The Optimal Linear Combination of Multiple Predictors Under the Generalized Linear Models

Hua Jin, Ph.D.[1] and **Ying Lu, Ph.D.**[2]

[1] School of Mathematical Sciences, South China Normal University, Guangzhou, China

[2] Department of Radiology, Department of Epidemiology and Biostatistics, and UCSF Helen Diller Family Comprehensive Cancer Center, University of California, San Francisco, CA, U. S. A

Summary

Multiple alternative diagnostic tests for one disease are commonly available to clinicians. It's important to use all the good diagnostic predictors simultaneously to establish a new predictor with higher statistical utility. Under the generalized linear model for binary outcomes, the linear combination of multiple predictors in the link function is proved optimal in the sense that the area under the receiver operating characteristic (ROC) curve of this combination is the largest among all possible linear combination. The result was applied to analysis of the data from the Study of Osteoporotic Fractures (SOF) with comparison to Su and Liu's approach.

1. Introduction

Multiple alternative diagnostic tests for one disease are commonly available to clinicians, and statistical methods are available for evaluating these alternatives. For instance, multivariate regression models can be used to determine the marginal advantage of additional tests (Richards, 1995), and this approach has been used in osteoporosis research to determine the relative risk and/or odds ratio of osteoporotic frature (Hans, 1999). While these parameters are important for determining future fracture risk or developing clinical predictor rules based on tests, multivariate regression does not directly provide information on test utility or optimizing test combinations.

Because sensitivity and specificity of a diagnostic test depend on the threshold used to define abnormal, the receiver operating characteristic (ROC) curve is often used to assess the utility of a diagnostic test (DeLong, 1988; Pepe, 1997). This methodology has been extended to multivariate tests by constructing a linear function and then evaluating a combined utility based on the ROC of the linear function. Su and Liu (1993) have shown that under a normal distribution assumption, the best linear combination of diagnostic tests to achieve maximum

Corresponding Author: JIN Hua, Ph.D., School of Mathematical Sciences, South China Normal University, Guangzhou 510631, China. Tel: 86-20-85215691; jinh1@163.com.

area under a ROC curve is the linear discriminant function. However, no literature covers the general situation.

In this article, under the generalized linear model, we provide the optimal linear combination among all possibilities under the criterion that the area under the corresponding ROC curve of the combination is maximized, which is named the ROC criterion. The main results are presented in Section 2. Section 3 considers the estimation of the optimal linear combination and its corresponding area under ROC curve. Section 4 demonstrates an application of the proposed method to data from the Study of Osteoporotic Fractures (SOF) with comparison to Su and Liu's approach. Discussion and conclusions are presented in the last section.

2. The Best Linear Combination under the ROC Criterion

In this section it is proved that the linear combination in the link function of a model for binary responses maximizes sensitivity uniformly at any given specificity under the generalized linear model. Here we only present results for a simple case with two predictors, which can be easily generalized to the situation of multiple diagnostic markers.

Let Z be the binary outcome, (X_1, X_2) be the random predictive variables. Suppose the generalized linear model for binary responses holds, that is to say,

$$P Z=1|(X_1, X_2)=h(\beta_0+\beta_1 X_1+\beta_2 X_2)$$

(1)

where $\beta = (\beta_0, \beta_1, \beta_2)'$ is the 3-vector of parameters and h is a known function that is bounded in the unit interval $(0,1)$. Although a wide choice of link function is available, the most commonly used in practical are the three models as follows:

1. the logistic regression model:

$$h(x)=\frac{\exp(x)}{1+\exp(x)}$$

2. the probit model:

$$h(x)=\Phi(x)$$

3. the complementary log-log model:

$$h(x)=1 - \exp(-\exp(x))$$

Let $\alpha_1 X_1 + \alpha_2 X_2$ be any linear combination, and their sensitivity and specificity are respectively *Sn* and *Sp*. First, we will show that the selected linear combination by the link function of the model, $\beta_1 X_1 + \beta_2 X_2$, dominates all the other possible linear combinations in the sense that it provides the highest sensitivity uniformly over the entire range of specificity and thus, it is the best linear combination that results in the largest area under the ROC curve (AUC) among all linear combinations. Then we give the mathematical formula to compute the corresponding largest AUC.

The following useful lemma was originally due to Chebyshev (see Hardy, 1959)

Lemma 1

Let U be a one-dimensional non-degenerate random variable. If g is bounded, positive, and strictly decreasing, then

$$E[U \cdot g(U)] < E(U) \cdot E(g(U))$$

holds provide the expectation of U exists.

Theorem 1

Suppose the generalized linear model (1) for binary responses holds and the function h is continuous and strictly monotone. If (X_1, X_2) is two-dimensional continuous variable with continuous probability density and its expectation exists, then for any given specificity Sp, the coefficients for the best linear combination that provides the highest sensitivity uniformly is $(\alpha_1, \alpha_2) \propto (\beta_1, \beta_2)$.

Theorem 1 may be easily extended to multiple-markers situation, which is presented in the following theorem without proof.

Theorem 2

Suppose the generalized linear model for binary responses holds, i.e.,

$$P Z=1|(X_1, \cdots, X_k)=h(\beta_0+\beta_1 X_1+ \cdots +\beta_k X_k) \tag{2}$$

where $\beta = (\beta_0, \beta_1, \cdots, \beta_k)'$ is the $(k + 1)$-vector of parameters with $k \geq 2$. Let $\alpha_1 X_1 + \alpha_k X_k$ be any linear combination. If h is a known function that is also continuous, strictly monotone and bounded in the unit interval $(0,1)$, $(X_1, \cdots, X_k)$ is k-dimensional continuous variable with continuous probability density and its expectation exists, then for any given specificity, the coefficients for the optimal linear combination that provides the highest sensitivity uniformly is $(\alpha_1, \cdots, \alpha_k) \propto (\beta_1, \cdots, \beta_k)$.

Theorem 3

Suppose the conditions in Theorem 2 hold. Then the best linear combination under the ROC criterion is also $(\alpha_1, \cdots, \alpha_k) \propto (\beta_1, \cdots, \beta_k)$. And if h is strictly increasing, then the largest AUC is

$$LAUC=\frac{1}{P(Z=1)P(Z=0)} \int G(x)h(x)dG(x) - \frac{P(Z=1)}{2P(Z=0)}$$

where $P(Z = 1) = \int h(x)dG(x)$, $P(Z = 0) = 1 - P(Z = 1)$ and $G(x)$ is the distribution function of $X = \beta_0 + \beta_1 X_1 + \cdots + \beta_k X_k$, which can be derived from the joint distribution of $(X_1, \cdots, X_k)$.

For rare diseases such as Osteoporotic Fractures discussed in the example section, we have the following simple formula to calculate the largest area under the ROC curve of the best linear combination, which can be directly derived from simple computation.

Corollary 1

If the prevalence of the disease, $P(Z = 1)$, is small enough, for example less than 5%, then the largest AUC may be approximated as follows:

$$LAUC \approx \frac{\int G(x)h(x)dG}{P(Z=1)}$$

whose error is about $P(Z = 1)/2$, say less than 2.5% if $P(Z = 1) \leq 5\%$.

3. Estimation of the Best Linear Combination and the largest AUC

In practical application, we need first to fit the generalized linear model from data. Suppose that for the study subject i, the response Z_i has the Bernoulli distribution

$$P(Z_i = z_i) = \theta_i^{z_i}(1 - \theta_i)^{1-z_i}, \quad z_i = 0, 1, \quad i = 1, \cdots, n$$

and θ_i is determined by the link function $\theta_i = h(\beta_0 + \beta_1 x_{i1} + \cdots + \beta_k x_{ik})$, where $\beta = (\beta_0, \beta_1, \cdots, \beta_k)'$ is the $(k + 1)$-vector of parameters with $k \geq 2$, h is a known continuous and strictly monotone function bounded in the unit interval $(0,1)$, and $(x_{i1}, \cdots, x_{ik})$ are the values of the two diagnostic predictors included in the model for the ith subject $(i = 1, \cdots, n)$.

It is common to make use of the likelihood function as follows to estimate the parameters $\beta = (\beta_0, \beta_1, \cdots, \beta_k)'$:

$$L(\beta; z) = \prod_{i=1}^{n} [h(\beta_0 + \beta_1 x_{i1} + \cdots + \beta_k x_{ik})]^{z_i} [1 - h(\beta_0 + \beta_1 x_{i1} + \cdots + \beta_k x_{ik})]^{1-z_i}$$

In fact, we can obtain the maximum likelihood estimate $\hat{\beta}$ of β by maximizing $L(\beta; z)$. An iterative algorithm, such as the popular Newton-Raphson method implemented in standard software from SAS and S-Plus, may be used to compute $\hat{\beta}$ numerically (Cox, 1989).

As for estimation of the largest AUC, we need further know the joint distribution of the diagnostic predictors. There are two approaches for this problem. One belongs to conventional parametric methods, among which the normal approximation is the most popular. Another is the standard nonparametric technique, which use the empirical distribution, based on the sample data, to directly estimate the population distribution. After that, we could derive the estimation of the distribution function of $\beta_0 + \beta_1 X_1 + \cdots + \beta_k X_k$, and the corresponding area under the ROC curve based on Theorem 3 in this paper.

4. Application

In this section, we apply our method to the data obtained from the Study of Osteoporotic Fractures (SOF) for illustration and compare it to the approach of Su and Liu (1993). From 1986 to 1988, SOF recruited 9,704 white women aged 65 years or older from four areas of the United States. At baseline, bone mineral density (BMD) was measured at the calcaneus, distal radius and proximal radius using single photon absorptiometry (SPA). At the second visit (1988–89), surviving participants had BMD measurements of the posterior-anterior (PA) spine (L1–L4) and proximal femur (neck, trochanter, total hip regions of interest) using dual x-ray absorptiometry (DXA). Fractures of the hip were recorded for each subject at each visit. More

details about the study design and the data have been published previously (Cummings, 1995).

We included 7,127 women from the study. All of these women had forearm, calcaneal, hip and spine bone mineral density (BMD) measurements. In addition to the BMD measurements, many other previously identified predictive variables at baseline were also investigated. Furthermore, they all had known 5-year hip fracture status: either they were followed for 5 years after visit 2 without hip fracture or they had hip fracture within five years after visit 2. Women who were lost to follow-up within 5 years without known hip fracture were excluded.

Although all these 43 variables, including patient demographic data, clinical BMD measured by DXA and SXA, medical history, X-ray for prevalent fracture and vertebral heights, functional status, vision test results, and nutrition, etc, have been identified as significant predictors of hip fracture risk and they may reflect different aspects of osteoporosis and aging and may help in understanding the etiology of osteoporosis and hip fracture, only a few of them are necessary to identify subjects with elevated fracture risks. Standard approaches to the analysis of binary data such as logistic regression and probit model show that a linear combination of age, the femoral neck BMD and loss of height could best predict the hip fracture. Our previous study also suggested that these 3 variables could be used to build a non-inferior classification rule to the optimum recursive partitioning rule (Jin, 2004). So here we consider these three predictors and want to find out the best linear combination under the ROC criterion.

Assume the generalized linear model (2) holds for the SOF data, where $Z_1 = 1$ stands for hip fracture, X_1, X_2 and X_3 denote age, femoral neck BMD and loss of height respectively. If we further assume $h(x) = \dfrac{\exp(x)}{1+\exp(x)}$, the standard software S-plus or SAS provides the following fitted logistic regression model:

$$P Z=1|(X_1, X_2, X_3)=1 - \frac{1}{1+\exp(-3.89+0.075X_1 - 8.90X_2+0.100X_3)}$$

It follows from our theorems 2 and 3 that $X^{(l)} = 0.075X_1 - 8.90X_2 + 0.100X_3$ is the optimal linear combination in the sense that it provides the highest sensitivity uniformly over the entire range of specificity and also under the ROC criterion. Hence, the estimated best coefficients are proportional to $(\alpha_1^l, \alpha_2^l, \alpha_3^l) \propto (1, -119, 1.33)$.

We need to estimate the joint distribution of (X_1, X_2, X_3) in order to get the corresponding largest area under the ROC curve. If we further assume that $(X_1, X_2, X_3)^\tau \sim N \mu,\Sigma$, where τ means the transpose of a vector, the standard normal estimation gives $\hat{\mu} = (71.06, 0.65, 3.36)^\tau$ and

$$\widehat{\Sigma}=\begin{pmatrix} 23.88 & -0.13 & 3.77 \\ -0.13 & 0.012 & -0.05 \\ 3.77 & -0.05 & 8.05 \end{pmatrix},$$

which leads to $-3.89 + 0.075X_1 - 8.90X_2 + 0.10X_3 \sim N -4.05, 1.22$. It follows from Theorem 3 that the largest AUC under the logistic regression model is **0.798.** The 250 times bootstrap standard error is 0.049.

We may obtain similar result under the probit model, that is to say, $h(x) = \Phi(x)$. Here, the fitted model of hip fracture risk is

$$P Z=1|(X_1, X_2, X_3)=\Phi - 2.16+0.036X_1 - 4.09X_2+0.048X_3,$$

which suggests $X^{(p)} = 0.036X_1 - 4.09X_2 + 0.048X_3$ be the best linear combination under the probit model. Hence, the estimated best coefficients are proportional to $(\alpha_1^p, \alpha_2^p, \alpha_3^p) \propto (1, -114, 1.34)$, and the corresponding largest AUC be **0.805** (with standard error 0.050). By the way, if we use the approximate formula in Corollary 1, these two largest AUCs are estimated to be **0.788** and **0.795**, just 1% lower than the true values.

We also can consider applying Su and Liu's method to the hip fracture data. Under their approach, the two conditional distributions are estimated as

$$\begin{pmatrix} X_1 \\ X_2 \\ X_3 \end{pmatrix} | Z=1 \sim N\left(\begin{pmatrix} 74.97 \\ 0.56 \\ 5.28 \end{pmatrix}, \begin{pmatrix} 36.83 & -0.079 & 8.28 \\ -0.079 & 0.007 & -0.069 \\ 8.28 & -0.069 & 13.23 \end{pmatrix} \right)$$

and

$$\begin{pmatrix} X_1 \\ X_2 \\ X_3 \end{pmatrix} | Z=0 \sim N\left(\begin{pmatrix} 70.93 \\ 0.66 \\ 3.30 \end{pmatrix}, \begin{pmatrix} 22.92 & -0.12 & 3.36 \\ -0.12 & 0.012 & -0.046 \\ 8.28 & -0.046 & 7.76 \end{pmatrix} \right)$$

The general method in Section 3 of their paper may give the best coefficients as follows, $(\alpha_1^s, \alpha_2^s, \alpha_3^s) \propto (1, -103, 1.05)$ with the estimated largest AUC **0.760**.

For comparison of fit goodness of the three models, we directly estimate the ROC curves corresponding to the three coefficients $(\alpha_1^l, \alpha_2^l, \alpha_3^l)$, $(\alpha_1^p, \alpha_2^p, \alpha_3^p)$ and $(\alpha_1^s, \alpha_2^s, \alpha_3^s)$ empirically, i.e. based on the SOF data itself, and then calculate the areas under the curves. It's not surprising that the real AUCs are all 0.805 because of little difference between the three optimal linear combinations resulted by different models. Therefore, the logistic regression and probit model fit the SOF data very well while Su and Liu's method underestimates the area under the ROC curve of the best linear combination.

5. Discussion and Conclusion

In medical applications, it is important to use all the good diagnostic predictors of a disease simultaneously to establish a new predictor with higher statistical utility. The linear combinations of multiple predictors are of particular interest to us. In this paper, we consider a common issue where the generalized linear models are utilized to fit the data with binary outcomes. Now that the linear combination identified by the link function is proved optimal under the ROC criterion, we just need to estimate the best linear combination using the standard procedure once the generalized linear models pass the statistical test. Therefore it is easy for clinicians to get the optimal linear combination of multiple diagnostic predictors under the condition of generalized linear models.

One referee of Su and Liu's paper pointed out, "Using logistic regression to identify tests that best predict presence or absence of disease is also common". Although the logistic regression is usually less efficient than the normal discriminant analysis with the normal assumption holding (Efron, 1975; Ruiz-Velasco, 1991), the generalized linear models for binary data will be more robust than the latter, because choice and estimation of the best linear combination need no assumption of the joint distribution of the multiple predictors. Su and Liu admitted this feature, exposed clearly by Cox and Snell (1989) that "once a vector of explanatory variables is given, then the probability that this individual belongs to one of the two groups is determined." Our SOF data could be a real example for illustration of that point.

There may be two directions for extension of this paper. On the one hand, we can extend our work from linear to non-linear models, which may explore a new concept--the ROC region to assess statistical utility of a diagnostic predictor instead of the ROC curve. On the other hand, we may extend our interest from binary data to more complex data with more than two outcomes, which was discussed by Yang and Carlin (2000) who used ROC surface approach. All these kinds of research are currently under investigation.

Acknowledgments

References

Cox, DR.; Snell, EJ. The Analysis of Binary Data. London: Chapman and Hall; 1989. p. 132

Cummings SR, Nevitt MC, Browner WS, et al. Risk factors for hip fracture in white women. Study of osteoporosis research group [see comments]. The New England Journal of Medicine 1995;332:767–773. [PubMed: 7862179]

DeLong ER, DeLong DM, Clark-Pearson DL. Comparing the areas under two or more correlated receiver operating characteristic curves: a non-parametric approach. Biometrics 1988;44:837–845. [PubMed: 3203132]

Efron B. The Efficiency of logistic Regression Compared to Normal Discriminant Analysis. Journal of the American Statistical Association 1975;70:892–898.

Hans D, Srivastav SK, Singal C, et al. Does combining the results from multiple bone sites measured by a new quantitative ultrasound device improve discrimination of hip fracture? Journal of Bone and Mineral Research 1999;14:644–651. [PubMed: 10234587]

Hardy, GH.; Littlewood, JE.; Polya, G. Inequalities. London and New York: Cambridge University Press; 1959. p. 43

Jin H, Lu Y, Stone KL, et al. Classification algorithms for hip fracture prediction based on recursive partitioning methods. Medical Decision Making 2004;24(4):386–397. [PubMed: 15271277]

Pepe MS. A regression modelling framework for receiver operating characteristic curves in medical diagnostic testing. Biometrika 1997;85:595–608.

Richards RJ, Hammitt JK, Tsevat J. Finding the optimal multiple-test strategy using a method analogous to logistic regression analysis. Medical Decision Making 1995;16:367–375. [PubMed: 8912298]

Ruiz-Velasco S. Asymptotic Efficiency of Logistic Regression Relative to Linear Discriminant Analysis. Biometrika 1991;78:235–243.

Su J, Liu J. Linear combinations of multiple diagnostic markers. Journal of the American Statistical Association 1993;88:1350–1355.

Yang H, Carlin D. ROC surface: A generalization of ROC curve analysis. Journal of Biopharmaceutical Statistics 2000;10:183–96. [PubMed: 10803724]

APPENDIX: PROOFS

Proof of Theorem 1

Without loss of generality, let $\alpha_2 = \beta_2$. For simplicity, let $\alpha_1 = \alpha$.

Suppose that $\beta_2 > 0$ and h is strictly increasing. Let $f(x_1, x_2)$ be the distribution density function of the two-dimensional continuous variable (X_1, X_2), and $\bar{h}(x) = 1 - h(x)$, then the specificity of the linear combination, $\alpha X_1 + \beta_2 X_2$, can be expressed as

$$Sp(c) = P(\alpha X_1 + \beta_2 X_2 \le c | Z=0) = \frac{1}{P(Z=0)} \iint_{\alpha x_1 + \beta_2 x_2 \le c} P(Z=0|(x_1, x_2)) f(x_1, x_2) dx_1 dx_2$$

$$= \frac{1}{P(Z=0)} \int_{-\infty}^{\infty} \left[\int_{-\infty}^{\frac{c-\alpha x_1}{\beta_2}} \bar{h}(\beta_0 + \beta_1 x_1 + \beta_2 x_2) \cdot f(x_1, x_2) dx_2 \right] dx_1$$

and, therefore,

$$\int_{-\infty}^{\infty} \left[\int_{-\infty}^{\frac{c-\alpha x_1}{\beta_2}} \bar{h}(\beta_0 + \beta_1 x_1 + \beta_2 x_2) \cdot f(x_1, x_2) dx_2 \right] dx_1 = S_p \cdot P(Z=0) \tag{A1}$$

For any given specificity Sp, the equation (A1) defines c as a differentiable function of α, denoted as $c(\alpha)$ below. Furthermore, it follows from differentiating both sides of equation (A1) with respective to α that

$$\int_{-\infty}^{\infty} \bar{h}(\beta_0 + \beta_1 x + c(\alpha) - \alpha x) \cdot f(x, \frac{c(\alpha) - \alpha x}{\beta_2}) \cdot \frac{c'(\alpha) - x}{\beta_2} dx = 0 \tag{A2}$$

Evaluating (A2) at the specific choice $\alpha = \beta_1$, and noting that the nonzero function $\bar{h}(\beta_0 + \beta_1 x + c(\beta_1) - \beta_1 x)$ does not depend on x and so can be taken out of the above integral, leads to the equality:

$$\int_{-\infty}^{\infty} \frac{c'(\beta_1) - x}{\beta_2} f(x, \frac{c(\beta_1) - \beta_1 x}{\beta_2}) dx = 0 \tag{A3}$$

where $c'(\alpha)$ denotes the derivative of $c(\alpha)$ with respect to α.

As for the sensitivity of X, it can be rewritten as

$$S_n(c(\alpha))=P(\alpha X_1+\beta_2 X_2>c(\alpha)|Z=1)=1-\frac{1}{P(Z=1)}\iint\limits_{\alpha x_1+\beta_2 x_2\leq c(\alpha)}P(Z=1|(x_1,x_2))f(x_1,x_2)dx_1dx_2$$

$$=1-\frac{1}{P(Z=1)}\int\limits_{-\infty}^{\infty}dx_1\left[\int\limits_{-\infty}^{\frac{c(\alpha)-\alpha x_1}{\beta_2}}h(\beta_0+\beta_1 x_1+\beta_2 x_2)f(x_1,x_2)dx_2\right]$$

$$=1-\frac{1}{P(Z=1)}\left(\int\limits_{-\infty}^{\infty}\left[\int\limits_{-\infty}^{\frac{c(\alpha)-\alpha x_1}{\beta_2}}f(x_1,x_2)dx_2\right]dx_1-S_p\cdot P(Z=0)\right)$$

by utilizing (A1). So, we just need to prove that $S(\alpha)=\int\limits_{-\infty}^{\infty}\left[\int\limits_{-\infty}^{\frac{c(\alpha)-\alpha x_1}{\beta_2}}f(x_1,x_2)dx_2\right]dx_1$ can reach the absolute minimum at the point $\alpha=\beta_1$ under the restriction of (A1) or (A2).

Because

$$S'(\alpha)=\int\limits_{-\infty}^{\infty}\frac{c'(\alpha)-x}{\beta_2}f(x,\frac{c(\alpha)-\alpha x}{\beta_2})dx, \tag{A4}$$

it follows from (A3) and (A4) that

$$S'(\beta_1)=\int\limits_{-\infty}^{\infty}\frac{c'(\beta_1)-x}{\beta_2}f(x,\frac{c(\beta_1)-\beta_1 x}{\beta_2})dx=0 \tag{A5}$$

On the other hand, it's easy to see that, for any fixed α, $\frac{1}{\beta_2}\int\limits_{-\infty}^{\infty}f(x,\frac{c(\alpha)-\alpha x}{\beta_2})dx$ is the value of the density function of the linear combination $\alpha X_1+\beta_2 X_2$ at point $c(\alpha)$. Let

$$p(x)=\frac{f(x,\frac{c(\alpha)-\alpha x}{\beta_2})}{\int\limits_{-\infty}^{\infty}f(x,\frac{c(\alpha)-\alpha x}{\beta_2})dx}$$

be the density function of a random variable U, and $g(x)=\bar{h}(\beta_0+\beta_1 x+c(\alpha)-\alpha x)$. So, when $\alpha<\beta_1$, g is bounded, positive, and strictly decreasing. From Lemma 1, we have

$$E[U\cdot g(U)]<E(U)\cdot E(g(U))$$

which leads to the following inequality:

$$\frac{\int\limits_{-\infty}^{\infty}xg(x)f(x,\frac{c(\alpha)-\alpha x}{\beta_2})dx}{\int\limits_{-\infty}^{\infty}g(x)f(x,\frac{c(\alpha)-\alpha x}{\beta_2})dx}<\frac{\int\limits_{-\infty}^{\infty}xf(x,\frac{c(\alpha)-\alpha x}{\beta_2})dx}{\int\limits_{-\infty}^{\infty}f(x,\frac{c(\alpha)-\alpha x}{\beta_2})dx} \tag{A6}$$

From (A2), we see that

$$\frac{\int\limits_{-\infty}^{\infty} xg(x)f(x, \frac{c(\alpha)-\alpha x}{\beta_2})dx}{\int\limits_{-\infty}^{\infty} g(x)f(x, \frac{c(\alpha)-\alpha x}{\beta_2})dx} = c'(\alpha) \tag{A7}$$

Combining (A4),(A6) and (A7), we obtain that $S'(\alpha) < 0$ when $\alpha < \beta_1$.

Similarly, we can prove that $S'(\alpha) > 0$ when $\alpha > \beta_1$. Therefore, combining with (A5), we finish the proof that $S(\alpha)$ has the absolute minimum, hence $Sn(c(\alpha))$ has the absolute maximum, at $\alpha = \beta_1$ when $\beta_2 > 0$ and h is strictly increasing.

In a similar way, it's easy to prove that the coefficients for the best linear combination that provides the highest sensitivity uniformly is $(\alpha_1, \alpha_2) \propto (\beta_1, \beta_2)$ when $\beta_2 < 0$ or h is strictly decreasing. Thus the proof of the theorem is complete.

Proof of Theorem 3

Once again, we let $\bar{h}(x) = 1 - h(x)$. It follows from (2) that

$$P\ Z=1|X=h(X)$$

So we have $P(Z = 1) = \int h(x)dG(x)$, and the specificity of the best linear combination can be expressed as

$$S\ p(c)=P\ X \leq c|Z=0=\frac{1}{P(Z=0)}\int\limits_{x\leq c} P\ Z=0|x)\ dG(x)=\frac{1}{P(Z=0)}\int\limits_{-\infty}^{c} \bar{h}(x)dG(x)$$

Then $dSp(c)=\dfrac{1}{P(Z=0)}\bar{h}(c)dG(c)$, and

$$\int \frac{1}{P(Z=1)}G(c)dSp(c)=\frac{1}{P(Z=1)P(Z=0)}\int G(c)\bar{h}(c)dG(c).$$

On the other hand, its sensitivity is

$$S\ n(c)=P\ X>c|Z=1=1 - \frac{1}{P(Z=1)}\int\limits_{x\leq c} P\ Z=1|x)\ dG(x)=1 - \frac{1}{P(Z=1)}\int\limits_{-\infty}^{c} h(x)dG(x)$$
$$=1+\frac{P(Z=0)}{P(Z=1)}S\ p(c) - \frac{1}{P(Z=1)}G(c)$$

Then the largest area under the ROC curve is given by

$$\begin{aligned}
AUC = \int S\,n(c)\,dSp(c) &= \int \left(1 + \frac{P(Z=0)}{P(Z=1)} S\,p(c)\right) dSp(c) - \int \frac{1}{P(Z=1)} G(c)\,dSp(c) \\
&= \left(1 + \frac{P(Z=0)}{2P(Z=1)}\right) - \frac{1}{P(Z=1)P(Z=0)} \int G(c)\overline{h}(c)\,dG(c) \\
&= \left(1 + \frac{P(Z=0)}{2P(Z=1)}\right) - \frac{1}{P(Z=1)P(Z=0)} \int G(c)(1 - h(c))\,dG(c) \\
&= \frac{1}{P(Z=1)P(Z=0)} \int G(c)h(c)\,dG(c) + \left(1 + \frac{P(Z=0)}{2P(Z=1)} - \frac{1}{2P(Z=1)P(Z=0)}\right) \\
&= \frac{1}{P(Z=1)P(Z=0)} \int G(x)h(x)\,dG(x) - \frac{P(Z=1)}{2P(Z=0)}.
\end{aligned}$$