Journal of Modern Applied Statistical Methods

Volume 16 | Issue 2 Article 19

December 2017

Approximating the Distribution of Indefinite Quadratic Forms in Normal Variables by Maximum Entropy Density Estimation

Ghasem Rekabdar
Abadan Branch, Islamic Azad University, Abadan, Iran, ghasem_rekabdar@yahoo.com

Rahim Chinipardaz
Shahid Chamran University of Ahvaz, Ahvaz, Iran, chinipardaz_r@scu.ac.ir

Recommended Citation
Rekabdar, G., & Chinipardaz, R. (2017). Approximating the Distribution of Indefinite Quadratic Forms in Normal Variables by Maximum Entropy Density Estimation. Journal of Modern Applied Statistical Methods, 16(2), 359-377. doi: 10.22237/jmasm/1509495540

Journal of Modern Applied Statistical Methods
November 2017, Vol. 16, No. 2, 359-377.
doi: 10.22237/jmasm/1509495540

ISSN 1538 − 9472

Approximating the Distribution of Indefinite Quadratic Forms in Normal Variables by Maximum Entropy Density Estimation

Ghasem Rekabdar
Abadan Branch, Islamic Azad University
Abadan, Iran

Rahim Chinipardaz
Shahid Chamran University of Ahvaz
Ahvaz, Iran

The quadratic form of non-central normal variables is presented based on a sum of weighted independent non-central chi-square variables. This presentation provides moments of quadratic form. The maximum entropy method is used to estimate the density function because distribution moments of quadratic forms are known. A Euclidean distance is proposed to select an appropriate maximum entropy density function. In order to compare with other methods some numerical examples were evaluated. Also, for discrimination between two groups by the Euclidean distances, we obtained a stochastic representation for the linear discriminant function using the quadratic form. The maximum entropy estimation was an acceptable method to approximate the distribution of quadratic forms in normal variables.

Keywords: Quadratic forms, maximum entropy density estimation, non-central chi-square distribution, linear discriminant analysis

Introduction

The distribution of quadratic forms in normal vectors or sums of weighted independent non-central chi-square variables are considered in some of applied statistical problems (Mathai & Provost, 1992). Researchers have introduced different methods to approximate the distribution of a weighted sum of chi-square variables. Distribution approximation based on moments is a simple method with suitable accuracy. This method is also used frequently for approximate distribution of the quadratic form. For the distribution of non-negative quadratic forms in non-central normal variables, Patnaik's two moments (Patnaik, 1949) and Pearson's

Ghasem Rekabdar is an Assistant Professor in the Department of Mathematics. Email them at: ghasem_rekabdar@yahoo.com. Rahim Chinipardaz is Faculty of Mathematical and Computer Science in the Department of Statistics. Email them at: chinipardaz_r@scu.ac.ir.

three moments (Pearson, 1959) central chi-square approximation are the simplest methods that are used in abundance. Recently, Liu, Tang and Zhang (2009) proposed a non-central chi-square approximation with the unknown degrees of freedom and non-centrality parameter determined by the first four cumulants of the quadratic forms. These methods are accurate for the approximate upper tail of definite (non-negative) quadratic forms and, in the case of indefinite quadratic forms, are not suitable. In the case of indefinite forms, the indefinite quadratic form can be written as the difference of two independent definite quadratic forms. In this case, the density function of a positive definite quadratic form can be approximated according to polynomial gamma or generalized gamma density functions (Mohsenipour & Provost, 2011). In light of this method the density function of the indefinite quadratic form can be approximated by the distribution of the difference of two polynomial gammas. In gamma-polynomial density approximation, we can use more than four moments to approximate the distribution of quadratic forms.

The maximum entropy density estimation is a flexible method to assign values of probability distributions based on limited information such as moments. In general, more limited information such as the percentiles of distribution are recommended. The maximum entropy approach, proposed by Jaynes (1957), is a flexible and powerful tool for density estimation. It was proposed for solving problems with little information about distribution. The maximum entropy method is a suitable tool to estimate properly all special distributions such as normal, exponential, Cauchy, etc. Expressed in Jaynes' language, all known special distributions represent an unbiased probability distribution when some information is not available (Zong, 2006). In this study, the maximum entropy method will be used to approximate indefinite quadratic forms distributions because the moments of weighted sums of non-central chi-square variable are known. As a result, this information can be used to maximize Shanon's entropy. Also, a new approach based on the distance between distributions is proposed to select the number of constraints and compared with the conventional method.

Stochastic Representation of Quadratic Forms

Recall some definitions and basic properties of indefinite quadratic forms in non-central normal variables: Let $\mathbf{X} = (X_1, \ldots, X_d)'$ be a multivariate normal random vector $\mathbf{X} \sim N_d(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ with mean vector $\boldsymbol{\mu}$ and positive definite covariance matrix $\boldsymbol{\Sigma}$. The quadratic form in the random variable $\mathbf{X}$ associated with a $d \times d$ real symmetric matrix $\mathbf{A}$ is defined by

$$Q = \mathbf{X}'\mathbf{A}\mathbf{X} \tag{1}$$

Because (1) is a nonlinear combination of correlated univariate normal random variables, it is hard to obtain any approximation directly from it. Instead, we can derive a stochastic representation for quadratic forms in terms of some simpler distribution, i.e.

$$Q = \sum_{i=1}^{n} \lambda_i \chi_1^2\left(\gamma_i^2\right) \tag{2}$$

where the $\chi_1^2\left(\gamma_i^2\right)$ are independent non-central chi-square random variables with one degree of freedom and non-centrality parameters γ_i^2. The weights $\lambda_1 \geq \ldots \geq \lambda_d$ are obtained by the spectral decomposition theorem, i.e.

$$\boldsymbol{\Sigma}^{\frac{1}{2}}\mathbf{A}\boldsymbol{\Sigma}^{\frac{1}{2}} = \mathbf{H}\boldsymbol{\Lambda}\mathbf{H}'$$

where $\boldsymbol{\Lambda}$ is a diagonal matrix for which the diagonal elements $\lambda_1,\ldots,\lambda_d$ are the eigenvalues of matrix $\boldsymbol{\Sigma}^{\frac{1}{2}}\mathbf{A}\boldsymbol{\Sigma}^{\frac{1}{2}}$, $\boldsymbol{\Sigma}^{\frac{1}{2}}$ denotes the symmetric square root of matrix $\boldsymbol{\Sigma}$, and $\mathbf{H}$ is an orthogonal matrix. The non-centrality parameters γ_i^2 are obtained by taking the square elements of the vector

$$\boldsymbol{\gamma} = \left(\gamma_1,\ldots,\gamma_d\right)' = \mathbf{H}'\boldsymbol{\Sigma}^{-\frac{1}{2}}\boldsymbol{\mu}$$

According to this stochastic representation, obtain the cumulant generating function of Q by

$$\mathrm{C}(t) = \frac{1}{2}\sum_{i=1}^{d}\log\left(1-2\lambda_i t\right) + \sum_{i=1}^{d}\frac{\lambda_i\gamma_i^2 t}{\left(1-2\lambda_i t\right)}$$

The formula for the r^{th} cumulant of the quadratic form is

$$\kappa_r = 2^{r-1}(r-1)!\sum_{i=1}^{d}\lambda_i^r\left(1+r\gamma_i^2\right) \tag{3}$$

The moments of the quadratic form can be obtained from its cumulants by means of the recursive relationship obtained by Smith (1995). According to this formula, the r^{th} moment of the quadratic form is given by

$$\mu'_r = \sum_{j=0}^{r-1} c_j^{r-1} \kappa_r \mu'_j \tag{4}$$

where $\mu'_j = \mathrm{E}\left(X^j\right)$ and $c_j^{r-1} = (r-1)!/\,j!(r-j-1)!$ are the j^{th} non-central moment and the combination j of $r-1$, respectively.

An Application

The linear discriminant function is used when the d-dimensional observation $\mathbf{y} = (y_1, \ldots, y_d)'$ in two independent populations Π_1 and Π_2 has multivariate normal densities $N_d(\boldsymbol{\mu}_1, \boldsymbol{\Sigma})$ and $N_d(\boldsymbol{\mu}_2, \boldsymbol{\Sigma})$. We use the notation $P(i \mid j)$ to denote the probability of misclassification of an observation $\mathbf{y}$ into group i when, in fact, it belongs to the group j, where $i, j = 1, 2$. For simplicity, suppose that the prior probabilities are taken to be equal, i.e. $p_1 = p_2 = 1/2$. By the Bayes optimal classification rule, the linear discriminant function is defined as

$$W = (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)' \boldsymbol{\Sigma}^{-1} \left(\mathbf{y} - \frac{1}{2}(\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2) \right) \tag{5}$$

Future observations $\mathbf{y}$ are assigned into the group Π_1 when $W \geq 0$. In the case of $W < 0$, this observation is assigned into the group Π_2. The total probability of misclassification (TMP) is given by

$$\begin{aligned} \mathrm{TPM} &= \frac{1}{2}\left(\mathrm{P}(2 \mid 1) + \mathrm{P}(1 \mid 2)\right) \\ &= \frac{1}{2}\left(\mathrm{P}(W < 0) + \mathrm{P}(W \geq 0)\right) \end{aligned}$$

Because the W distribution is univariate normal, the TPM is obtained as (Johnson & Wichern, 2007, p. 297)

$$\mathrm{TPM}_W = \frac{1}{2}\left(\Phi\left(-\frac{\Delta}{2}\right)+\Phi\left(-\frac{\Delta}{2}\right)\right)=\Phi\left(-\frac{\Delta}{2}\right) \tag{6}$$

where $\Phi(.)$ is the cumulative probability function of the standard normal distribution and Δ is the Mahalanobis distance between the two mean vectors $\boldsymbol{\mu}_1$ and $\boldsymbol{\mu}_2$, i.e.

$$\Delta^2\left(\boldsymbol{\mu}_1,\boldsymbol{\mu}_2\right)=\left(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2\right)'\boldsymbol{\Sigma}^{-1}\left(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2\right) \tag{7}$$

Clearly, the Bayes classification rule is equivalent to classification between the two groups by the minimum Mahalanobis distance. In this case, the discrimination variable W can be obtained if we equate the squared Mahalanobis distance between group means and observation $\mathbf{y}$, i.e. $\Delta^2(\mathbf{y}, \boldsymbol{\mu}_1) = \Delta^2(\mathbf{y}, \boldsymbol{\mu}_2)$.

The squared Euclidean distances between means of the two groups is defined by

$$\delta^2\left(\boldsymbol{\mu}_1,\boldsymbol{\mu}_2\right)=\left(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2\right)'\left(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2\right) \tag{8}$$

The minimum Euclidean distance of observation $\mathbf{y}$ from the group means, i.e. $\delta^2(\mathbf{y}, \boldsymbol{\mu}_1) = \delta^2(\mathbf{y}, \boldsymbol{\mu}_2)$, can be used for discrimination. In this case the discriminant variable is given by

$$W' = \left(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2\right)'\left(\mathbf{y}-\frac{1}{2}\left(\boldsymbol{\mu}_1+\boldsymbol{\mu}_2\right)\right) \tag{9}$$

The TPM of W' is given by

$$\mathrm{TPM}_{W'} = \frac{1}{2}\left(\Phi\left(\frac{-\delta^2}{2\Delta_1}\right)+\Phi\left(\frac{-\delta^2}{2\Delta_1}\right)\right)=\Phi\left(\frac{-\delta^2}{2\Delta_1}\right) \tag{10}$$

where the notation Δ_1^2 is defined in equation (7) by replacing $\boldsymbol{\Sigma}^{-1}$ with $\boldsymbol{\Sigma}$ (for more details see Appendix A).

In practice, the parameters $\boldsymbol{\mu}_1$, $\boldsymbol{\mu}_2$, and $\boldsymbol{\Sigma}$ are unknown. Estimate these parameters by means of independent random "training samples". Suppose we have N_1 observations $y_1^{(1)},\ldots y_{N_1}^{(1)}$ drawn from Π_1 and N_2 observations $y_1^{(2)},\ldots y_{N_2}^{(2)}$ drawn

from Π_2, where N_1, $N_2 > d$. We estimate $\boldsymbol{\mu}_1$, $\boldsymbol{\mu}_2$ by the unbiased sample means $\bar{\mathbf{y}}_1$ and $\bar{\mathbf{y}}_2$, respectively, and estimate $\boldsymbol{\Sigma}$ by the pooled sample covariance matrix $\mathbf{S}_p$. Then the discriminant functions (5) and (9) can be modified as $\hat{W}$, and $\hat{W}'$ yields a "plug-in" discriminant functions. The sample distribution of $\hat{W}$ has been studied by several authors (Atakan, 2009).

The stochastic representations for the exact distribution of $\hat{W}$ in terms of elements of two independent 2×2 central and non-central Wishart matrices has been studied by Bowker (1961). Similarly, we can obtain stochastic representations for the plug-in discriminant function $\hat{W}'$ in terms of a sum of weighted independent non-central chi-square random variables where the weights are paired with different sign (see Appendix B).

Density Estimation Based on Moments

Maximum Entropy Density Estimation

Let X be a continuous variable with probability density function f(x). Then Shanon's entropy is defined by

$$\mathrm{H}(q) = -\int_{-\infty}^{\infty} \mathrm{f}(q) \log \mathrm{f}(q)\, dq \tag{11}$$

where

$$\int_{-\infty}^{\infty} \mathrm{f}(q)\, dq = 1$$

A maximum entropy density function can be obtained by maximizing the Shanon's entropy subject to known moment constraints, i.e.

$$\mu_j' = \int_{-\infty}^{\infty} q^j \mathrm{f}(q)\, dq, \quad j = 0,1,\ldots,k \tag{12}$$

where k is the number of known moments. So, to get the maximum entropy density, we have a non-linear system of equations. The solution of this system is a maximum entropy density function that can be obtained by a variational principle used for f(q) according to the Lagrange method (Singh, 2013). The maximum entropy density is an exponential function is given by

$$\hat{\mathrm{f}}\left(x \mid k\right) = \exp\left(-\theta_0 - \theta_1 q - \ldots - \theta_k q^k\right) \tag{13}$$

where θ_1, θ_2,…, θ_k are determined so that (13) is a proper density function and satisfies all the $k + 1$ moment constraints. With the use of Lagrange's method to solve the non-linear equation system, $k + 1$ Lagrage parameters are obtained by solving $k + 1$ non-linear equations which maximum entropy density function (13) substituting into equation (12), i.e.

$$\mathrm{g}_j\left(\theta\right) = \int_{-\infty}^{\infty} q^j \hat{\mathrm{f}}\left(q \mid k\right) dq = \mu_j', \quad j = 0,1,\ldots,k \tag{14}$$

The solution of equation (14) can be obtained through numerical methods such as the Newton-Raphson algorithm and dual approach (Golan, Judge, & Miller, 1996). In this paper, the Newton-Raphson's algorithm is used for obtaining the Lagrange vector $\boldsymbol{\theta} = (\theta_0,\ldots, \theta_k)'$. Let $\boldsymbol{\theta}^0$ be the initial vector, which may have elements that are zero. The first step is to compute the first order Taylor expansion of $\mathrm{g}_j(\boldsymbol{\theta})$ around the initial trial $\boldsymbol{\theta}^0$; we have

$$\boldsymbol{\theta}^1 = \boldsymbol{\theta}^0 + \mathbf{G}^{-1}\mathbf{v} \tag{15}$$

where the elements of vector $\mathbf{v}$ are given by

$$v_j = \mu_j' - \mathrm{g}_j\left(\boldsymbol{\theta}^0\right), \quad j = 0,1,\ldots,k$$

and the elements g_{ij} of the Hessian matrix $\mathbf{G}$ are

$$g_{ij} = g_{ji} = \int_{-\infty}^{\infty} x^{i+j} \hat{\mathrm{f}}\left(q \mid k\right) dq, \quad i, j = 0,1,\ldots,k$$

Because the Hessian matrix is positive definite in each iteration of the algorithm, equation (15) has a unique solution for $\boldsymbol{\theta}^1$. In each iteration, the current estimation replaces $\boldsymbol{\theta}^0$ and the non-linear equations will be solved again. The iterative algorithm stops when the Euclidean norm of the difference between the current and previous parameter values is less than a small value, for example $\left\|\boldsymbol{\theta}^1 - \boldsymbol{\theta}^0\right\| \leq 10^{-5}$.

The determination of the number of constraints is important to the selection of the optimal density. If the number of constraints are taken to be over-large or over-small, then the accuracy of the estimated density function will be lost (Zong,

2006). Theoretically, among calculated densities, the function with the lowest entropy is selected. It should be mentioned, for a variable that is known information such as its moments, lower entropy indicates less uncertainty to the probability distribution of the variable. However, an increase in the number of constraints leads to a decrease of entropy (Singh, 2013), i.e.

$$\mathrm{H}_{k_2}(q) \le \mathrm{H}_{k_1}(q), \quad k_2 > k_1$$

The Kullback-Leibler and information discrimination are other criteria that can be used to select a maximum entropy density function (Soofi, Ebrahimi, & Habibullah, 1995). The stochastic representation of the quadratic form in equation (2) can be performed by a Monte Carlo simulation to estimate percentiles of the distribution. Therefore, to determine the constraints *k*, the density is chosen in such a way that its percentiles have the minimum distance to the empirical percentiles obtained by Monte Carlo simulations. In other words, for $\alpha = 0.01, 0.02, \ldots, 0.99$, the percentiles t_α of the empirical distribution function are calculated by Monte Carlo simulations. Percentiles of distribution for the known constraints *k* are calculated by

$$\mathrm{F}_k(t_\alpha) = \int_{q_1}^{t_\alpha} \hat{\mathrm{f}}(q \mid k)\, dq \tag{16}$$

where $\hat{\mathrm{f}}(q \mid k)$ is estimated by the density function (13). In this study, the lowest integral bound is determined by the minimum observation in the Monte Carlo simulation. Therefore, the number of constraints *k* is chosen so that the following Euclidean distance has the minimum value:

$$\hat{k} = \min \left\| \mathrm{F}_k(t) - \mathrm{G}(t) \right\|, \quad k = 1, 2, \ldots \tag{17}$$

where G(*t*) is the empirical distribution.

In summary, to approximate the distribution of quadratic forms by maximum entropy estimation, the following algorithm can be utilized to approximate the density function:

Step 1. Using the right-hand side of equation (2), we can perform a Monte Carlo simulation with 10^6 iterations.
Step 2. The theoretical moments are obtained by equation (3).
Step 3. The maximum entropy density function is obtained by (13).

Step 4. The number of constraints $\hat{k}$ will be estimated by equation (17).
Step 5. Cumulative probability (16) of the point t is calculated using numerical integration.

Modified Pearson's Approximation

When the quadratic form is indefinite some weights can be negative and, by equation (4), the third cumulant κ_3 will have a negative value. The skewness coefficient, i.e. $\beta_1 = \kappa_3 / \kappa_2^{3/2}$, is negative. Therefore, it is unsuitable to use gamma and chi-square distributions to approximate the distribution of indefinite quadratic forms directly. In this case, the proposed methods by Houshmand (1993) are remarkable. However, such positive distributions may be modified to approximate the distribution of indefinite quadratic forms. For example, in modified Pearson's tree moments, we write $Q; b\chi_\nu^2 + a$, where the symbol (;) means "is approximately distributed as". Equate the first three cumulants on both sides and determine *a*, *b*, and *v*, respectively. If we define

$$c_r = \sum_{j=1}^{d} \lambda_j^r \left(1 + r\delta_j^2\right)$$

then $a = -\left(c_2^2 / c_3\right) + c_1$, $b = c_3 / c_2$, and the degrees of freedom is estimated as $\nu = c_2^3 / c_3^2 = 8 / \beta_1^2$. Since the degrees of freedom v can be fractional, for computational purposes we use a gamma distribution with shape and scale parameters $v / 2$ and $|b|$. As a result, the tail probability is given by (Houshmand, 1993)

$$\mathrm{P}\left(Q \le t\right) \approx \begin{cases} \mathrm{P}\left(Y \le t - a\right) & \text{if } b > 0 \\ 1 - P\left(Y < a - t\right) & \text{if } b < 0 \end{cases} \tag{18}$$

where $Y = |b|\, \chi_\nu^2\left(\delta_i^2\right)$.

Numerical Examples

Two artificial examples are provided to illustrate the usefulness of the maximum entropy density estimation method. For the well-known Egyptian skull data (Manly, 1994), the estimation methods of misclassification probabilities were compared. It

is further noted that Mathematica software was used to write program codes for numerical calculation. The package is available from the authors upon request.

Model Selection

In this example, a quadratic form is demonstrated as a weighted sum of non-central chi-square variables. The criteria (11) and (17) are compared to obtain the most appropriate probability density of the maximum entropy. Suppose in equation (1) the variable **X** has the multivariate normal distribution with mean $\boldsymbol{\mu}$ = (7, 12, -3), covariance

$$\boldsymbol{\Sigma} = \begin{bmatrix} 2 & -1 & 1 \\ -1 & 6 & 1 \\ 1 & 1 & 2 \end{bmatrix}$$

and matrix of quadratic form given by

$$\mathbf{A} = \begin{bmatrix} 2 & -5 & -1 \\ -5 & 1 & 1 \\ -1 & 1 & 6 \end{bmatrix}$$

The weights of the non-central chi-square variables are λ_1 = 28.3786, λ_2 = 9.4582, and λ_3 = -5.8833. The non-centrality parameters of the non-central chi-square variables are $\gamma_1^2 = 0.161495$, $\gamma_2^2 = 28.5401$, and $\gamma_3^2 = 144.965$. In Table 1, different models are given to obtain of the maximum entropy density function. The Shanon's entropy decreased with an increasing number of constraints. Minimal change is between 4 to 9 constraints. From the table, k = 12 minimizes H, meaning that the best density is given by 12 constraints. The distance D is minimized by 9 constraints (D = 0.00265). Clearly, by obtained information about H and D, the constraints k = 9 provides the suitable maximum entropy density function.

An Exact Density of Quadratic Form

Imhof (1961) obtained the exact distribution of the indefinite quadratic form where the weights of the central chi-square variables are paired and, in each pair, the weights are equal. In this particular case, suppose that the eigenvalues are $\lambda_1 = \lambda_2 = 16$, $\lambda_3 = \lambda_4 = 7$, $\lambda_5 = \lambda_6 = 3$, $\lambda_7 = \lambda_8 = 1$, $\lambda_9 = \lambda_{10} = -2$, $\lambda_{11} = \lambda_{12} = -6$,

$\lambda_{13} = \lambda_{14} = -14$, $\lambda_{15} = \lambda_{16} = -32$, and $\lambda_{17} = \lambda_{18} = -70$. The stochastic representation of the indefinite quadratic form is given by

$$Q_1 = 16\chi_2^2 + 7\chi_2^2 + 3\chi_2^2 + \chi_2^2 - 2\chi_2^2 - 6\chi_2^2 - 14\chi_2^2 - 32\chi_2^2 - 70\chi_2^2$$

Table 1. Model selection values for number of constraints *k*; Shanon's entropy H; and proposed method *D*

k	H	*D*	*k*	H	*D*
1	7.49458	1.91482	7	6.61249	0.00283
2	6.61282	0.01816	8	6.61248	0.00294
3	6.61276	0.01754	9	6.61248	0.00265
4	6.61249	0.00424	10	6.61245	0.0036
5	6.61249	0.00281	11	6.61243	0.00458
6	6.61249	0.00277	12	6.61198	0.01108

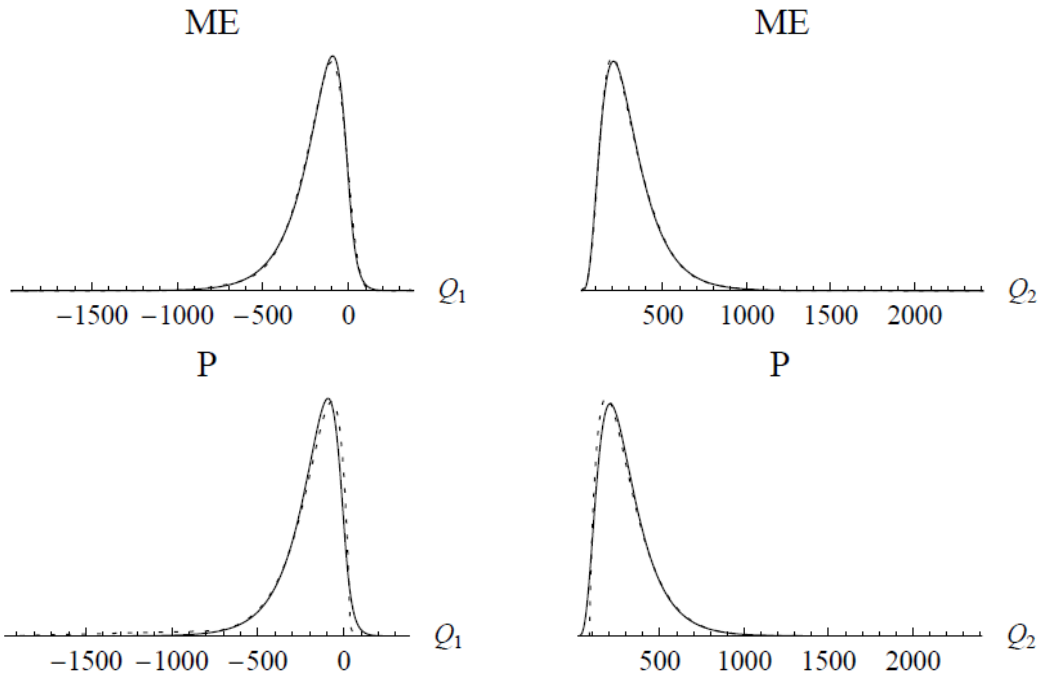

Figure 1. Exact density (solid line) and approximated methods (dotted): maximum entropy (ME) and Pearson's method (P)

Considering the absolute value of weights there is a definite quadratic form Q_2. The best maximum entropy density function of Q_1 is obtained by the constraints $k = 10$, where the estimated criteria are H = 6.35872 and $D = 0.029954$. The discrepancy criterion D are estimated to be 0.109882 and 0.110786 for the modified Pearson's methods. The definite quadratic form Q_2 has minimum Shanon's entropy at $k = 11$, while the minimum value of the discrepancy criterion is 0.020173 for $k = 10$. The discrepancies are estimated for the Pearson's methods to be 0.006198 and 0.088602. Figure 1 displays the exact density functions of the quadratic forms Q_1 and Q_2 alongside the three methods of approximation. For the indefinite quadratic form Q_1, the density approximated the maximum entropy method is quite accurate given the skewness of the exact density. The maximum entropy density is more accurate than the Pearson's method. In the definite form, the Pearson's method for upper tail is a suitable approximation.

Egyptian Skulls Data

The Egyptian skull data consists of four measurements: maximal breadth (X_1), basibregmatic height (X_2), basialveolar length (X_3), and nasal height (X_4), which were measured on skulls of ancient Egyptian males from five different time periods (4000 BC, 3300 BC, 1850 BC, 200 BC, 150 AD). Also, each time period consists of 30 observations. This data can be found in many books on applied multivariate statistical methods, e.g. Manly (1994). In this study, the observations are categorized into two groups: BC and AD. To determine whether or not the underlying assumption of normality of the groups is satisfied, Mardia's multivariate skewness statistic (Mardia, 1974) applied to test the hypothesis that each training sample is drawn from a multivariate normal population. By this statistic, the two groups BC and AD are normal at the 5% level. The test failed to reject the null hypothesis in either case. The smallest p-value for the Mardia's skewness statistic for either Π_1 or Π_2 was $p = 0.181$. The homogeneity of covariance matrices is evaluated with Box's M test, which was not significant at the 5% level ($M = 11.015$, $p = 0.403$). Therefore, the assumption of homogeneity of covariance matrices was satisfied. Hotelling's T^2 statistic is used to test for differences between means of two groups. The equality of means is rejected at the 1% level ($T^2 = 23.4985$, $p = 0.00025$). Therefore, the three assumptions of the linear discriminant analysis are satisfied.

Summarized in Table 2 are the misclassification probabilities of the Egyptian skull data. The apparent error rate (APER) is defined as the fraction of observations in the training sample that are classified by the plug-in discriminant function. APER

Table 2. Estimated misclassification probabilities for Egyptian skull data

Method	P(AD | BC)	P(BC | AD)	TPM
APER of $\hat{W}$	0.291667	0.400000	0.345833
APER of $\hat{W}'$	0.275000	0.366667	0.320833
Normal			0.312231
Monte-Carlo simulation of $\hat{W}'$	0.317874	0.317848	0.317861
ME	0.317860	0.318204	0.318032

does not depend on the distribution of the groups and that can be calculated for any classification procedure. Unfortunately, APER tends to underestimate the actual error rate and the problem does not disappear unless the sample sizes N_1 and N_2 are very large (Jhonson & Wichern, 2007, p. 598 ff.). From the Table 2, we see that the estimates of P(AD | BC), P(BC | AD), and TPM obtained by the APER of the discriminant function $\hat{W}$ are 0.275, 0.366667 and 0.320833, respectively. These estimates are less than the corresponding estimates obtained by $\hat{W}$.

The sample variance and squared Euclidean distance were estimated to be $\hat{\Delta}_1 = 0.989498$ and $\hat{\delta}^2 = 29.1546$, respectively. Estimate the misclassification probability of the plug-in discriminant function $\hat{W}'$ by normal approximation, TPM = 0.312231. The stochastic representation of $\hat{W}'$, if $y \in \Pi_1$, is given by

$$\hat{W}_1' = 3.224\chi_1^2(12.079) + 2.426\chi_1^2(1.4498) + 1.952\chi_1^2(0.4005) + 0.954\chi_1^2(0.327) - 0.954\chi_1^2(0.218) - 1.952\chi_1^2(0.266) - 2.426\chi_1^2(0.965) - 3.224\chi_1^2(8.036)$$

and, if $y \in \Pi_2$, then

$$\hat{W}_1' = 3.224\chi_1^2(8.036) + 2.426\chi_1^2(1.4498) + 1.952\chi_1^2(0.266) + 0.954\chi_1^2(0.218) - 0.954\chi_1^2(0.327) - 1.952\chi_1^2(0.4005) - 2.426\chi_1^2(0.965) - 3.224\chi_1^2(12.079)$$

since $\hat{W}_1' = -\hat{W}_2'$, we have P(AD | BC) = P(BC | AD). Estimate the TPM by performing 10^6 iterations of a Monte Carlo simulation of the stochastic representation $\hat{W}_1'$. From Table 2, note the estimate of TPM through Monte Carlo simulations is 0.317861. The maximum entropy method approximated the best density for $\hat{W}_1'$ and $\hat{W}_2'$ with $k = 9$. As can be seen from the table, the estimation of

misclassification probabilities obtained by maximum entropy method and Monte Carlo simulation are close to each other.

Conclusion

Determining the exact distribution of the indefinite quadratic form in normal variables is complicated, and its distribution remains an area of investigation. Fortunately, higher order moments of the quadratic form can be computed, and then the approximate distribution based on the moments approach. The maximum entropy density approximation method can be an alternative method to approximate the distribution. Despite its versatility, the maximum entropy density has not been widely used in empirical studies. One possible reason is that there is generally no analytical solution for the maximum entropy density problem and numerical estimation, such as the Newton-Raphson's algorithm, must be used. However, new computer systems are provided and computational speed has increased significantly. We proposed a criterion for selecting the number of constraints by the discrepancy between the approximated maximum entropy density and the empirical distribution of the stochastic representation of the quadratic form. This criterion can be used to select the maximum entropy density when it has minimum value. The results of the examples reveal that the approximated density of indefinite quadratic forms via maximum entropy are suitable.

References

Atakan, C. (2009). Bootstrap percentile confidence intervals for actual error rate in linear discriminant analysis. *Hacettepe Journal of Mathematics and Statistics, 38*(3), 357-372. Retrieved from http://www.hjms.hacettepe.edu.tr/uploads/6a0741d6-290b-4f18-b16d-a7b0351fba9b.pdf

Bowker, A. H. (1961). A representation of Hotelling's T^2 and Anderson's classification statistic W in terms of simple statistics. In H. Solomon (Ed.), *Studies in item analysis and prediction* (pp. 285-292). Stanford, CA: Stanford University Press.

Golan, A., Judge, G. G., & Miller, D. (1996). *Maximum entropy econometrics: Robust estimation with limited data*. New York, NY: John Wiley.

Houshmand, A. A. (1993). Misclassification probabilities for quadratic discriminant function. *Communication in Statistics – Simulation and Computation, 22*(1), 81-98. doi: 10.1080/03610919308813082

Imhof, J. P. (1961). Computing the distribution of quadratic forms in normal variables. *Biometrika, 48*(3/4), 419-426. doi: 10.2307/2332763

Jaynes, E. T. (1957). Information theory and statistical mechanics. *Physics Review, 106*(4), 620-630. doi: 10.1103/physrev.106.620

Johnson, R. A., & Wichern, D. W. (2007). Applied multivariate statistical analysis. Upper Saddle River, NJ: Pearson Prentice Hall.

Liu, H., Tang, Y., & Zhang, H. H. (2009). A new chi-square approximation to the distribution of non-negative definite quadratic forms in non-central normal variables. *Computational Statistics & Data Analysis, 53*(4), 853-856. doi: 10.1016/j.csda.2008.11.025

Manly, B. F. J. (1994). *Multivariate statistical methods: A primer*. New York, NY: Chapman & Hall.

Mardia, K. V. (1974). Applications of some measures of multivariate skewness and kurtosis in testing normality and robustness studies. *Sankhyā: The Indian Journal of Statistics, Series B, 36*(2), 115-128. Available from http://www.jstor.org/stable/25051892

Mardia, K. V., Kent, T. J., & Bibby, J. (1979). *Multivariate analysis*. London, UK: Academic Press.

Mathai, A. M., & Provost, S. B. (1992).*Quadratic forms in random variables, Theory and applications*. New York, NY: Marcel Dekker.

Mohsenipour, A. A., & Provost, S. B. (2011). On approximating the distribution of indefinite quadratic expressions in singular normal vectors. *Acta et Commentationes Universitatis Tartuensis de Mathematica, 15*(1), 61-86. Retrieved from http://math.ut.ee/acta/15-1/MohsenipourProvost1.pdf

Patnaik, P. B. (1949). The non-central χ^2- and *F*-distributions and their application. *Biometrika, 36*(1/2), 202-232. doi: 10.2307/2332542

Pearson, E. S. (1959). Note on an approximation to the distribution of non-central χ^2. *Biometrika, 46*(3/4), 364. doi: 10.2307/2333533

Singh, V. P. (2013). *Entropy theory and its application in environmental and water engineering*. New York, NY: John Wiley.

Smith, P. J. (1995). A recursive formulation of the old problem of obtaining moments from cumulants and vice versa. *The American Statistician, 49*(2), 217-219. doi: 10.2307/2684642

Soofi, E., Ebrahimi, N., & Habibullah, M. (1995). Information distinguishability with application to analysis of failure data. *Journal of the American Statistical Association, 90*(43), 657-668. doi: 10.2307/2291079

Zong, Z. (2006). *Information-theoretic methods for estimating complicated probability distributions*. Boston, MA: Elsevier.

Appendix A: TPM of W'

Let **y** be from the population Π_1. The variable $x = (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)'\mathbf{y}$ has univariate normal distribution with mean $\mu_1 = (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)'\,\boldsymbol{\mu}_1$ and variance $\sigma^2 = \Delta_1^2$, where Δ_1^2 is the squared Mahalanobis distance in (7) by substituting $\boldsymbol{\Sigma}$ for $\boldsymbol{\Sigma}^{-1}$. In this case,

$$\begin{aligned} \mathrm{P}(2\,|\,1) &= P(\mathrm{W}' < 0) = \mathrm{P}\left(x < \frac{1}{2}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)'(\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2) \right) \\ &= \mathrm{P}\left(\frac{x - \mu_1}{\sigma} < \frac{\frac{1}{2}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)'(\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2)' - (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)'\boldsymbol{\mu}_1}{\Delta_1} \right) \\ &= \mathrm{P}\left(z < -\frac{\frac{1}{2}\delta^2}{\Delta_1} \right) \\ &= \Phi\left(-\frac{\delta^2}{2\Delta_1} \right) \end{aligned} \tag{A1}$$

where δ^2 is the squared Euclidean distance in (8). Similarly,

$$\begin{aligned} \mathrm{P}(1\,|\,2) &= \mathrm{P}(\mathrm{W}' \geq 0) \\ &= \mathrm{P}\left(z \geq \frac{\frac{1}{2}\delta^2}{\Delta_1} \right) \\ &= 1 - \Phi\left(\frac{\delta^2}{2\Delta_1} \right) \\ &= \Phi\left(-\frac{\delta^2}{2\Delta_1} \right) \end{aligned} \tag{A2}$$

substituting equations (A1) and (A2) in (5),

$$\mathrm{TPM}_{W'} = \Phi\left(-\frac{\delta^2}{2\Delta_1}\right)$$

In the linear discriminant analysis $\boldsymbol{\mu}_1 \neq \boldsymbol{\mu}_2$, then $\delta^2 > 0$. The TPM of W' is decreasing in Δ_1 if δ^2 is fixed. Therefore it follows that $\mathrm{TPM}_{W'}$ is tending to zero if $\Delta_1 \to 0$.

Appendix B: Stochastic Representations of $\hat{W}'$

The discriminant function $\hat{W}'$ can be written as a quadratic form $\hat{W}' = \mathbf{x}'\mathbf{A}\mathbf{x}$, where $\mathbf{x} = \left(\bar{\mathbf{y}}_1 - \bar{\mathbf{y}}_2, \mathbf{y} - \frac{1}{2}\left(\bar{\mathbf{y}}_1 + \bar{\mathbf{y}}_2\right)\right)$ and

$$\mathbf{A} = \begin{bmatrix} \mathbf{0} & \frac{1}{2}\mathbf{I}_d \\ \frac{1}{2}\mathbf{I}_d & \mathbf{0} \end{bmatrix}$$

If the vector $\mathbf{y}$ is from the population Π_j, i.e. $\mathbf{y} \in \Pi_j$, $j = 1, 2$, then the vector $\mathbf{x}$ is distributed as $N_{2d}(\boldsymbol{\eta}_j, \boldsymbol{\Omega})$ with means $\boldsymbol{\eta}_1 = \left(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2, \frac{1}{2}\left(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2\right)\right)$ and $\boldsymbol{\eta}_2 = \left(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2, -\frac{1}{2}\left(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2\right)\right)$ and covariance matrix

$$\boldsymbol{\Omega} = \begin{bmatrix} \left(\frac{N_1 + N_2}{N_1 N_2}\right)\boldsymbol{\Sigma} & \mathbf{0} \\ \mathbf{0} & \left(\frac{N_1 + N_2 + 4N_1N_2}{4N_1N_2}\right)\boldsymbol{\Sigma} \end{bmatrix}$$

Therefore, apply the results of the quadratic form stochastic representation to estimate the TPM of W'. The eigenvalues of $\boldsymbol{\Omega}^{\frac{1}{2}}\mathbf{A}\boldsymbol{\Omega}^{\frac{1}{2}}$ are equal with the eigenvalues of matrix

$$\boldsymbol{\Omega} = \begin{bmatrix} \mathbf{0} & \left(\frac{N_1 + N_2}{2N_1N_2}\right)\boldsymbol{\Sigma} \\ \left(\frac{N_1 + N_2 + 4N_1N_2}{8N_1N_2}\right)\boldsymbol{\Sigma} & \mathbf{0} \end{bmatrix} \tag{B1}$$

If $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_d > 0$ are the eigenvalues of $\boldsymbol{\Sigma}$, then the characteristic equation of (B1), by applying equation (A.2.3.j) in Mardia, Kent, and Bibby (1979), has the eigenvalues

$$\pm \frac{\lambda_j \sqrt{\left(N_1 + N_2\right)\left(N_1 + N_2 + 4N_1N_2\right)}}{4N_1N_2}$$

where $j = 1, 2, \ldots, d$.