Research Article

Statistics in Medicine

Received 24 June 2010, Accepted 15 December 2010 Published online 21 February 2011 in Wiley Online Library

(wileyonlinelibrary.com) DOI: 10.1002/sim.4196

Equivalence of improvement in area under ROC curve and linear discriminant analysis coefficient under assumption of normality

Olga V. Demler,[a*†] Michael J. Pencina[b] and Ralph B. D'Agostino, Sr.[c]

In this paper we investigate the addition of new variables to an existing risk prediction model and the subsequent impact on discrimination quantified by the area under the receiver operating characteristics curve (AUC of ROC). Based on practical experience, concerns have emerged that the significance of association of the variable under study with the outcome in the risk model does not correspond to the significance of the change in AUC: that is, often the variable is significant, but the change in AUC is not. This paper demonstrates that under the assumption of multivariate normality and employing linear discriminant analysis (LDA) to construct the risk prediction tool, statistical significance of the new predictor(s) is equivalent to the statistical significance of the increase in AUC. Under these assumptions the result extends asymptotically to logistic regression. We further show that equality of variance–covariance matrices of predictors within cases and non-cases is not necessary when LDA is used. However, our practical example from the Framingham Heart Study data suggests that the finding might be sensitive to the assumption of normality. Copyright © 2011 John Wiley & Sons, Ltd.

Keywords: linear discriminant analysis; risk prediction model; AUC; ROC; logistic regression

1. Introduction

Statistical models are powerful tools in modern clinical and preventive medicine. Among their common applications is the assessment of risk for individuals for developing certain types of medical conditions. For example, the 10-year risk of Coronary Heart Disease is estimated using Framingham Risk Score [1], which has been developed as part of the Framingham Heart Study [2–4]. This score is based on a range of factors and medical tests including age, total and HDL cholesterol, systolic blood pressure and smoking status. In cancer research, statistical models are used to select patients in the high-risk category for more precise but also more invasive tests (see, for example, Gail model for 5-year risk of breast cancer [5, 6]).

Generally, the goal of risk prediction procedures is to use information available from the medical tests to distinguish between people who will (cases or events) and will not (non-cases, non-events or controls) develop the medical condition of interest and assign them accurate predicted probabilities for having or developing the condition. When a unique test is available, determining the high-risk group is straightforward. When several medical tests may be performed, combining results into a single composite risk score is more challenging and requires statistical modeling techniques, such as linear discriminant analysis (LDA) or logistic regression. Higher risk scores correspond to higher probabilities

[a] *Department of Biostatistics, Boston University, 801 Massachusetts Avenue, Boston, MA 02118, U.S.A.*
[b] *Department of Biostatistics, Boston University, Harvard Clinical Research Institute, 801 Massachusetts Avenue, Boston, MA 02118, U.S.A.*
[c] *Department of Mathematics and Statistics, Boston University, 111 Cummington Street, Boston, MA 02215, U.S.A.*
[*] *Correspondence to: Olga V. Demler, Department of Biostatistics, Boston University, 801 Massachusetts Avenue, Boston, MA 02118, U.S.A.*
[†] *E-mail: demler@bu.edu*

of developing or having the adverse medical condition. When the composite risk score exceeds a certain threshold value, people can be assigned to the high-risk group.

There is an ongoing search for new predictors that improve the accuracy of risk evaluation algorithms. Numerous biomarkers and genetic factors have been postulated and more are being discovered, proposed and studied. However, what 'improving accuracy' should really mean remains a subject of considerable debate [7]. Before we can quantify improved accuracy, we need to decide which metrics should be used to assess performance of the risk prediction tool at hand. D'Agostino *et al.* [8] give a comprehensive review of performance measures that can be used for models with binary outcomes. Often, these measures are categorized as those focusing on discrimination and calibration. Discrimination measures how well a given model separates events from non-events, whereas calibration focuses on how close the estimated risks are to the observed event rates.

The most commonly used measure of discrimination [8–11] is the area under the receiver operating characteristics curve (ROC), also known as the AUC. This measure gained popularity following Bamber's [9], and Hanley and McNeil's [10] observations that the AUC can be interpreted as the probability that a randomly picked non-event has a lower risk score than a randomly picked event. Following these papers, the AUC has become a well-accepted measure of model performance in many areas of medical diagnostics and prevention. Although in recent years, the AUC has been more scrutinized [12], it still remains widely used, and improvement in the AUC is a popular tool for comparing models. The *c* statistic [13] (corresponding to the Wilcoxon–Mann–Whitney statistic [10] in this case) is commonly used as a non-parametric estimator of the AUC. Researchers usually test statistical significance of a new candidate predictor by adding it to a model with 'standard' risk factors and comparing *c* statistics for this new model to the standard.

Following the relationship present in linear regression, where statistical significance of the regression coefficient is equivalent to the significance in improvement in performance described by the model's coefficient of determination (R^2), it is natural to ask whether such a relationship holds for models with binary outcomes. It is reasonable to hypothesize that variables more strongly associated with the outcome should lead to a greater improvement in model performance. This has been discussed by Ware [14]. However, empirical results do not seem to support this intuitively logical hypothesis: in many practical applications, researchers have observed that the statistical significance of new variables added to particular risk prediction models does not translate into statistically significant or numerically appreciable increase in the *c* statistic. For example, Wang *et al.* [15] demonstrated that a multiple biomarker score can be significantly associated with the risk of cardiovascular disease, but increases the *c* statistic only from 0.76 to 0.77. Pencina and D'Agostino *et al.* [7] give an example in which HDL cholesterol is highly significant in the proportional hazards regression model, but does not significantly increase the *c* statistic.

In this paper we investigate this phenomenon in the context of multivariate normality of predictors (paralleling the context of linear regression). In Section 2 we show that under the assumption of multivariate normality statistical significance of a new variable must lead to a significant improvement in AUC and vice versa. Therefore, we prove the existence of a direct link between significance of the predictor and significance of the change in AUC under the assumptions of this paper. We illustrate this finding using numerical simulations in Section 3 where we also show that it may not hold using real, non-normal data from the Framingham Heart Study. Generalizations of our findings beyond our assumptions as well as extension to logistic regression are discussed in Section 4. We summarize our findings in Section 5.

2. Equivalence property under assumptions of normality and equal covariance matrices

In this paper we show that if data are normally distributed and we need to compare two nested models, then statistical significance of the added predictor variables in the larger of the two models is equivalent to statistically significant change in the AUC. In this section we prove this result for equal covariance matrices and later in Section 4 we extend it to a more general case of unequal, non-proportional covariance matrices.

When data are normally distributed in event and non-event subgroups, LDA is the best method to use. It has been shown in [16] that a linear model with LDA coefficients has the highest AUC among

all other linear models under the assumptions of conditional normality in the subgroups. For this reason we use LDA as a way to construct risk scores and to prove the main result.

Let D be an outcome of interest: with $D=1$ for cases and $D=0$ for non-cases. Our goal is to predict the event status using p test results which we denote as $\boldsymbol{x}=x_1,\ldots,x_p$. Assume D and $\boldsymbol{x}$ are available for N patients. Furthermore, we assume multivariate normality of test results conditional on the disease status: $\mathbf{x}|D=0\sim \mathrm{N}(\boldsymbol{\mu}_0,\Sigma)$ and $\mathbf{x}|D=1\sim \mathrm{N}(\boldsymbol{\mu}_1,\Sigma)$, where $\boldsymbol{\mu}_0$ and $\boldsymbol{\mu}_1$ are vectors of means for the p test results among non-events and events, respectively and Σ is the variance–covariance matrix which we assume to be the same in the two subgroups. Normality assumption and equality of within-group covariance matrices allow us to use traditional LDA solution. Extension of our results for unequal covariance matrices is discussed in Section 4.

Denote difference in group means as $\boldsymbol{\delta}=\boldsymbol{\mu}_1-\boldsymbol{\mu}_0$. We want to test whether the model with p predictors discriminates between the two subgroups better than the model with only the first $p-k$ predictors. Using AUC as a measure of model performance we formulate the following hypothesis:

$$H_0^{\mathrm{AUC}}:\mathrm{AUC}_p=\mathrm{AUC}_{p-k},$$

$$H_a^{\mathrm{AUC}}:\mathrm{AUC}_p\neq\mathrm{AUC}_{p-k}.$$

Here, AUC_p is the AUC of the full model and AUC_{p-k} is the AUC of the model with only the first $p-k$ predictors. Let $\boldsymbol{\alpha}'=(\alpha_1,\alpha_2,\ldots,\alpha_p)$ be the LDA coefficients for the full model and $\boldsymbol{\beta}'=(\beta_1,\beta_2,\ldots,\beta_{p-k})$ be the LDA coefficients for the reduced model with first $p-k$ predictors. Denoting the first $p-k$ coefficients of the full model as $\boldsymbol{\alpha}_{p-k}$ and the remaining k coefficients as $\boldsymbol{\alpha}_k$ we have the following proposition, the main result of our paper:

Proposition
Equality of two AUCs for two nested discriminant models is equivalent to discriminant coefficients equal to zero for variables not shared by the two models:

$$\mathrm{AUC}_p=\mathrm{AUC}_{p-k}\Leftrightarrow\boldsymbol{\alpha}_k=\mathbf{0}.$$

We note that this in turn implies that statistically significant improvement in nested AUCs is equivalent to statistical significance of at least one LDA coefficient among the added predictors.

The complete proof of this proposition is given in the Appendix and relies on the following facts:

1. LDA solution to the discrimination problem given in [17] has the form:

$$\boldsymbol{\alpha}=\Sigma^{-1}\boldsymbol{\delta}. \tag{1}$$

2. AUC for the LDA solution can be written explicitly as in [16]:

$$\mathrm{AUC}=\Phi\left(\sqrt{\frac{\boldsymbol{\delta}'\Sigma^{-1}\boldsymbol{\delta}}{2}}\right), \tag{2}$$

   where $\Phi(\cdot)$ is a standard normal cumulative distribution function.
3. AUC is a function of Mahalanobis D^2 [17].

   In formula (2) $\delta'\Sigma^{-1}\delta$ is the Mahalanobis D^2 a measure of the distance between two multivariate normal distributions [17]. Since $\Phi(\sqrt{\cdot})$ in (2) is a strictly increasing function, AUC is a strictly monotone function of the Mahalanobis distance. So significance of the AUC improvement is equivalent to a significant change in the Mahalanobis distance. Equivalently, this result can be formulated as:

$$\mathrm{AUC}_p=\mathrm{AUC}_{p-k}\Leftrightarrow\Delta_p=\Delta_{p-k}, \tag{3}$$

   where Δ_p, Δ_{p-k} are the Mahalanobis Δ^2 based on models with all p and only first $p-k$ predictors, respectively.
4. Mahalanobis D^2 for the full model (D_p) can be decomposed into a sum of Mahalanobis D^2 for the reduced model (Δ_{p-k}) plus a non-negative term [17]:

$$\Delta_p=\Delta_{p-k}+\Delta_{k|p-k}. \tag{4}$$

Using (3) and (4) the equivalence property follows and can be formulated as:

$$\mathrm{AUC}_p = \mathrm{AUC}_{p-k} \Leftrightarrow \Delta_p = \Delta_{p-k} \Leftrightarrow \boldsymbol{\alpha}_k = \mathbf{0}.$$

In other words: significance of predictor variables, significance of the change in Mahalanobis D^2 and significance of the change in AUC are all equivalent.

From the proposition follows a simple corollary: statistical significance of the change in AUC under the LDA assumptions can be tested by the well-known F-test [17]. This provides a simple way of determining significance and produces a result that mirrors the relationship of R^2 and model coefficient in linear regression. We emphasize the importance of using the F-test to check the AUC improvement, as the F-test is applicable to finite size samples and is exact in this problem. We note that it follows from the proof that AUC is non-decreasing for nested models and thus the hypothesis reduces to:

$$H_0^{\mathrm{AUC}} : \mathrm{AUC}_p = \mathrm{AUC}_{p-k},$$

$$H_a^{\mathrm{AUC}} : \mathrm{AUC}_p > \mathrm{AUC}_{p-k}.$$

The proof presented in this section relies on true parameters: group means and variance–covariance matrix. Extension to the case of unknown values of true parameters is presented in Section 4.

3. Numerical examples

To illustrate the results of the previous section we perform numerical simulations. Two five-dimensional multivariate normal vectors with the same correlation structure given as

$$\begin{pmatrix} 1.0 & .06 & .37 & .44 & .20 \\ & 1.0 & .07 & -.06 & -.12 \\ & & 1.0 & .25 & .18 \\ & & & 1.0 & .74 \\ & & & & 1.0 \end{pmatrix}$$

were simulated: one for cases (size 640) and one for non- cases (size 8365). The vector of means for non-cases was set to zero, whereas for cases it was (0.65 −0.47 0.48 0.62 0.42). These parameters correspond to the correlation structure and effect sizes of the Framingham Heart Study data described in the next paragraph. The first $p-k=4$ variables were used for the reduced model and all $p=5$ for the full model. Theoretical AUCs were computed according to formula (2) and discriminant analysis coefficients were estimated using (1). Following the corollary of the previous section, significance of the increase in AUC was tested with an F-test. For comparison, we applied bootstrap [18–20] with 1000 replications to obtain the p-value for testing significance of the discriminant coefficient. This process was repeated 1000 times and p-values from the F-test (comparing AUC) and bootstrap of the coefficient are plotted in Figure 1. As expected, they are nearly identical (the F-test p-values differed from the bootstrap p-values of significance of the coefficient only in the second or third decimal places), forming a 45° line in the plot, illustrating that testing significance of the discriminant coefficient is equivalent to testing the increase in AUC. Of note, a theoretical proof does not need empirical validation so these simulation results have been presented here primarily as an illustration but can also be viewed as a validation of the bootstrap approach to testing the significance of discriminant coefficient.

Our results, proven and illustrated under theoretical multivariate normality, leave the question of how robust they might be under some mild violations of normality occurring in practical applications. For this purpose, we investigate the issue in an example from the Framingham Heart Study data. A total of 8365 observations on people free of cardiovascular disease at baseline examination in the 1970s were available. Measurements of risk factors and results of medical tests were obtained, including age, total and HDL cholesterol, systolic and diastolic blood pressure (DBP). Participants of the study were followed for 12 years for the development of coronary heart disease (CHD) and were categorized as cases if they developed the CHD or non-cases otherwise. To correct for skewness of the predictors, simple

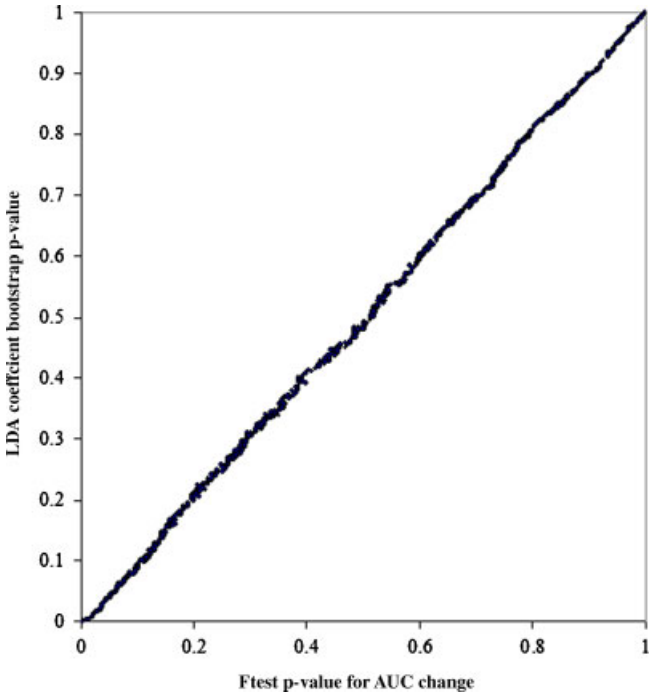

Figure 1. P-value from bootstrap test of discriminant coefficient versus p-value from F-test for the change in AUC for 1000 simulated multivariate normal data sets. Two models were fit to each data set: one with all five predictors and the other with the 5th predictor omitted.

Table I. Significance of change in AUC versus significance of discriminant coefficient based on real data with mild violation of normality.

Excluded variable	Full model AUC	Reduced model AUC	F-test p-value	Discriminant coefficient	Bootstrap p-value for coefficient
Age	0.765	0.738	<.001	1.36	<0.001
HDL cholesterol	0.765	0.731	<0.001	−0.94	<0.001
Total cholesterol	0.765	0.753	<0.001	0.90	<0.001
Systolic blood pressure	0.765	0.762	0.003	0.96	<0.001
Diastolic blood pressure	0.765	0.765	0.155	0.46	0.01

logarithmic transformations were applied. The resulting distributions were unimodal with moderate degree of skewness; however, normality could still be rejected using the Shapiro–Wilks test [21, 22]. Variance–covariance matrices between cases and non-cases could not be assumed identical, so LDA with separate matrices was used (5). Our full model included age, total and HDL cholesterol, systolic and DBP, as predictors of CHD. Five reduced models were considered, omitting one of the predictors each time. Theoretical AUCs based on formula (2) and F-test-based p-values for the difference between AUC of full and reduced model as well as the LDA coefficient and its bootstrap-based p-values are presented in Table I.

We observe that for variables with a strong association with the outcome, both the F-test and bootstrap p-value of the coefficient produce identical results. However, when the association is weaker, as in the case of logarithm of DBP—the relationship no longer holds: the F-test gives a p-value of 0.155, whereas the test of significance of the discriminant coefficient has a p-value of 0.009. The likely cause of this discrepancy lies in violation of normality rather than in the inequality of the variance–covariance matrices in cases and non-cases, as is illustrated in the following section.

4. Generalizations

In the previous section we saw that the equivalence in significance of AUC improvement and LDA coefficient works perfectly when all the assumptions are satisfied and may not work when they are not.

It was not clear whether a violation of normality or equality of the variance–covariance matrices is the problem. It turns out that equality of the variance–covariance matrices for cases and non-cases is not necessary. It can be shown easily that the result holds for proportional covariance matrices. Moreover, in the case of non-proportional covariance matrices, Su and Liu [16] have showed that an optimal solution to the discrimination problem is given by

$$\boldsymbol{\alpha}=(\Sigma_{D=0}+\Sigma_{D=1})^{-1}\boldsymbol{\delta}, \tag{5}$$

where $\Sigma_{D=0}, \Sigma_{D=1}$ denote the variance–covariance matrices for non-cases and cases, respectively. This solution maximizes the AUC over all possible values of linear discriminant coefficients $\boldsymbol{\alpha}$ and so it is the best solution if we believe that the AUC is the most appropriate measure of discrimination. When matrices are proportional or equal, Su and Liu's LDA solution reduces to the traditional linear discriminant solution. The equivalence property can then be extended to the case of unequal, non-proportional covariance matrices, provided we estimate the linear coefficients by Su and Liu's formula (5). The complete proof is given in Appendix B.

The proofs presented in this paper rely on true parameters: variance–covariance matrices and group means. In most cases we do not know the value of the true parameters and they should be estimated. We know that transformation by a continuous function preserves consistency [23]. All functions used in the proof of equivalence proposition are continuous. Thus, it suffices to use consistent estimators for means and within-group variance–covariance matrices and all results will hold asymptotically in probability.

Under the assumptions of normality and equal covariance matrices in the subgroups, the equivalence property holds asymptotically in probability for logistic regression. Indeed, under these assumptions, according to [24] Lemmas 2 and 3, linear coefficients, estimated by logistic regression, and linear coefficients produced by LDA with an estimated mean and covariance matrix converge in probability to the same vector of true coefficients $\boldsymbol{\alpha}$. Since the AUC can be written as $\text{AUC}=\Pr(\boldsymbol{\alpha}'\mathbf{x}_{D=0}<\boldsymbol{\alpha}'\mathbf{x}_{D=1})$, LDA and logistic regression have the same true AUC. So it follows that if we use any consistent estimator for the AUC, the equivalence property stated in this paper holds asymptotically in probablity for logistic regression as well under the above assumptions. The commonly used c statistic is an example of a consistent non-parametric estimator of the AUC [25]. Alternatively, if we use normality of the data, we can calculate the AUC as $\text{AUC}=\Phi(\boldsymbol{\alpha}'(\boldsymbol{\mu}_1-\boldsymbol{\mu}_0)/\sqrt{2\boldsymbol{\alpha}'\Sigma\boldsymbol{\alpha}})$. If we replace the vector of true linear coefficients with its logistic regression estimator (a consistent estimator), the resulting expression becomes a consistent estimator of the true AUC.

Finally, we note that even though the examples of Section 3 focused on one predictor added to the risk prediction model, the results are valid for any number of additional predictors, as presented in Section 2.

It is not clear how sensitive our results are to the normality assumption. Our example implies that it may not always hold for mildly skewed unimodal distributions. In the literature several authors noted that the LDA solution can be extended to a more general elliptically contoured family of distribution functions [16, 26]. Whether the same is true for the equivalence property requires further research. Furthermore, we did not address the practically important situations, where variables comprising the risk score and the new predictor(s) come from different distributions (for example, adding normal or categorical predictors to risk scores based on a combination of normal and categorical variables).

5. Summary

Previous research employing the AUC has unearthed a practical concern that the significance of association of an additional variable in a risk model does not correspond to the significance of the change in the AUC: often the variable is significant, but the change in the AUC is not. In this paper we studied this situation under the assumption of multivariate normality. In this context we used the optimal LDA approach to prove that under these assumptions the significance of the AUC improvement is equivalent to the significance of the new predictor variable(s). The result was illustrated by empirical simulations of multivariate normal data. We have also shown that the equality of the variance–covariance matrices is not necessary— proportional matrices suffice, or even if they are not proportional, the result holds using the form of discriminant solution proposed by Su and Liu [16]. Furthermore, we argued that under multivariate normality and equality of covariance matrices, the result holds asymptotically for logistic regression.

Extensions to categorical predictors and even continuous non-normal data require further research. The numeric example in Section 3 for moderately skewed non-normal Framingham Heart Study data shows that results are sensitive to non-normality. The extensions will be complicated by the fact that formula (2) which provides a connection between the theoretical AUC and the Mahalanobis D^2 does not hold for the empirical AUC and non-normal data. The empirical AUC is the popularly known form of the AUC based on ranks of the estimated risks (also known as c statistic or Mann–Whitney statistic). The lack of closed-form solutions to the logistic regression problem further complicates the matter. Thus, the simplicity and elegance of our proposition may not be attainable.

Appendix A

A.1. Proof of equivalence proposition for equal covariance matrices

We adopt notation of Section 2. We further denote differences in group means

$$\boldsymbol{\delta}=\boldsymbol{\mu}_1-\boldsymbol{\mu}_0 \quad \text{as} \begin{bmatrix}\boldsymbol{\delta}_1\\ \boldsymbol{\delta}_2\end{bmatrix},$$

where $\boldsymbol{\delta}_2$ is the vector of differences in the means of the last k predictors.

Su and Liu [16] showed that under the assumption of multivariate conditional normality, LDA produces AUC that can be expressed explicitly as

$$\text{AUC}=\Phi\left(\sqrt{\frac{\boldsymbol{\delta}'\Sigma^{-1}\boldsymbol{\delta}}{2}}\right),$$

where $\Phi(\cdot)$ is a standard normal cumulative distribution function. We can write the common covariance matrix Σ as

$$\Sigma=\begin{bmatrix}\Sigma_{11} & \Sigma_{12}\\ \Sigma_{21} & \Sigma_{22}\end{bmatrix},$$

where Σ_{11} is the variance–covariance matrix for the first $p-k$ predictors, Σ_{12}-covariance matrix of the first $p-k$ predictors with the new k predictors. Since $\Phi(\cdot)$ is a monotone increasing function we have $\text{AUC}_p=\text{AUC}_{p-k}\Leftrightarrow\boldsymbol{\delta}'\Sigma^{-1}\boldsymbol{\delta}=\boldsymbol{\delta}_1'\Sigma_{11}^{-1}\boldsymbol{\delta}_1$. Using Mahalanobis Distance Decomposition [17] we can write the left-hand side of the last equality as

$$\boldsymbol{\delta}'\Sigma^{-1}\boldsymbol{\delta}=\begin{bmatrix}\boldsymbol{\delta}_1\\ \boldsymbol{\delta}_2\end{bmatrix}'\begin{bmatrix}\Sigma_{11} & \Sigma_{12}\\ \Sigma_{21} & \Sigma_{22}\end{bmatrix}^{-1}\begin{bmatrix}\boldsymbol{\delta}_1\\ \boldsymbol{\delta}_2\end{bmatrix}=\boldsymbol{\delta}_1'\Sigma_{11}^{-1}\boldsymbol{\delta}_1+(\boldsymbol{\delta}_2-\Sigma_{21}\Sigma_{11}^{-1}\boldsymbol{\delta}_1)'(\Sigma_{22}-\Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12})^{-1}$$
$$\times(\boldsymbol{\delta}_2-\Sigma_{21}\Sigma_{11}^{-1}\boldsymbol{\delta}_1).$$

The last formula implies that

$$\boldsymbol{\delta}'\Sigma^{-1}\boldsymbol{\delta}=\boldsymbol{\delta}'\Sigma_{11}^{-1}\boldsymbol{\delta}_1\Leftrightarrow(\boldsymbol{\delta}_2-\Sigma_{21}\Sigma_{11}^{-1}\boldsymbol{\delta}_1)'(\Sigma_{22}-\Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12})^{-1}(\boldsymbol{\delta}_2-\Sigma_{21}\Sigma_{11}^{-1}\boldsymbol{\delta}_1)=0.$$

Since matrix $(\Sigma_{22}-\Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12})^{-1}$ is positive definite, then $(\boldsymbol{\delta}_2-\Sigma_{21}\Sigma_{11}^{-1}\boldsymbol{\delta}_1)'(\Sigma_{22}-\Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12})^{-1}(\boldsymbol{\delta}_2-\Sigma_{21}\Sigma_{11}^{-1}\boldsymbol{\delta}_1)$ is a positive-definite quadratic form, which is equal to zero if and only if $\boldsymbol{\delta}_2-\Sigma_{21}\Sigma_{11}^{-1}\boldsymbol{\delta}_1=\mathbf{0}$. Thus, we established that $(\boldsymbol{\delta}_2-\Sigma_{21}\Sigma_{11}^{-1}\boldsymbol{\delta}_1)'(\Sigma_{22}-\Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12})^{-1}(\boldsymbol{\delta}_2-\Sigma_{21}\Sigma_{11}^{-1}\boldsymbol{\delta}_1)=\mathbf{0}\Leftrightarrow(\boldsymbol{\delta}_2-\Sigma_{21}\Sigma_{11}^{-1}\boldsymbol{\delta}_1)=\mathbf{0}$. To complete the proof we need to show that $\boldsymbol{\delta}_2-\Sigma_{21}\Sigma_{11}^{-1}\boldsymbol{\delta}_1=\mathbf{0}\Leftrightarrow\boldsymbol{\alpha}_2=\mathbf{0}$. We recall that $\boldsymbol{\alpha}$ is the solution to the LDA problem and is given by $\boldsymbol{\alpha}=\Sigma^{-1}\boldsymbol{\delta}$. Partitioning $\boldsymbol{\alpha},\boldsymbol{\delta},\Sigma$ and using the notation of Lemma A1 (see below) we can write:

$$\boldsymbol{\alpha}=\Sigma^{-1}\boldsymbol{\delta}=\begin{bmatrix}\Sigma^{11} & \Sigma^{12}\\ \Sigma^{21} & \Sigma^{22}\end{bmatrix}\begin{bmatrix}\boldsymbol{\delta}_1\\ \boldsymbol{\delta}_2\end{bmatrix}=\begin{bmatrix}\boldsymbol{\alpha}_1\\ \boldsymbol{\alpha}_2\end{bmatrix}.$$

So

$$\boldsymbol{\alpha}_2=\mathbf{0}\Leftrightarrow[\Sigma^{21}\Sigma^{22}]\begin{bmatrix}\boldsymbol{\delta}_1\\ \boldsymbol{\delta}_2\end{bmatrix}=\Sigma^{21}\boldsymbol{\delta}_1+\Sigma^{22}\boldsymbol{\delta}_2=\mathbf{0}\Leftrightarrow(\Sigma^{22})^{-1}\Sigma^{21}\boldsymbol{\delta}_1+\boldsymbol{\delta}_2=\mathbf{0}. \tag{A1}$$

By Lemma A1: $\Sigma^{21}=-\Sigma^{22}\Sigma_{21}\Sigma_{11}^{-1}$. Then (A1) is equivalent to $-(\Sigma^{22})^{-1}\Sigma^{22}\Sigma_{21}\Sigma_{11}^{-1}\boldsymbol{\delta}_1+\boldsymbol{\delta}_2=\mathbf{0}$ or $\boldsymbol{\delta}_2-\Sigma_{21}\Sigma_{11}^{-1}\boldsymbol{\delta}_1=\mathbf{0}$. □

Lemma A1 (*see Mardia et al.* [*17*])
Consider a non-singular, $n\times n$ matrix $\mathbf{A}\in\mathbf{R}^{n\times n}$. Decompose $\mathbf{A}$ as

$$\mathbf{A}=\begin{bmatrix}\mathbf{A}_{11} & \mathbf{A}_{12}\\ \mathbf{A}_{21} & \mathbf{A}_{22}\end{bmatrix},$$

where $\mathbf{A}_{11}\in\mathbf{R}^{(n-k)\times(n-k)}$, for some $k<n$, $\mathbf{A}_{12}\in\mathbf{R}^{(n-k)\times k}$, $\mathbf{A}_{21}\in\mathbf{R}^{k\times(n-k)}$, $\mathbf{A}_{22}\in\mathbf{R}^{k\times k}$.

Denote

$$\mathbf{A}^{-1}=\begin{bmatrix}\mathbf{A}^{11} & \mathbf{A}^{12}\\ \mathbf{A}^{21} & \mathbf{A}^{22}\end{bmatrix}.$$

Then

$$\mathbf{A}^{11}=[\mathbf{A}_{11}\ -\mathbf{A}_{12}\ \mathbf{A}_{22}^{-1}\ \mathbf{A}_{21}]^{-1},$$
$$\mathbf{A}^{22}=[\mathbf{A}_{22}\ -\mathbf{A}_{21}\ \mathbf{A}_{11}^{-1}\ \mathbf{A}_{12}]^{-1},$$
$$\mathbf{A}^{12}=-\mathbf{A}^{11}\ \mathbf{A}_{12}\ \mathbf{A}_{22}^{-1}=-\mathbf{A}_{11}^{-1}\ \mathbf{A}_{12}\ \mathbf{A}^{22},$$
$$\mathbf{A}^{21}=-\mathbf{A}_{22}^{-1}\ \mathbf{A}_{21}\ \mathbf{A}^{11}=-\mathbf{A}^{22}\ \mathbf{A}_{21}\ \mathbf{A}_{11}^{-1}.$$

Appendix B

B.1. Equivalence property for unequal covariance matrices

We adopt the same notation as above but this time $\mathbf{x}|D=0\sim\mathrm{N}(\boldsymbol{\mu}_0,\Sigma_{D0})$ and $\mathbf{x}|D=1\sim\mathrm{N}(\boldsymbol{\mu}_1,\Sigma_{D1})$. Su and Liu [16] showed that under the assumption of multivariate normality in the subgroups for any covariance matrices Σ_{D0}, Σ_{D1} LDA coefficients $\boldsymbol{\alpha}=(\Sigma_{D0}+\Sigma_{D1})^{-1}\boldsymbol{\delta}$ produce a linear model with the largest AUC over all other linear models. When matrices are equal then Su and Liu's solution is the same as the traditional LDA solution. We use Su and Liu's solution for coefficients $\boldsymbol{\alpha}$ to prove the equivalence proposition for unequal covariance matrices.

The proof for equal covariance matrices relied upon the observation that the AUC formula for the reduced model uses a submatrix of a variance–covariance matrix that is used in the full model AUC formula. This allowed us to apply decomposition of the Mahalanobis D^2. The same observation is true for Su and Liu's solution for the discrimination problem with no restriction on within-group covariance matrices. The AUC can be expressed explicitly as

$$\mathrm{AUC}_p=\Phi\left(\sqrt{\boldsymbol{\delta}'(\Sigma_{D0}+\Sigma_{D1})^{-1}\boldsymbol{\delta}}\right) \tag{B1}$$

where $\Phi(\cdot)$ is a standard normal cumulative distribution function. Covariance matrices Σ_{D0}, Σ_{D1} can be split into blocks:

$$\Sigma_{Di}=\begin{bmatrix}\Sigma_{Di11} & \Sigma_{Di12}\\ \Sigma_{Di21} & \Sigma_{Di22}\end{bmatrix},$$

$i=0$, 1. Note that Σ_{Di11} is the covariance matrix for the first $p-k$ predictors for $D=i$ group (i is 0 or 1), Σ_{Di12} is the variance–covariance matrix of the first $p-k$ predictors with the new k predictor variables. Notice that in (B1) AUC_p is a monotone function of the Mahalanobis $D^2:\boldsymbol{\delta}'(\Sigma_{D0}+\Sigma_{D1})^{-1}\boldsymbol{\delta}=\boldsymbol{\delta}'\Sigma^{*-1}\boldsymbol{\delta}$ with the matrix $\Sigma^*=\Sigma_{D0}+\Sigma_{D1}$. The AUC of the reduced model can also be written as

a monotone function of the Mahalanobis D^2 but with a submatrix of $\Sigma^* = \Sigma_{D0} + \Sigma_{D1}$: $\mathrm{AUC}_{p-k} = \Phi(\sqrt{\boldsymbol{\delta}_1'(\Sigma_{D011} + \Sigma_{D111})^{-1}\boldsymbol{\delta}_1}) = \Phi(\sqrt{\boldsymbol{\delta}_1'\Sigma_{11}^{*-1}\boldsymbol{\delta}_1})$, where Σ_{11}^* is a submatrix of Σ^*. We again have $\mathrm{AUC}_p = \mathrm{AUC}_{p-k} \Leftrightarrow \boldsymbol{\delta}'\Sigma^{*-1}\boldsymbol{\delta} = \boldsymbol{\delta}_1'\Sigma_{11}^{*-1}\boldsymbol{\delta}_1$. Because Σ_{11}^* is a submatrix of Σ^* we can also apply the decomposition of the Mahalanobis D^2 as in [17]. We can now follow the proof of the proposition for equal covariance matrices line by line with just one modification: we have to replace symbol Σ with Σ^*.

References

1. Executive Summary of the Third Report of the National Cholesterol Education Program (NCEP) expert panel on detection, evaluation, and treatment of high blood cholesterol in adults (Adult Treatment Panel III). *Journal of the American Medical Association* 2001; **285**:2486–2497.
2. Wilson P, D'Agostino R, Levy D, Belanger AM, Silbershatz H, Kannel WB. Prediction of coronary heart disease using risk factor categories. *Circulation* 1998; **97**:1837–1847.
3. Anderson KM, Odell PM, Wilson PWF, Kannel WB. General cardiovascular risk profile for use in primary care. *American Heart Journal* 1991; **121**:293–298.
4. D'Agostino RB, Vasan RS, Pencina MJ, Wolf PA, Cobain MR, Massaro JM, Kannel WB. General cardiovascular risk profile for use in primary care: the Framingham Heart Study. *Circulation* 2008; **117**:743–753.
5. Gail MH, Brinton LA, Byar DP, Corle DK, Green SB, Schairer C, Mulvihill JJ. Projecting individualized probabilities of developing breast cancer for white females who are being examined annually. *Journal of the National Cancer Institute* 1989; **81**:1879–1886.
6. Gail MH, Costantino JP, Pee D, Bondy M, Newman L, Selvan M, Anderson GL, Malone KE, Marchbanks PA, McCaskill-Stevens W, Norman SA, Simon MS, Spirtas R, Ursin G, Bernstein L. Projecting individualized absolute invasive breast cancer risk in African American women. *Journal of the National Cancer Institute* 2007; **99**(23): 1782–1792.
7. Pencina MJ, D'Agostino SrRB, D'Agostino Jr RB, Vasan RS. Evaluating the added predictive ability of a new marker: from area under ROC curve to reclassification and beyond. *Statistics in Medicine* 2008; **27**:157–173.
8. D'Agostino RB, Griffith JL, Schmidt CH, Terrin N. Measures for evaluating model performance. *Proceedings of the Biometrics Section*. American Statistical Association, Biometrics Section: Alexandria, VA. U.S.A., 1997; 253–258.
9. Bamber D. The area above the ordinal dominance graph and the area below the receiver operating characteristic graph. *Journal of Mathematical Psychology* 1975; **12**:387–415.
10. Hanley JA, McNeil BJ. The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology* 1982; **143**:29–36.
11. Pepe MS. *The Statistical Evaluation of Medical Tests for Classification and Prediction*. Oxford University Press: Oxford, 2004; 77–79.
12. Hand DJ. Measuring classifier performance: a coherent alternative to the area under the ROC curve. *Machine Learning* 2009; **77**(1):103–123.
13. Harrel Jr FE. *Regression Modeling Strategies*: *With Applications to Linear Models*, *Logistic Regression and Survival Analysis*. Springer: New York, 2001.
14. Ware JH. The limitations of risk factors as prognostic tool. *The New England Journal of Medicine* 2006; **355**(25): 2616–2617.
15. Wang TJ, Gona P, Larson MG, Tofler GH, Levy D, Newton-Cheh C, Jacques PF, Rifai N, Selhub J, Robins SJ, Benjamin EJ, D'Agostino RB, Vasan RS. Multiple biomarkers for the prediction of first major cardiovascular events and death. *The New England Journal of Medicine* 2006; **355**(25):2631–2639.
16. Su JQ, Liu JS. Linear combinations of multiple diagnostic markers. *Journal of the Acoustical Society of America* 1993; **88**(424):1350–1355.
17. Mardia KV, Kent JT, Bibby JM. *Multivariate Analysis*. Academic Press: London, 1979; 78–79.
18. Efron B, Tibshirani R. *An Introduction to Bootstrap*. Chapman & Hall/CRC: London, Boca Raton, FL, 1993.
19. Tibshirani R, Hall P, Wilson SR. Bootstrap hypothesis testing. *Biometrics* 1992; **48**(3):969–970.
20. Hall P, Wilson SR. Two guidelines for bootstrap hypothesis testing. *Biometrics* 1991; **47**:757–762.
21. Shapiro SS, Wilk MB. An analysis of variance test for normality (complete samples). *Biometrika* 1965; **52**(3/4): 591–611.
22. Pearson ES, D'Agostino RB, Bowman RO. Tests for departure from normality comparison of powers. *Biometrika* 1977; **64**(2):231–246.
23. Casella B, Berger R. *Statistical Inference*. Duxbury: Pacific Grove, CA, 2002; 233–234.
24. Efron B. The efficiency of logistic regression compared to normal discriminant analysis. *Journal of the Americal Statistical Association* 1975; **70**(352):892–898.
25. Lee AJ. *U-Statistics*: *Theory and Practice*. Marcel Dekkel: New York and Basel, 1990.
26. Wakaki H. Discriminant analysis under elliptical populations. *Hiroshima Mathematical Journal* 1994; **24**:257–298.