stichting
mathematisch
centrum

AFDELING MATHEMATISCHE STATISTIEK
(DEPARTMENT OF MATHEMATICAL STATISTICS)

SW 96/83 AUGUSTUS

A.W. AMBERGEN & W. SCHAAFSMA

INTERVAL ESTIMATES FOR POSTERIOR PROBABILITIES IN A MULTIVARIATE NORMAL CLASSIFICATION MODEL

kruislaan 413 1098 SJ amsterdam

Printed at the Mathematical Centre, Kruislaan 413, Amsterdam, The Netherlands.

The Mathematical Centre, founded 11 February 1946, is a non-profit institution for the promotion of pure and applied mathematics and computer science. It is sponsored by the Netherlands Government through the Netherlands Organization for the Advancement of Pure Research (Z.W.O.).

1980 Mathematics subject classification: 62H30

Interval estimates for posterior probabilities in a multivariate normal classification model

by

A.W. Ambergen & W. Schaafsma*)

ABSTRACT

This paper is devoted to the asymptotic distribution of estimators for the posterior probability that a p dimensional observation vector originates from one of k normal distributions with identical covariance matrices. The estimators are based on training samples from the k distributions involved. Observation vector and prior probabilities are regarded as given constants. The validity of various estimators and approximate confidence intervals is investigated by simulation experiments.

KEY WORDS & PRHASES: *Estimating posterior probabilities, classification, discriminant analysis, multivariate normal distributions*

*) Department of Mathematics, University of Groningen

1. INTRODUCTION

Suppose that an observation x comes from one of k populations Π_h, $h = 1,\dots,k$, which are characterized by p-dimensional multivariate normal distributions with equal covariance matrices. Accordingly let f_h denote the p.d.f. of $N_p(\mu_h,\Sigma)$, $h = 1,\dots,k$. The parameters $\mu_1,\dots,\mu_k$, Σ are unknown. We assume that past experience is available in the form of outcomes of independent random vectors $X_{h1},\dots,X_{hn_h}$, $h = 1,\dots,k$, X_{hi} having density f_h. Let ρ_h denote the prior probability that the observation comes from Π_h, $h = 1,\dots,k$ $\Sigma_{h=1}^k \rho_h = 1$. For $\rho_1,\dots,\rho_k$ and x given the posterior probabilities

$$(1.1) \qquad \rho_{t|x} = \rho_t f_t(x) \Big/ \sum_{h=1}^{k} \rho_h f_h(x), \quad t = 1,\dots,k$$

are considered as unknown parameters which are to be estimated from the training samples.

Let $R_{\cdot|x} = (R_{1|x},\dots,R_{k|x})^T$ denote any of the estimators for $\rho_{\cdot|x} = (\rho_{1|x},\dots,\rho_{k|x})^T$ to be defined in section 2. We shall prove that $n^{\frac{1}{2}}(R_{\cdot|x}-\rho_{\cdot|x})$ is asymptotically normal with expectation zero and a singular dispersion matrix. Application to practice requires that the unknown parameters in the asymptotic covariance matrix are replaced by suitable estimates. The diagonal elements of the thus obtained estimated asymptotic covariance matrix provide the means of constructing asymptotic confidence intervals for the posterior probabilities separarely. The whole matrix is needed if one wants to apply a Scheffé-type method for judging linear combinations. Pairwise comparisons might be treated by applying theory for the case k = 2 where certain exact moments can be exploited (see e.g. SCHAAFSMA-VAN VARK (1979)). The main purpose of this paper is to present the asymptotic variances and covariances of the $R_{t|x}$'s as means of expressing the involved uncertainties.

Most of the literature about estimating posterior probabilities deals with the case k = 2. The case p = 1, k = 2 is considered in SCHAAFSMA-VAN VARK (1977). The case $p \geq 1$, k = 2 can be found in SCHAAFSMA-VAN VARK (1979). In AMBERGEN-SCHAAFSMA (1983) the extension to $p \geq 1$, $k \geq 2$, with no assumption about the equality of covariance matrices, is considered. The corresponding theory is easier than that presented in this paper ($p \geq 1$,

$k \geq 2$, $\Sigma_1 = \ldots = \Sigma_k$) because here we have dependence of the density estimators. Apart from the "estimative" methods used in this paper the "predictive" method of GEISSER (1964) has been discussed in the literature. AITCHISON, HABBEMA and KAY (1977) is a comparison of the two methods. McLACHLAN (1977) studies the bias of sample based posterior probabilities. McLACHLAN (1979) compares the bias of classical plug-in estimators with that of predictive estimators. A recent reference is RIGBY (1982) who constructs credibility intervals for the posterior probabilities in order to compare the estimative and predictive estimators.

2. DEFINITION OF THE ESTIMATORS

The densities of the populations are given by

$$(2.1) \qquad f_h(x) = |2\pi\, \Sigma|^{-\frac{1}{2}} \exp(-\tfrac{1}{2}\Delta^2_{x;h}), \qquad h = 1,\ldots,k$$

where

$$(2.2) \qquad \Delta^2_{x;h} = (x-\mu_h)^T \Sigma^{-1} (x-\mu_h).$$

Hence

$$(2.3) \qquad \rho_{t|x} = \rho_t \exp(-\tfrac{1}{2}\Delta^2_{x;t}) / \sum_{h=1}^{k} \rho_h \exp(-\tfrac{1}{2}\Delta^2_{x;h})$$

For $k = 2$ is it useful to rewrite (2.3) as

$$(2.4) \qquad \rho_{t|x} = [1 + \rho_{3-t}\rho_t^{-1} \exp\{\tfrac{1}{2}(\Delta^2_{x;t} - \Delta^2_{x;3-t})\}]^{-1}, \qquad t = 1,2$$

because then an approximate confidence interval for $\rho_{t|x}$ can be obtained by transforming the approximate confidence interval for $\Delta^2_{x;t} - \Delta^2_{x;3-t}$ based on exact moments (see SCHAAFSMA-VAN VARK (1979) and RIGBY (1982), who used a similar approach in a Bayesian context).

Let X_h denote the mean of the h-th sample and S the pooled matrix of cross-products:

$$(2.5) \qquad X_h = h_h^{-1} \sum_{i=1}^{n_h} X_{hi}, \quad S = \sum_{h=1}^{k} \sum_{i=1}^{n_h} (X_{hi} - X_h)(X_{hi} - X_h)^T.$$

It sometimes happens that extra samples are available for estimating S. Therefore, instead of $S \sim W_p\ (n-k,\Sigma)$ where $n = \Sigma_i n_i$, we shall work with $S_f \sim W_p\ (f,\Sigma)$.

The maximum likelihood estimator $R^{(0)}_{t|x}$ for $\rho_{t|x}$ is obtained by plugging in the estimators X_h for μ_h, $(f+k)^{-1}S$ for Σ and $\hat{\Delta}^{2(0)}_{x;h} = (f+k)V^2_{x;h}$ for $\Delta^2_{x;h}$ where

$$(2.6) \qquad V^2_{x;h} = (x-X_h)^T S_f^{-1}(x-X_h).$$

Other estimators for $\rho_{t|x}$ are obtained by plugging in unbiased estimators for various parameters in (2.2) or (2.3). This gives us the following estimators for $\Delta^2_{x;h}$:

$$(2.7) \qquad \begin{aligned} \hat{\Delta}^{2(1)}_{x;h} &= fV^2_{x;h} \\ \hat{\Delta}^{2(2)}_{x;h} &= (f-p-1)V^2_{x;h} \\ \hat{\Delta}^{2(3)}_{x;h} &= (f-p-1)V^2_{x;h} - pn_h^{-1} \end{aligned}$$

based on $ES_f = f\Sigma$, $ES_f^{-1} = (f-p-1)^{-1}\Sigma^{-1}$ and $EV^2_{x;h} = (f-p-1)^{-1}\Delta^2_{x;h} + n_h^{-1}(f-p-1)^{-1}p$ respectively. By plugging into (2.3) we obtain the estimators

$$(2.8) \qquad R^{(j)}_{t|x} = \rho_t \exp(-\tfrac{1}{2}\hat{\Delta}^{2(j)}_{x;t}) / \sum_{h=1}^{k} \rho_h \exp(-\tfrac{1}{2}\hat{\Delta}^{2(j)}_{x;h})$$

$(t=1,\ldots,k;\ j=0,\ldots,3)$.

3. THE ASYMPTOTIC DISTRIBUTION OF THE ESTIMATORS

All estimators for $\Delta^2_{x;h}$, suggested in section 2, are asymptotically equivalent. In this section it is therefore sufficient to focus on $fV^2_{x;h}$. The corresponding estimator $R_{t|x} = R^{(1)}_{t|x}$ for $\rho_{t|x}$ has the same asymptotic distribution as each of the other estimators $R^{(j)}_{t|x}$ (j=0,2,3). $(R_{1|x},\ldots,R_{k|x})^T$ is asymptotically efficient and the asymptotic covariance matrix follows from Fisher's information matrix. Elaborating on this and related principles we have to consider the inverse Wishart distribution.

LEMMA 3.1. *If* $W_f \sim W_p(f,\Sigma)$ *then*

$$\mathcal{L} f^{\frac{1}{2}}(f^{-1}\mathrm{vec}(W_f) - \mathrm{vec}(\Sigma)) \to N_{p^2}(0,A) \tag{3.1}$$

and

$$\mathcal{L} f^{\frac{1}{2}}(f\,\mathrm{vec}(W_f^{-1}) - \mathrm{vec}(\Sigma^{-1})) \to N_{p^2}(0,B) \tag{3.2}$$

with

$$A_{ijk\ell} = \sigma_{ik}\sigma_{j\ell} + \sigma_{i\ell}\sigma_{jk}$$

$$B_{ijk\ell} = \sigma^{ik}\sigma^{j\ell} + \sigma^{jk}\sigma^{i\ell}$$

where we use the notations $M_{ijk\ell} = M_{(j-1)p+i,(\ell-1)p+k}$ *for* M *a* $p^2 \times p^2$ *matrix*, $\sigma_{ij} = \Sigma_{ij}$ *and* $\Sigma^{ij} = (\Sigma^{-1})_{ij}$.

PROOF. (3.1) is an immediate consequence of the multivariate central limit theorem. (3.2) follows from the δ-method:

$$B_{ijk\ell} = \sum_{rstu} \left(\frac{\partial\sigma^{ij}}{\partial\sigma_{rs}}\right) A_{rstu} \left(\frac{\partial\sigma^{k\ell}}{\partial\sigma_{tu}}\right).$$

The proof is completed by using

$$\frac{\partial\sigma^{ij}}{\partial\sigma_{\alpha\beta}} = -\sigma^{i\alpha}\sigma^{\beta j}. \quad \square$$

LEMMA 3.2. *If* $f \to \infty$, $n_h/f \to b_h > 0$ $(h=1,\ldots,k)$, $V_x^2 = (V_{x;1}^2,\ldots,V_{x;k}^2)$ *and* $\Delta_x^2 = (\Delta_{x;1}^2,\ldots,\Delta_{x;k}^2)^T$ *then*

$$\mathcal{L} f^{\frac{1}{2}}(fV_x^2 - \Delta_x^2) \to N_k(0,\Gamma) \tag{3.3}$$

where Γ *is determined by*

$$\begin{aligned} \Gamma_{h,h} &= 4\Delta_{x;h}^2/b_h + 2\Delta_{x;h}^4 \\ \Gamma_{h,t} &= 2\{(x-\mu_h)^T\Sigma^{-1}(x-\mu_t)\}^2 \quad h \neq t. \end{aligned} \tag{3.4}$$

PROOF. The independent random variables $X_1,\ldots,X_k,S_f$ satisfy $f^{\frac{1}{2}}(X_h-\mu_h) \xrightarrow{L} N_p(0,b_h^{-1}\Sigma)$ and $f^{\frac{1}{2}}(fS_f^{-1}-\Sigma^{-1}) \xrightarrow{L} N_{p^2}(0,B)$ where B is defined in lemma 3.1 and S_f in section 2. Consider the partial derivatives

$$\frac{\partial\Delta^2_{x;i}}{\partial\mu_j} = -2\Sigma^{-1}(x-\mu_i)\delta_{ij} \quad i,j = 1,\ldots,k$$

and

$$\frac{\partial\Delta^2_{x;i}}{\partial\sigma^{\alpha\beta}} = (x-\mu_i)_\alpha(x-\mu_j)_\beta \quad i = 1,\ldots,k; \quad \alpha,\beta = 1,\ldots,p$$

where $\delta_{ij} = 1$ if $i = j$ and $= 0$ if $i \neq j$.
The δ-method gives

$$\Gamma_{ii} = \left(\frac{\partial\Delta^2_{x;j}}{\partial\mu_j}\right)^T b_i^{-1}\,\Sigma\left(\frac{\partial\Delta^2_{x;i}}{\partial\mu_i}\right) + \sum_{rstu}\left(\frac{\partial\Delta^2_{x;i}}{\partial\sigma^{rs}}\right)B_{rstu}\left(\frac{\partial\Delta^2_{x;i}}{\partial\sigma^{tu}}\right)$$

and

$$\Gamma_{ij} = \sum_{rstu}\left(\frac{\partial\Delta^2_{x;i}}{\partial\sigma^{rs}}\right)B_{rstu}\left(\frac{\partial\Delta^2_{x;j}}{\partial\sigma^{tu}}\right) \quad i \neq j$$

(3.4) follows by simple computation.

THEOREM 3.3. *If* $f \to \infty$, $n_h/f \to b_h > 0$ $(h=1,\ldots,k)$ $R_{.|x} = (R_{1|x},\ldots,R_{k|x})^T$ *and* $\rho_{.|x} = (\rho_{1|x},\ldots,\rho_{k|x})^T$ *then*

(3.7) $$Lf^{\frac{1}{2}}(R_{.|x}-\rho_{.|x}) \to N_k(0,\Psi\Gamma\Psi)$$

where Γ *is determined by (3.5) and* Ψ *by*

(3.8) $$\Psi_{t,t} = \tfrac{1}{2}\rho_{t|x}(-1+\rho_{t|x})$$
$$\Psi_{t,h} = \tfrac{1}{2}\rho_{t|x}\rho_{h|x} \quad t \neq h.$$

PROOF. With the δ-method where Ψ is the matrix of partial derivatives.

4. FOUR METHODS TO CONSTRUCT CONFIDENCE INTERVALS

An approximate $100(1-\alpha)\%$ confidence interval for $\rho_{t|x}$ is given by

(4.1) $$[R_{t|x}^{(j)} - \tfrac{1}{2}L_{t|x}^{(j)},\ R_{t|x}^{(j)} + \tfrac{1}{2}L_{t|x}^{(j)}],\quad j = 0,\ldots,3$$

where $R_{t|x}^{(j)}$ has been defined in (2.8) and

(4.2) $$L_{t|x}^{(j)} = 2u_{\frac{1}{2}\alpha} f^{-\frac{1}{2}} \{\hat{\Psi}^{(j)}\hat{\Gamma}^{(j)}\hat{\Psi}^{(j)}\}_{t,t}^{\frac{1}{2}}$$

with $u_{\frac{1}{2}\alpha}$ defined by $P(U > u_{\frac{1}{2}\alpha}) = \frac{1}{2}\alpha$ if U has a standard normal distribution. The estimators $\hat{\Gamma}^{(j)}$ and $\hat{\Psi}^{(j)}$ for the corresponding parameters in (3.4) and (3.8) are obtained by plugging in the estimators $R_{t|x}^{(j)}$ for $\rho_{t|x}$, $\hat{\Delta}^{2(j)}_{x;h}$ for $\Delta^2_{x;h}$ and the parameter $(x-\mu_h)^T \Sigma^{-1} (x-\mu_t)$ in (3.4) is estimated by $b^{(j)}(x-X_h)S^{-1}(x-X_t) + c^{(j)}$, $j = 0,\ldots,3$, where $b^{(0)} = f+k$, $b^{(1)} = f$, $b^{(2)} = f-p-1$, $b^{(3)} = f-p-1$, $c^{(0)} = c^{(1)} = c^{(2)} = 0$ and $c^{(3)} = -pn_h^{-1}\delta_{h,t}$ with $\delta_{h,t} = 1(0)$ if $h = t$ $(h \neq t)$.

5. SIMULATION EXPERIMENT

An overall comparison of small sample performance of the estimators $R_{t|x}^{(j)}$ and the approximate confidence intervals $R_{t|x}^{(j)} \pm \frac{1}{2}L_{t|x}^{(j)}$ $(j=0,\ldots,3)$ is rather complicated because the performance depends on the very large number of parameters

(5.1) $$p,k,n_1,\ldots,n_k,t,\alpha,x,\rho_1,\ldots,\rho_k,\mu_1,\ldots,\mu_k,\Sigma$$

where t indicates the number of the density from which the score vector has been drawn. We selected 500 parameter points for the simulation experiments to be performed. The results for the chosen parameter points are rather accurate because we did the following for each point: compute $\rho_{t|x}$, generate 1000 times a set of training samples and compute each time $R_{t|x}^{(j)}$, $L_{t|x}^{(j)}$ $(j=0,\ldots,3)$. Count the number of times the interval contains the true value $\rho_{t|x}$. This number, divided by 10, should be compared with the value $100(1-\alpha)$. The 500 points were grouped into 25 clusters of each 20 points. Within a cluster only the x vectors differ because they were drawn independently. For the points within a cluster the same training set was used. We made the restrictions $t = 1$, $\alpha = 0.05$, $\mu_1 = 0_p$, $\Sigma = I_p$, $\rho_h = k^{-1}$ $(h=1,\ldots,k)$ and considered only $\rho_{1|x}$

which is the most important, because largest, posterior probability. For each cluster we averaged the results of the 20 points. These averaged results with their standard deviations are presented in table 1; a cluster corresponds with a row in the table. In order to get a nice layout of the table we introduce the following notations:

$$n = (n_1,\dots,n_k),\quad \mu = (\mu_1;\dots;\mu_k)$$

$$a = (0,0,0,0)^T;\quad b = (2,0,0,0)^T;\quad c = (0,2,0,0)^T;\quad d = (1,1,1,1)^T$$

$$e = (1,1,0,0)^T;\quad f = (0,0,2,0)^T;\quad g = (0,0,0,2)^T;\quad h = (0,0,1,1)^T$$

$$1_4 = (1,1,1,1);\quad 1_8 = (1_4;1_4);\quad m_4 = (0,1,0,1);\quad m_8 = (m_4;m_4)$$

Bias, m.s.e. and m.a.d. of the point estimators $R^{(j)}_{t|x}$ (j=0,...,3; t=1) were also studied.

CONCLUSIONS

For the chosen parameter points we conclude that the m.ℓ. estimator $R^{(0)}_{t|x}$ has smaller bias, smaller m.a.d. and smaller m.s.e. than its competitors, at least, on the average. Table 1 shows that the confidence intervals for j = 1,2 and 3 are slightly more reliable than those based on the m.ℓ. estimator (j=0). Sample sizes should certainly not be smaller than 50 (25) if one requires that the true confidence coefficient of the interval based on the m.ℓ. estimator and $1-\alpha = .95$, should not be smaller than .90 (.85).

ACKNOWLEDGEMENTS

The motivation for this paper arose from discussions with the anthropologist G.N. van Vark. Several computations were verified by A.G.M. Steerneman.

Input parameter values for the clusters.		Averaged confidence coefficients with standard deviations for the four procedures.			
p,k,μ	n	j = 0	j = 1	j = 2	j = 3
p = 4	50.1_4	92.0 2.0	92.8 1.8	93.3 1.6	93.0 1.6
k = 4	50.1_4-25m_4	90.8 2.6	91.8 2.3	92.8 1.9	92.3 2.0
μ = (abcd)	25.1_4+25m_4	90.1 2.9	91.2 3.0	92.1 2.7	91.6 2.4
	25.1_4	88.9 3.2	90.6 2.7	92.0 2.1	91.4 1.9
	15.1_4	84.3 4.5	87.2 4.1	89.7 3.4	88.4 3.2
p = 4	50.1_8	92.4 2.2	93.1 1.7	93.3 1.4	92.9 1.2
k = 8	50.1_8-25m_8	92.3 2.2	93.1 1.4	93.4 1.1	92.4 1.2
μ =	25.1_8+25m_8	90.0 3.0	91.8 2.3	92.2 1.8	91.7 1.5
(ab...gh)	25.1_8	90.5 3.3	91.6 2.3	92.2 1.6	91.3 1.3
	15.1_8	87.7 4.8	89.8 3.3	90.0 2.2	89.4 1.6
p = 8	50.1_4	88.7 1.6	89.3 1.5	90.9 1.5	90.4 1.6
k = 4	50.1_4-25m_4	87.0 2.4	87.9 2.2	90.0 2.1	89.4 1.9
μ = $\begin{pmatrix}abcd\\aaaa\end{pmatrix}$	25.1_4+25m_4	86.3 2.4	87.5 2.2	89.6 2.0	89.1 2.0
	25.1_4	83.4 2.4	85.0 2.2	88.2 2.3	87.2 2.3
	15.1_4	76.2 4.0	78.7 3.8	84.8 3.3	83.1 3.3
p = 8	50.1_4	86.4 2.3	88.0 2.1	91.4 2.7	91.1 2.7
k = 4	50.1_4-25m_4	83.8 2.5	86.0 2.4	90.4 2.8	90.5 3.1
μ = $\begin{pmatrix}abcd\\abcd\end{pmatrix}$	25.1_4+25m_4	84.4 2.6	86.2 2.7	90.3 3.5	89.2 3.2
	25.1_4	80.2 2.7	83.3 2.8	89.0 4.4	88.6 4.6
	15.1_4	73.3 3.8	77.5 4.3	86.5 6.3	85.6 6.7
p = 8	50.1_8	89.4 2.2	90.0 1.8	90.8 1.4	90.3 1.4
k = 8	50.1_8-25m_8	89.1 2.1	90.0 2.1	90.9 1.6	89.4 1.7
μ =	25.1_8+25m_8	85.9 3.1	86.7 2.6	87.9 1.9	88.2 1.7
$\begin{pmatrix}ab...gh\\aa...aa\end{pmatrix}$	25.1_8	85.0 4.2	86.6 3.2	88.0 2.3	87.3 2.2
	15.1_8	79.2 5.2	82.0 3.6	84.8 2.8	83.5 2.6

Table 1. The reliability of the confidence intervals.

[1] AITCHISON, J., J.D.F. HABBEMA & J.W. KAY (1977), *A critical comparison of two methods of statistical discrimination*. Appl. Statist., 26, no. 1, pp. 15-25.

[2] AMBERGEN, A.W. & W. SCHAAFSMA (1983), *Interval estimates for posterior probabilities, applications to Border Cave*. In Multivariate Statistical Methods, (G.N. van Vark and W.W. Howells, Ed.). Reidel, Dordrecht, The Netherlands.

[3] GEISSER, S. (1964), *Posterior odds for Multivariate Normal Classification*. J. Roy. Statist. Soc. Ser. B 26, pp. 69-76.

[4] McLACHLAN, G.J. (1977), *The bias of sample based posterior probabilities*. Biom. J. vol. 19, no. 6, pp. 421-426.

[5] McLACHLAN, G.J. (1979), *A comparison of the estimative and predictive methods of estimating posterior probabilities*. Comm. Statist. A-Theory Methods, 8(9), pp. 919-929.

[6] RIGBY, R.A. (1982), *A credibility interval for the probability that a new observation belongs to one of two multivariate normal populations*. J.R. Statist. Soc. B, 44, no. 2, pp. 212-220.

[7] SCHAAFSMA, W. & G.N. VAN VARK (1977), *Classification and discrimination problems with applications*, part I, Statist. Neerlandica, 31, pp. 25-45.

[8] SCHAAFSMA, W. & G.N. VAN VARK (1979), *Classification and discrimination problems with applications*, part II^a. Statist. Neerlandica, 33, pp. 91-126.

[9] VAN VARK, G.N. & W.W. HOWELLS (1983), *Multivariate Statistical Methods in Physical Anthropology*. Reidel, Dordrecht, The Netherlands.