eta}) = \nabla q \left(\frac{\hat{\eta} + \eta_1}{2} \right) - \nabla q \left(\frac{\hat{\eta} + \eta_2}{2} \right) = [\nabla^2 q(\eta^\dagger)] \left(\frac{\eta_1 - \eta_2}{2} \right),$$

where $\eta^\dagger$ lies on the line connecting $(\hat{\eta} + \eta_1)/2$ to $(\hat{\eta} + \eta_2)/2$. Therefore,

$$\text{(A.2)} \qquad \begin{aligned} 4(\eta_1 - \eta_2)' \nabla Q(\hat{\eta}) &= (\eta_1 - \eta_2)' [\nabla^2 q(\eta^\dagger)] (\eta_1 - \eta_2) \\ &= (\eta_1 - \eta_2)' [-\operatorname{Var}_{\eta^\dagger} T(\boldsymbol{x})] (\eta_1 - \eta_2) < 0, \end{aligned}$$

since $\operatorname{Var}_{\eta^\dagger} T(\boldsymbol{x})$ is positive definite. This proves part (a). Now we prove the first part of part (b). By another Taylor series expansion,

$$\begin{aligned} Q(\hat{\eta}) &= Q(\eta_1) + [\nabla Q(\eta_1)]' (\hat{\eta} - \eta_1) + O_p(N^{-1}) \\ &= \log \rho(F_1, F_2) + \frac{1}{2} \left[\nabla q(\eta_1) - \nabla q \left(\frac{\eta_1 + \eta_2}{2} \right) \right]' (\hat{\eta} - \eta_1) + O_p(N^{-1}) \end{aligned}$$

as $N \to \infty$. Using Lemma A.1 the result follows. The second part of part (b) follows similarly. □

PROOF OF THEOREM 5.2. Fix a value η^*. Taking gradients with respect to $\hat{\eta}$, we note that

$$2\nabla \log \frac{\rho_2(\hat{F}, F^*)}{\rho_2(F, F^*)} = \nabla q(\hat{\eta}) - \nabla q \left(\frac{\hat{\eta} + \eta^*}{2} \right) = [\nabla^2 q(\eta^\dagger)] \left(\frac{\hat{\eta} - \eta^*}{2} \right),$$

where $\eta^\dagger$ lies on the line connecting $\hat{\eta}$ to $(\hat{\eta} + \eta^*)/2$, by a Taylor series expansion. Therefore, if $\hat{\eta} \neq \eta^*$, then

$$\begin{aligned} 4(\hat{\eta} - \eta^*)' \nabla \log \frac{\rho_2(\hat{F}, F^*)}{\rho_2(F, F^*)} &= (\hat{\eta} - \eta^*)' [\nabla^2 q(\eta^\dagger)] (\hat{\eta} - \eta^*) \\ &= (\hat{\eta} - \eta^*)' [-\operatorname{Var}_{\eta^\dagger} T(\boldsymbol{x})] (\hat{\eta} - \eta^*) < 0, \end{aligned}$$

since $\mathrm{Var}_{\eta^\dagger} \boldsymbol{T}(\boldsymbol{x})$ is positive definite. □

PROOF OF THEOREM 5.3. By a Taylor series expansion, under the hypothesis that $F_3 = F_1$,

$$2[U(\hat{F}_1, \hat{F}_2, \hat{F}_3) - \log \rho_2(F_1, F_2)] = (\hat{\eta}_3 - \eta_1)' \left[\nabla q(\eta_1) - \nabla q\left(\frac{\eta_1 + \eta_2}{2}\right)\right]$$
$$+ (\hat{\eta}_2 - \eta_2)' \left[\nabla q(\eta_2) - \nabla q\left(\frac{\eta_1 + \eta_2}{2}\right)\right] + \sum_{j=1}^{3} O_p(\{\hat{\eta}_j - \eta_j\}'\{\hat{\eta}_j - \eta_j\})$$

as $N \to \infty$. Using Lemma A.1 and noting the independence between $\hat{\eta}_2$ and $\hat{\eta}_3$, the first part of the theorem is proved. The second part of the theorem is proved similarly. □

REFERENCES

Anderson, T. W. (1984). *An Introduction to Multivariate Statistical Analysis*, 2nd ed., Wiley, New York.

Bar-Hen, A. and J. J. Daudin (1998). Asymptotic distribution of Matusita's distance: Application to the location model, *Biometrika*, **85**, 447–481.

Beran, Rudolf (1977). Minimum Hellinger distance estimates for parametric models, *Ann. Statist.*, **5**, 445–463.

Bickel, Peter J. and Doksum, Kjell A. (1977). *Mathematical Statistics: Basic Ideas and Selected Topics*, Holden-Day, Oakland, California.

Kotz, Samuel, Johnson, N. L. and Boyd, D. W. (1967). Series representations of distributions of quadratic forms in normal variables. I. Central case, *Ann. Math. Statist.*, **38**, 823–837.

Krzanowski, W. J. (1975). Discrimination and classification using both binary and continuous variables, *J. Amer. Statist. Assoc.*, **70**, 782–790.

Krzanowski, W. J. (1983). Distance between populations using mixed continuous and categorical variables, *Biometrika*, **70**, 235–243.

Krzanowski, W. J. (1984). On the null distribution of distance between two groups, using mixed continuous and categorical variables, *J. Classification*, **1**, 243–253.

Krzanowski, W. J. (1987). A comparison between two distance-based discriminant principles, *J. Classification*, **4**, 73–84.

LeCam, L. (1970). On the assumptions used to prove asymptotic normality of maximum likelihood estimates, *Ann. Math. Statist.*, **41**, 802–828.

Matusita, K. (1955). Decision rules based on the distance, for problems of fit, two samples, and estimation, *Ann. Math. Statist.*, **26**, 631–640.

Matusita, K. (1966). A distance and related statistics in multivariate analysis, *Multivariate Analysis* (ed. P. R. Krishnaiah), 187–200, Academic Press, New York.

Matusita, K. (1967*a*). Classification based on distance in multivariate Gaussian cases, *Proc. Fifth Berkeley Symp. on Math. Statist. Prob.*, Vol. 1, 299–304, Univ. of California Press, Berkeley.

Matusita, K. (1967*b*). On the notion of affinity of several distributions and some of its applications, *Ann. Inst. Statist. Math.*, **19**, 181–192.

Matusita, K. (1971). Some properties of affinity and applications, *Ann. Inst. Statist. Math.*, **23**, 137–155.

Matusita, K. (1973). Discrimination and the affinity of distributions, *Discriminant Analysis and Applications*, (ed. T. Cacoullos), 213–223, Academic Press, New York.

McLachlan, Geoffrey J. (1992). *Discriminant Analysis and Statistical Pattern Recognition*, Wiley, New York.

Olkin, I. and R. F. Tate (1961). Multivariate correlation models with mixed discrete and continuous variables, *Ann. Math. Statist.*, **32**, 448–465. (Correction: *ibid.* (1965). **36**, 343–344).