Um Modelo de Classificação com Solução Aproximada por Técnica de Otimização

Henry Rossi de Almeida

ITA – Instituto Tecnológico da Aeronáutica
São José dos Campos - Brasil
hralmeid@cacapava.com.br

Abstract

In this work we consider the problem about classification of objects (products, individuals, companies) in two or more groups.

The traditional approach is based mainly on models like Logit or Discriminant Analysis. It results in this way, either a cutting score or a measure of distance from the center of each group, defined with the simultaneous use of all the sample.

In this model we consider that the classification variables are limited, inferiorly or superiorly. Through this hypothesis, it results a border for each group, that is the limit of the group.

We intend to determine the border for the group and, also, a value of probabilistic nature for the measure (score) of each unit, relatively to the boundary of the group that it belongs. After that, we do the classification of a new unit, in the group where it presents the greater measure.

Resumo

Neste trabalho consideramos o problema de classificação de objetos (indivíduos, empresas, produtos) em dois ou mais grupos.

A abordagem tradicional se fundamenta principalmente em modelos como Logit ou Análise Discriminante. Resulta deste modo, uma medida estatística – ponto de corte (escore) ou uma medida de distância padronizada ao centro do agrupamento ou uma hipersuperfície de separação – definido com a utilização simultânea de toda a amostra.

No modelo proposto consideramos que as variáveis de classificação são limitadas, inferior ou superiormente. Desta hipótese resulta uma fronteira, limite da região domínio de cada grupo em observação.

Objetivamos determinar as fronteiras dos grupos analisados e, também, um valor de natureza probabilística para a medida de cada unidade observada, que especifique quanto a mesma está inserida em seu respectivo grupo, relativamente à fronteira deste. Em seguida efetuamos o teste para classificar nova unidade, no grupo em que a mesma apresente a maior medida de inserção.

Keywords: Discriminant Analysis**,** Score Classification**,** Linear Programming**,** Data Envelopment Analysis (DEA)

Title: An approximate solution Classification Model through Optimization Techniques

1 Introdução

Neste trabalho abordamos o problema de classificação de unidades (DMU *Decision Making Unit*) em dois ou mais grupos (populações) distintos. Consideramos ainda que as variáveis discriminatórias são contínuas e limitadas, superior ou inferiormente, definindo fronteiras para cada particular grupo.

A motivação do trabalho é apresentar uma proposta complementar ao procedimento adotado em modelos tradicionais, que obtém a medida estatística de classificação utilizando todos os elementos da amostra, indistintamente. Por exemplo, ao procurarmos definir o escore para classificação de um grupo de empresas com característica de solvência e outro de insolvência, através de Análise Discriminante, as empresas com excelente e com péssima situação financeira, onde não residem dúvidas quanto à real classificação, vão influenciar sensivelmente na decisão de classificação de nova unidade de teste. Isto porque afetaram o escore, por terem participado na determinação deste, com a mesma importância dada às empresas em situação financeira duvidosa que, de fato, são aquelas que apresentam dificuldades de classificação. Apesar de todas participarem, como integrantes da amostra, na determinação do escore, apresentaremos um modelo diferenciando a participação de cada DMU.

Com base nas considerações iniciais, não deveremos encontrar nenhuma unidade de determinado grupo, fora da fronteira deste, resultando uma região vazia, seja pelo baixo valor de uma determinada variável com característica favorável ao grupo (que denominaremos variável favorável), seja pelo valor muito elevado de uma determinada variável desfavorável.

Consideramos ainda que uma mesma variável pode ter simultaneamente característica favorável a uma população e desfavorável a outra população. Assim, ao analisarmos uma colônia de microorganismos anaeróbicos, a variável "teor de oxigênio" deve ser considerada uma variável desfavorável a esse grupo, tal que acima de um teor reduzido, haveria impossibilidade de a colônia sobreviver. Já, na análise de uma colônia de microorganismos aeróbicos, "teor de oxigênio" deve ser considerado uma variável favorável para identificação deste grupo e, abaixo de certo nível, esta colônia não sobreviveria.

Os objetivos do presente trabalho, atendendo à nossa motivação, são:

- Determinar as fronteiras que delimitam cada população, a partir da observação dos valores das variáveis favoráveis e desfavoráveis, das amostras de elementos de cada população.
- Determinar uma medida da probabilidade de inserção de cada elemento da amostra em relação à sua fronteira. Deste modo desenvolveremos uma escala de medidas partindo da fronteira para dentro, não aquela usual do centro do agrupamento para fora.
- Classificar novas DMUs na população que apresentar maior medida de inserção em cada fronteira construída.
- Aprimorar as fonteiras, sempre que novas informações assim possibilitarem.

Neste trabalho consideramos que a seqüência de modelagem para classificação deve manter a proposta de [Duda et al, 2001], qual seja:

- Pré-processamento – Para levantamento das características de cada população.
- Extração – Definição das principais características e seus valores.
- Classificação – Tomada de decisão com base nos valores das características.
- Pós-Processamento – Ajuste definitivo, considerando qualidade e custos do sistema.

Para manter esta seqüência vamos necessitar de uma amostra original para calibração do modelo e uma segunda amostra para efetuar classificação e principalmente o pós-processamento, objetivando o aprimoramento das fronteiras originalmente obtidas. Porém estamos considerando que a utilização de um modelo de classificação não se encerra após uso inicial, de modo que o pós-processamento de modelagem deve ser dinâmico e, quando novas informações se agregarem, os parâmetros definidores das fronteiras podem e devem ser atualizados.

Dada a complexidade de cálculo, consideramos fundamental apresentar um procedimento para obtenção de solução aproximada. Nos apoiaremos nas técnicas de otimização, utilizando assim algoritmos conhecidos**.** Faremos uso de modelos de fronteira, cuja estrutura matemática é semelhantes à técnica de Análise de Envoltória de Dados DEA (do inglês *Data Envelopment Analysis)*. No entanto adequaremos os resultados para as condições probabilísticas de nosso problema.

O trabalho está organizado da seguinte maneira:

A seção 2 conceitua o que consideramos como variável favorável e variável desfavorável.

A seção 3 desenvolve o modelo a partir da medida de inserção, definida por curvas de mesma probabilidade (ou curvas de indiferença), definidoras de uma escala de medida.

A seção 4 estabelece os procedimentos para obtermos uma solução aproximada e sua utilização com base nos dados da amostra de calibração do modelo.

A seção 5 estabelece a metodologia de classificação, para a solução aproximada.

A seção 6 expõe considerações sobre *outliers* e aprimoramento das fronteiras.

A seção 7 é dedicada a um estudo de caso, com o desenvolvimento do modelo e comparação com outro modelo da literatura, evidenciando situações onde ele não se aplica.

A seção 8 apresenta as conclusões sobre a utilização e do modelo.

2 Definição das Variáveis

Nesta seção apresentamos o conceito adotado para as variáveis atuantes no processo de analisar uma DMU, conforme foi exposto na Seção 1. Assim:

- Se uma variável é favorável à ocorrência de elementos da população, então quanto maior for seu valor numérico para uma DMU, melhor é a condição desta no grupo.
- Se uma variável é desfavorável para a população, então quanto menor for seu valor, melhor é a condição da DMU no grupo.

A Figura 1 explicita uma população influenciada por duas variáveis, uma de característica favorável e outra de característica desfavorável. A situação se traduz em uma região H, definindo o domínio das variáveis VD e VF para esta população, bem como uma região I, onde as condições são extremamente desfavoráveis à ocorrência de uma entidade desta população.

VF = variável favorável (X)
VD = variável desfavorável (Y)
H = Região com ocorrência de DMUs
I = Região ausente de DMUs
y_S = Limite superior da variável Y
x_I = Limite inferior da variável X

Uma variável pode apresentar:
- Característica favorável (X) a uma população e
- Característica desfavorável (Y) a outra população

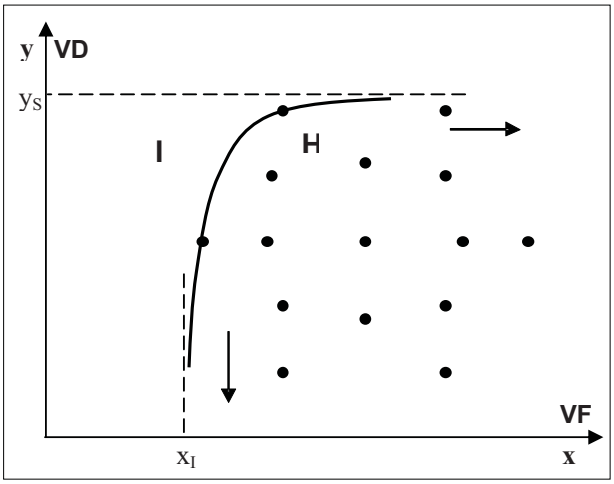

Figura 1 – Região H, domínio das variáveis

2.1 Ajuste das variáveis

Para trabalharmos somente com variáveis de características equivalentes, substituiremos cada valor x_i da variável favorável X_i por $x_i - x_{Ii}$ e cada valor y_r da variável desfavorável Y_r por $y_{Sr} - y_r$, considerando m variáveis favoráveis e s variáveis desfavoráveis.

Resultarão novas variáveis, todas apresentando característica favorável (X) e serão definidas por $X_i = \{x_i \in R \,/\, x_i \geq 0 \,,\, i = 1,...,m+s\}$, delimitando uma região, conforme Figura 2, para o caso de duas variáveis, X_1 e X_2.

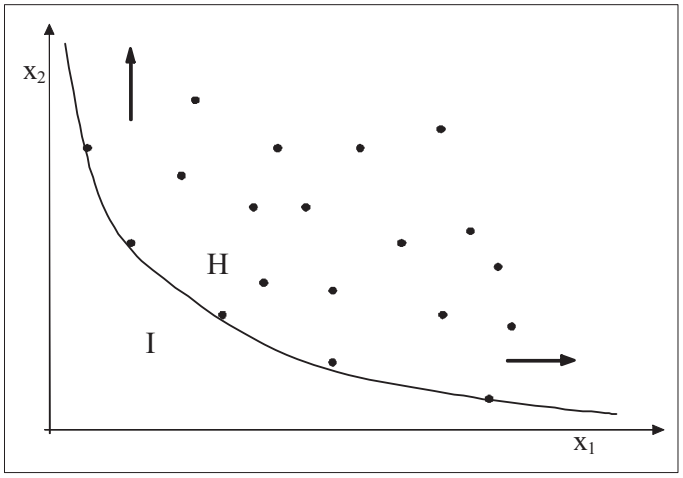

Figura 2 – Região H, domínio das variáveis ajustadas

3 Desenvolvimento de um Modelo Probabilístico de Classificação

Esta seção conceitua o grau de inserção que posiciona uma DMU, medido sobre a função de distribuição de probabilidades (fdp) desta população, relativamente à fronteira. A Figura 3, representando o domínio da fdp conjunta $f(x_1,x_2)$ das variáveis ajustadas X_1 e X_2, retrata esta situação.

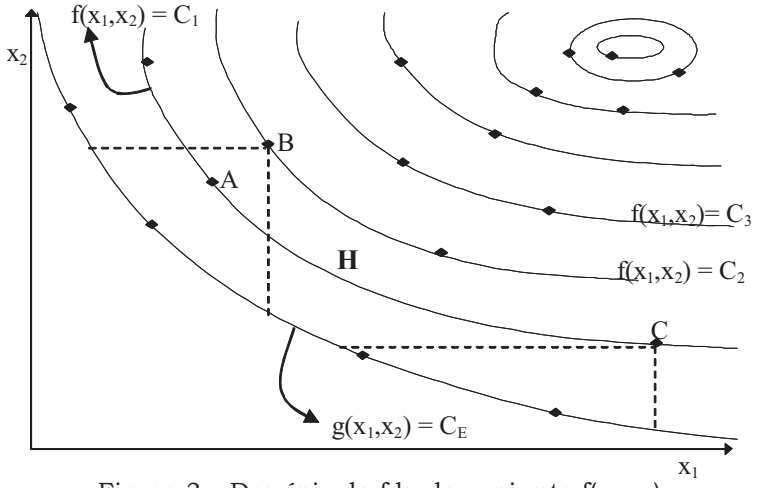

Figura 3 – Domínio da fdp da conjunta $f(x_1,x_2)$

A envoltória definida pela função $g(x_1,x_2) = C_E$ representa a fronteira assintótica aos eixos x_1 e x_2, estabelecendo o domínio da fdp.

A DMU, representada pelo ponto B, está mais inserida na região H, do que A, pois:

$x_{1B} > x_{1A}$ e $x_{2B} > x_{2A}$

Se $\iint_B dP$ representa $\iint f(x_1, x_2)dx_1dx_2 = P(B)$, entre a curva envoltória $f(x_1, x_2) = C_E$ e o ponto B

e $\iint_A dP$ representa $\iint f(x_1, x_2)dx_1dx_2 = P(A)$, entre a curva envoltória $f(x_1, x_2) = C_E$ e o ponto A,

Então: $x_{1B} > x_{1A}$ e $x_{2B} > x_{2A} \Rightarrow P(x_B) > P(x_A)$

Diferentemente, na comparação entre as DMUs representadas por B e C, teremos:

$x_{1B} < x_{1C}$ e $x_{2B} > x_{2C}$

Neste caso, as coordenadas não são suficientes para definir uma regra de decisão sobre medida de inserção. Deveremos formular uma nova medida, que será baseada em critérios probabilísticos, para o modelo proposto. Assim:

HIPÓTESE DA PROBABILIDADE:

- **Se $\iint_B dP = \iint_C dP \Rightarrow$ As DMUs B e C estão igualmente inseridas**
- **Se $\iint_B dP = \iint_C dP > \iint_A dP \Rightarrow$ B e C estão mais inseridas que A**

A Hipótese da Probabilidade estabelece as bases para definirmos a Regra de Decisão que estará orientando o processo de classificar nova DMU em determinadas populações candidatas, ou seja, quando ocorre interesse em definir sobre duas ou mais populações, qual aquela em que a DMU deve ser classificada.

3.1 Curvas de indiferença

Conforme exposto na seção anterior, $\iint_B dP$ entre a curva envoltória $g(x_1, x_2) = C_E$ e um particular ponto B define uma probabilidade α. Assim, pontos de mesma probabilidade descrevem curvas no plano das variáveis x_1 e x_2, conforme Figura 4.

Pelas proposições da seção anterior, estas curvas serão assintóticas aos eixos x_1 e x_2, pois a probabilidade cai a zero sobre a envoltória e tende a um ao aumentarmos indefinidamente as variáveis (que por hipótese são contínuas e ilimitadas à direita).

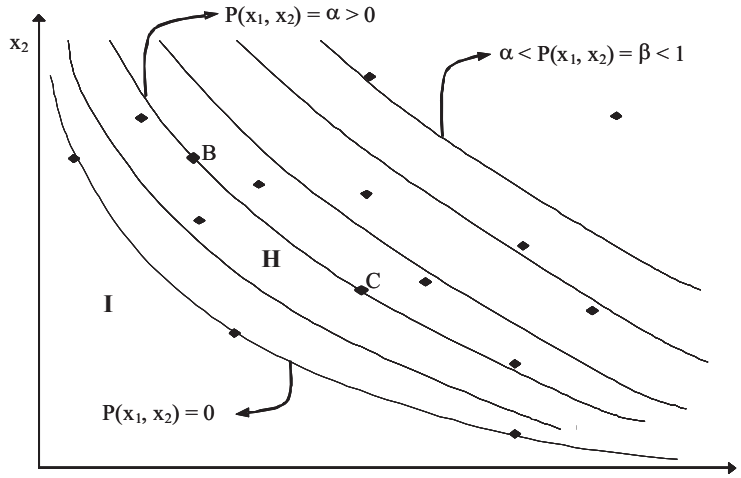

Figura 4 – Curvas de Indiferença

As DMUs representadas pelos pontos B e C na Figura 4, situados sobre a mesma curva, estão, por hipótese, igualmente inseridas dentro do domínio da fdp, pois apresentam a mesma probabilidade α. Trata-se de uma curva de indiferença no posicionamento de B e C.

Para $P(x_1, x_2) = 0$ estamos exatamente sobre a envoltória.

As curvas de indiferença (hipersuperfícies em caso n-dimensional) crescem segundo um gradiente, possibilitando serem utilizadas como medida de inserção, através do valor α.

3.2 Critério para classificação

Em acordo com as definições apresentadas na seção 3.1, podemos estabelecer um critério para classificar DMUs em diversas populações candidatas:

a) A DMU_0 está sobre a envoltória ou sobre uma curva de indiferença na região H (domínio)
   - Resulta $P(x_1, x_2, \ldots, x_{m+s}) = \alpha \geq 0$.

b) A DMU_0 está fora da região H
   - $\nexists$ $P(x_1, x_2, \ldots, x_{m+s})$.

Finalmente:
   - A DMU_0 será classificada no grupo que resultar o valor $\beta > \alpha$, para $P(x_1, x_2, \ldots, x_{m+s})$ calculado sobre a envoltória e a fdp de cada população.

4 Obtenção de uma Particular Solução Aproximada por Técnica de Otimização

A solução analítica do modelo foi apresentada na seção 3, quando as fdp são conhecidas, resultando o processo final de classificação exposto em 3.2.

No desconhecimento das fdp, podemos obter diversas soluções aproximadas, conforme as hipóteses simplificadoras adotadas. Vamos propor uma aproximação através da função que define as curvas de indiferença.

Quando representamos um ponto I de uma particular curva de indiferença $C_I(\alpha)$, em coordenadas polares, encontramos um ponto correspondente E, sobre a envoltória, de tal modo que $\overrightarrow{OI} = k\overrightarrow{OE}$, exposto na Figura 5.

A distância radial $\overline{EI}$ entre os diversos pontos da envoltória C_E e uma particular curva de indiferença $C_I(\alpha)$ é função do ângulo diretor θ e da probabilidade α desejada para esta curva, conforme projeção efetuada sobre o plano x_1x_2 da Figura 5:

$k = f(\alpha,\theta)$ $\alpha = P(x_{1I}, x_{2I}, \ldots, x_{(m+s)I})$ na generalização para m+s dimensões.

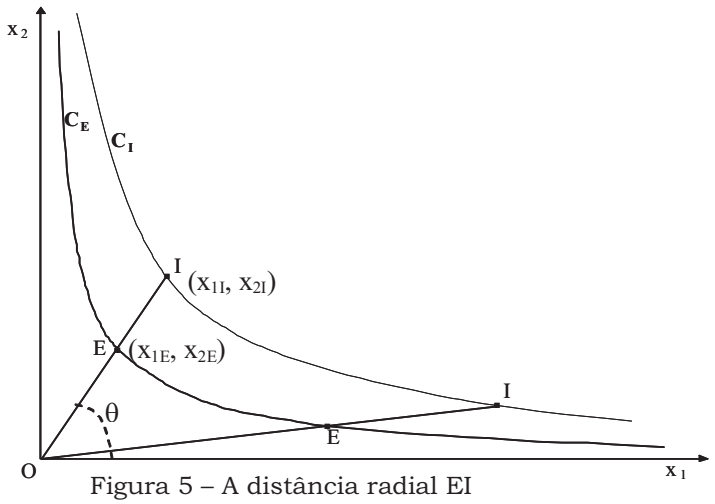

Figura 5 – A distância radial EI

Para todo $E \in$ Envoltória e $I \in C_I(\alpha)$ resulta, para uma dada probabilidade α tal que:

$$\begin{pmatrix} x_{1I} \\ x_{2I} \\ . \\ . \\ x_{mI} \end{pmatrix} = k \begin{pmatrix} x_{1E} \\ x_{2E} \\ . \\ . \\ x_{mE} \end{pmatrix} \quad \text{e, para o vetor EI:} \quad \begin{pmatrix} x_{1I} - x_{1E} \\ x_{2I} - x_{1E} \\ . \\ . \\ x_{mI} - x_{mE} \end{pmatrix} = (k-1) \begin{pmatrix} x_{1E} \\ x_{2E} \\ . \\ . \\ x_{mE} \end{pmatrix}$$

Ou: $x_{iI} - x_{iE} = (k-1)\, x_{iE}$; $i = 1, \ldots, m$ $\Rightarrow$ $\dfrac{x_{iI} - x_{iE}}{x_{iI}} = \dfrac{k-1}{k} = h(\alpha,\theta)$; $i = 1, \ldots, m$

$$\therefore \quad \frac{x_{1I} - x_{1E}}{x_{1I}} = \frac{x_{2I} - x_{2E}}{x_{2I}} = .. = \frac{x_{mI} - x_{mE}}{x_{mI}} = h(\alpha,\theta) \qquad (1)$$

A posição da curva de indiferença que explicita a probabilidade α desejada, pode ser representada pela medida h da equação (1).

Consideremos agora: $\overline{k(\alpha)} = \dfrac{\int_0^{2\pi} k(\alpha,\theta)\,d\theta}{\pi/2}$

Podemos observar que nos casos onde k for independente de θ resulta $k(\alpha,\theta) \equiv \overline{k(\alpha)}$.

Para as fdp em que $k(\alpha,\theta)$ seja contínua e apresente derivadas de qualquer ordem, podemos desenvolver $k(\alpha,\theta)$ em série de Taylor, a partir de um ponto θ_0. Vamos considerar os casos em que esta série converge, em todo o domínio $0 < \theta < \pi/2$.

Consideremos primeiramente θ_0 de modo que $k(\alpha,\theta_0) = \overline{k(\alpha)}$.

Temos assim (na particularização bidimensional) a série:

$$k(\alpha,\theta) = \overline{k(\alpha)} + \frac{\partial\, k(\alpha,\theta_0)}{\partial\,\theta}(\theta - \theta_0) + \frac{\partial^2 k(\alpha,\theta_0)}{\partial\,\theta^2}\frac{(\theta - \theta_0)^2}{2!} + \ldots.$$

Deste modo, a utilização de $\overline{k(\alpha)}$, para determinação de k(α,θ), é o primeiro passo para uma melhora, por aproximações sucessivas, da solução h(α,θ) da equação (1).

Adotaremos esta solução como definitiva, no desconhecimento de melhores informações sobre a fdp, ou seja, admitindo, por hipótese, que k(α,θ) independe de θ.

A equação (1) assume então a forma apresentada na equação (2), utilizando h(α) em vez de h(α,θ), como aproximação para as curvas de indiferença, na generalização para mais que duas dimensões:

$$\frac{x_{1I} - x_{1E}}{x_{1I}} = \frac{x_{2I} - x_{2E}}{x_{2I}} = .. = \frac{x_{mI} - x_{mE}}{x_{mI}} = h(\alpha) \qquad (2)$$

Ao utilizarmos h, para compararmos DMUs de um mesmo grupo, teremos uma escala crescente, partindo de zero, indicando quais DMUs estão mais inseridas, pois $h_2 > h_1$ equivale a $\alpha_2 > \alpha_1$ na comparação entre DMU_2 e DMU_1, mesmo desconhecendo α_2 e α_1.

Quando trabalharmos com duas ou mais populações, ao tentarmos escolher em qual delas classificarmos uma DMU, vamos necessitar uma medida padronizada. Vamos tratar este tópico na seção 4.1, considerando características de dispersão para cada população, visto que, na hipótese acima, não procuramos evidenciar a relação h = f(α), para cada particular grupo.

4.1 Generalização h_G, da escala h de medida

Pretendemos nesta seção estabelecer uma medida relativa de posicionamento das curvas de indiferença, incluindo as características de dispersão para cada população, definindo assim o procedimento para classificação de DMUs em diferentes grupos. Vamos utilizar novamente uma construção bi-dimensional, conforme Figura 6, cientes porém da possibilidade de generalização.

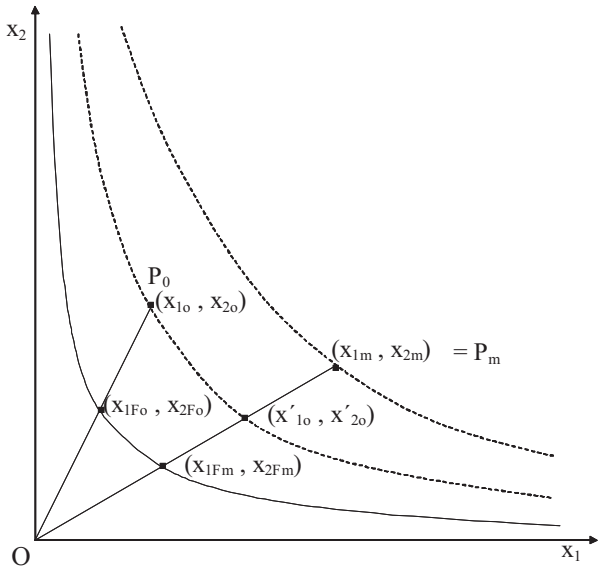

Figura 6 – Posicionamento das Curvas de Indiferença

Na Figura 6, (x_{10}, x_{20}) define a DMU_0 em análise e (x_{1m}, x_{2m}) define o centro P_m da distribuição, estimado pelo ponto médio da amostra. As três curvas representam a fronteira, a curva de indiferença onde se situa a DMU_0 e a curva de indiferença relativa ao ponto médio.

Para a DMU_0 temos, conforme equação (2):

$$h_0 = \frac{x_{10} - x_{1F_0}}{x_{10}} = \frac{x_{20} - x_{2F_0}}{x_{20}}$$

Podemos, pela mesma equação (2), obter h_0 através da DMU de apoio (x'_{10} , x'_{20}), intercessão da a reta $\overline{OP_m}$ com curva de indiferença comum às duas DMUs. Assim:

$$h_0 = \frac{x'_{10} - x_{1F_m}}{x'_{10}} = \frac{x'_{20} - x_{2F_m}}{x'_{20}}$$

$$\therefore \quad x'_{10} = \frac{x_{1F_m}}{(1-h_0)} \qquad \text{e} \qquad x'_{20} = \frac{x_{2F_m}}{(1-h_0)} \tag{3}$$

Pela equação (2), relativamente ao ponto médio P_m da distribuição:

$$h_m = \frac{x_{1m} - x_{1F_m}}{x_{1m}} = \frac{x_{2m} - x_{2F_m}}{x_{2m}}$$

$$\therefore \quad x_{1F_m} = (1-h_m)x_{1m} \quad \text{e} \quad x_{2F_m} = (1-h_m)x_{2m} \tag{4}$$

Vamos conceituar uma medida $\mathbf{h_{G_o}}$ que posicione P_0 em relação à fronteira e também em relação à curva de indiferença onde se localiza o ponto médio P_m.

Definimos assim, considerando que P_0 e (x'_{10} , x'_{20}) estão na mesma curva de indiferença:

$$h_{G_o} = \frac{x'_{10} - x_{1F_m}}{x_{1m} - x_{1F_m}} = \frac{x'_{20} - x_{2F_m}}{x_{2m} - x_{2F_m}} \tag{5}$$

Substituindo (3) e (4) em (5) resulta:

$$h_{G_o} = \frac{\dfrac{h_0}{1-h_0}}{\dfrac{h_m}{1-h_m}} \equiv \frac{h_0}{h_m}\frac{1-h_m}{1-h_0} \tag{6}$$

A equação (6) fornece a medida $\mathbf{h_G}$, que será nosso escore, no processo de comparar DMUs. Como h_{G_o} é função apenas de h_0 e h_m , então identifica a curva de indiferença de P_0, tanto em relação ao extremo como ao centro da distribuição.

Deste modo não determinamos quantitativamente o valor da probabilidade associada a P_0, conforme exposto no fim da seção anterior. Mas agora podemos comparar a posição da DMU_0 relativamente à diversos grupos candidatos para classificação, onde a mesma se encontra inserida. Isto porque estamos considerando a hipótese de os grupos apresentarem a mesma fdp, diferindo apenas na variância. A medida h_{G_o} tem o efeito de padronizar a dispersão. Resultará um valor de h_{G_o} para cada grupo, posicionando a DMU em uma diferente curva de indiferença, conforme o grupo. A DMU será classificada no grupo que apresentar maior valor de h_{G_o} .

4.2 Estimativas dos limites das variáveis

Esta seção considera que a formulação das variáveis se fará a partir de uma amostra representativa, para cada grupo.

Devemos assim estimar inicialmente os limites de cada variável por grupo. O estimador de máxima verossimilhança para o limite superior de uma variável resulta em

$\hat{Y} = \max(Y_1, Y_2, ...Y_n)$, sendo n o tamanho da amostra. Vamos adotá-lo, embora seja uma estimativa tendenciosa [Meyer P.L., 2000]. Do mesmo modo, o estimador de máxima verossimilhança para o limite inferior de uma variável resulta em $\hat{X} = \min(X_1, X_2, ...X_n)$.

Mantendo as características da seção 2.1, substituiremos cada valor X_{ij} de cada variável favorável X_i por $X_{ij} - \hat{X}_i$ e cada valor Y_{rj} da variável desfavorável Y_r por $\hat{Y}_r - Y_{rj}$. Resultarão novas variáveis, todas apresentando característica favorável (X) e serão definidas por $X_i = \{X_i \in R \ / \ X_i \geq 0 \ , \ i = 1,...,m+s\}$, delimitando a região, conforme representado na Figura 7, equivalente à Figura 2, para o caso de duas variáveis, X_1 e X_2 e um particular grupo.

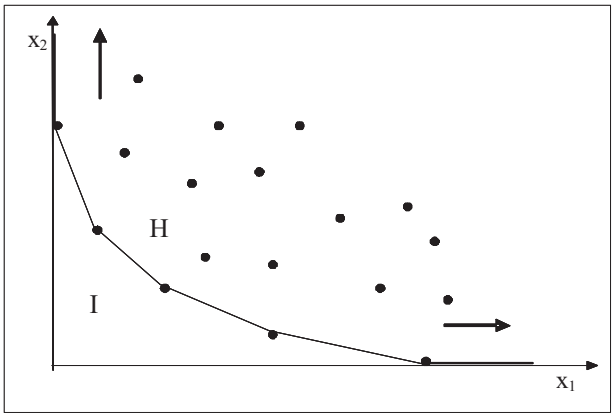

Figura 7 – Região H, domínio das variáveis ajustadas

4.3 Curvas de indiferença – Aproximação por poligonais como escala de medida

Esta seção fornece as bases para contornarmos as dificuldades de integração das fdp, necessário à determinação das curvas de indiferença, para obtenção do escore h da equação (2). Pretendemos obter diretamente o valor de h através de poligonais, cujos vértices estão sobre as curvas de indiferença.

Nos apoiaremos na Figura 8, em uma representação bidimensional, plausível de generalização para m variáveis. A envoltória, obtida a partir da amostra, está apoiada em quatro pontos (DMUs), onde identificamos o ponto P_E de coordenadas (x_{1E} , x_{2E}). Como determinar os pontos da envoltória em um ambiente m-dimensional será tópico da seção 6.1, quando estivermos expondo o apoio da técnica DEA ao modelo.

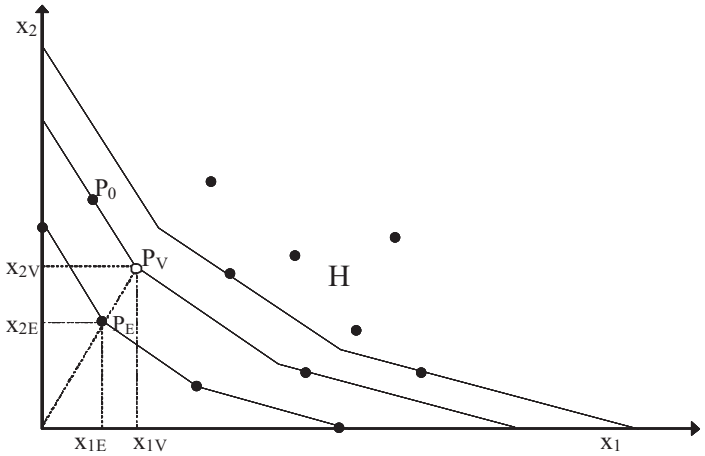

Figura 8 – Poligonais de indiferença

Vamos evidenciar que, quando construirmos as poligonais conforme procedimento abaixo, elas serão representativas das poligonais de indiferença. O procedimento oriunda de uma medida DEA, denominada distância direcional [Joro et al. 1998], que será detalhada na seção 5.1 e aqui introduzida através da Figura 8. A medida, que também denominaremos h, é obtida da equação (7) abaixo:

$$h = (x_{1V} - x_{1E})/x_{1V} = (x_{2V} - x_{2E})/x_{2V} \qquad (7)$$

A solução DEA fornece o valor de h que, no caso de somente variáveis favoráveis, resulta:

h = 0 para DMUs situadas sobre a envoltória e 0 < h < 1 para todas DMUs internas.

Na equação (7), x_{1E} e x_{2E} representam as coordenadas do ponto P_E sobre a envoltória e x_{1V} e x_{2V} as coordenadas de um ponto interno P_V, construído pela extrapolação de um segmento de reta que une a origem ao ponto P_E conhecido.

Devemos observar que (7) é a própria expressão da equação (2), que apresentou h.

Podemos assim determinar h para o ponto P_V, por técnica DEA.

Temos assim a relação biunívoca entre a probabilidade α e h, de característica crescente. Todas as DMUs obtidas por extrapolação dos vértices da envoltória, com o mesmo valor de h, apresentam o mesmo valor α e assim, representam os vértices de uma poligonal sobre uma curva de indiferença, podendo ser considerada uma poligonal de indiferença. A partir desta consideração, o ponto P_0 da Figura 8, que, pela técnica DEA, apresenta o mesmo valor de h, por estar sobre a mesma poligonal, deverá ser considerado "tão inserido quanto" o ponto P_V, no domínio H, conceituando a aproximação da curva pela poligonal.

A técnica DEA não nos fornece a relação $\alpha = f^{-1}(h)$, possibilitando obter o valor efetivo da probabilidade associada. No entanto proporciona uma escala monotônica crescente entre h e α, descrevendo as poligonais internas. A generalização através do escore h_G, conforme seção 4.1, possibilitará comparar diversos grupos.

Em posse desta forma de medir, podemos agora iniciar os procedimentos para classificação, que serão expostos a seguir.

5 Procedimentos para classificação

Nesta seção apresentamos os procedimentos para classificação ao utilizarmos o modelo proposto na seção 4.

5.1 O Modelo DEA proposto

A formulação geral deste modelo [Joro et al, 1998], está representada no sistema de equações (8), em forma matricial, onde σ representa a distância direcional, que denominamos h, na equação (1). As variáveis X e Y são denominadas variáveis de *input* e de *output*, respectivamente.

$$\begin{aligned} &\max \quad \sigma + \varepsilon 1^T (s^+ + s^-) \\ &\text{s.t} \\ &Y\lambda - \sigma Y_0 - s^+ = Y_0 \\ &X\lambda + \sigma X_0 + s^- = X_0 \\ &1^T \lambda = 1 \\ &\lambda, s^-, s^+ \geq 0 \\ &\varepsilon \geq 0 \end{aligned} \qquad (8)$$

A solução deste sistema é a unidade virtual (Y_f, X_f), referência para (X_o, Y_o), representada exatamente pelo ponto $(Y\lambda, X\lambda)$ quando o vetor λ assumir os valores da solução. O sistema de inequações (8) permite uma particularização, do ponto de vista matemático, ou seja, a utilização somente de variáveis de *input*. Reescrevendo em forma não matricial, para $\varepsilon = 0$ e acertando a denominação σ para h:

$$\begin{aligned} &\max \ h \\ &\text{sujeito a} \\ &\sum_{j=1}^{n} \lambda_j x_{ij} + h x_{ij_o} \leq x_{ij_o} \qquad i = 1,2,.....,m \\ &\sum_{j=1}^{n} \lambda_j = 1 \\ &\lambda_j \geq 0 \end{aligned} \qquad (9)$$

As equações (10) explicitam a solução deste sistema:

$$h = \frac{(x_{ij_o} - x_{ij_f})}{x_{ij_o}} \qquad i = 1,2,.....,m \qquad (10)$$

Quando posicionamos as assíntotas sobre os eixos, a equação (10) reproduz a equação (2), restabelecendo as variáveis ajustadas conforme seção 4.2.

O PPL definido pelo sistema de equações (9), com solução dada pela equação (10), é a base para construirmos o modelo proposto na seção 4, para classificação, através das técnicas DEA.

5.2 Metodologia de classificação

O objetivo desta seção é, inicialmente, definirmos a medida de posição das DMUs das amostras de calibração de cada grupo, relativamente à sua fronteira e, posteriormente, considerarmos o método para classificar uma DMU teste, relativamente às k fronteiras. Este é.

O procedimento envolve uma série de passos, a saber:

- Efetuar o ajuste das variáveis originais, conforme seção 4.2.
- Executar o PPL dado pelo sistema de inequações (9), para cada população em estudo.
- Todas as DMUs de uma determinada população que apresentem h = 0 definirão sua fronteira. O PPL desta população será reformulado, a partir deste momento, contendo somente estas DMUs da fronteira (e uma particular DMU que será testada). Isto porque a contribuição de todas as outras DMUs deixa de ser relevante na determinação da medida h desta particular DMU de teste. Definimos assim o sistema representativo desta população. Este sistema será denominado PPL_e para esta população (PPL definido sobre envoltória desta população).
- Obteremos tantos PPL_e, quanto forem as populações em estudo.
- Caso se objetive avaliar uma nova entidade frente a uma certa população, ela o será através do PPL_e desta.
  O PPL_e para cada população, contendo m restrições, formadas pelas m variáveis, terá a seguinte forma, ao testar a DMU:

PPL_e

O índice zero será assumido por cada DMU a ser testada.
O índice j refere-se às entidades da envoltória da particular população em teste.

$$\begin{aligned} &\max\ h \\ &\text{sujeito a} \\ &\sum_{j\in\text{Envoltória}} \lambda_j x_{ij} + h x_{io} \le x_{io} \qquad i = 1, \ldots\ldots, m \quad \text{(m linhas)} \\ &\sum_{j\in\text{Envoltória}} \lambda_j = 1 \\ &\lambda_j \ge 0 \end{aligned} \tag{11}$$

Adotaremos o conceito de Comparação Bilateral [Cooper et al, 2000] em DEA, para testar uma nova DMU:

- A DMU é introduzida apenas como dado adicional x_{ij_o} (índice zero), nas inequações do PPL_e da população em que ela será testada, após ajuste das variáveis conforme seção 4.2.
- A DMU não deve ser introduzida como uma restrição adicional em $\sum_{j=1}^{n} \lambda_j x_{ij}$, nas inequações do PPL_e. Deste modo fica impossibilitado a inclusão desta DMU como uma entidade de envoltória, ou seja, impedindo modificação na forma atual da envoltória (índice $j \neq 0$).
  Este conceito foi proposto por [Seiford e Zhu, 1998], complementando a proposta de [Troutt et al, 1996], em teste de uma entidade frente a uma única população, para verificar se a entidade testada pertencia ou não à população. Em vez da distância direcional h, utilizou-se diretamente a medida DEA de eficiência CCR, obtendo-se eficiência superior a um (100%) caso não pertencesse, porem não resultou uma escala adequada de medida, para classificação frente a duas ou mais populações.

- Definição de h:
  a) A entidade (DMU) está sobre a envoltória atual ou no interior da região H (domínio)
  - Resulta PPL_e viável com $h \ge 0$
  b) Uma nova entidade está fora da região H
  - Resulta PPL_e viável com $h < 0$.

- Classificação de novas entidades

Uma vez definidas as fronteiras que delimitam as populações em estudo, as DMUs atuais e novas, de quaisquer das populações, poderão ser testadas para:

a) Verificação da possibilidade de uma região comum a duas ou mais populações.
   - DMUs que estão dentro da região comum, conforme sugerido na Figura 9, apresentarão $h_k \geq 0$ quando medidas em relação a cada uma das k fronteiras, das k populações. Observar que na Figura 9, a região comum está apresentada com as variáveis originais, antes do ajuste.
   - DMUs que estão fora de uma das fronteiras apresentarão $h_k < 0$ em relação a esta fronteira.

b) Classificação de nova DMU através dos respectivos PPL_e de cada população (Equações 11), determinando em qual população esta DMU deve ser classificada, caso a mesma não esteja localizada na região comum. A classificação obedece o Critério do Maior h_G, ou seja, a nova DMU será classificada no grupo onde o PPL_e resultar maior h_G, após generalizarmos o valor de h, conforme seção 4.1.

Na representação geométrica da Figura 9 temos duas populações onde as variáveis V_1 e V_2 se comportam como variáveis favoráveis para a População I (a redução delas abaixo de um limite mínimo inviabiliza a ocorrência de elementos da População I) e como variáveis desfavoráveis à População II (o aumento delas acima de um limite máximo inviabiliza a ocorrência de elementos da População II).

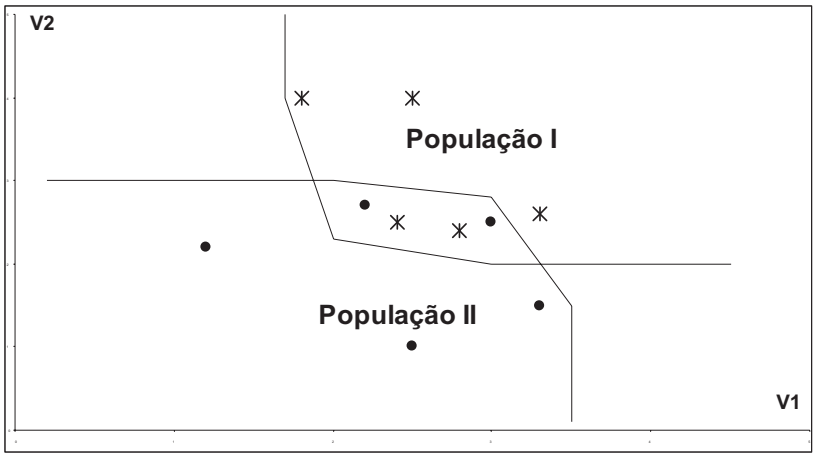

Figura 9 – Distribuição de duas Populações. Variáveis não ajustadas

A Figura 9 apresenta uma zona comum, onde é possível, para uma DMU, determinarmos um valor de h_G, em relação à cada uma das duas envoltórias, separadamente.

Para k grupos, utilizamos h_{Gk} para explorar a possibilidade de classificar uma nova DMU em uma das k populações (aquela que apresentar maior valor para h_{Gk}**)**, em um ambiente m dimensional, representado pelas m variáveis. Um grande facilitador resulta da existência de diversos algoritmos para solução DEA, possibilitando obter h_k para todas as DMUs, em relação a todas as fronteiras, sem grandes dificuldades, agilizando o processo de classificação. A obtenção de h_{Gk} é resultado de uma simples transformação algébrica sobre h_k.

6 *Outliers* e aprimoramento dinâmico das fronteira

A fronteira de cada grupo é de importância fundamental no modelo proposto. Assim, o seu reconhecimento deve ser considerado sob diversos aspectos. Este fato é tópico desta seção.

6.1 O problema de *outliers* no estabelecimento das fronteiras

No nosso caso o problema adquire fundamental importância, pois a existência prévia de um *outlier* na amostra original pode alterar totalmente o delineamento das fronteiras e conseqüentemente o processo de classificação.

O conceito de *outlier* pode ser considerado dentro do exposto na Figura 10.

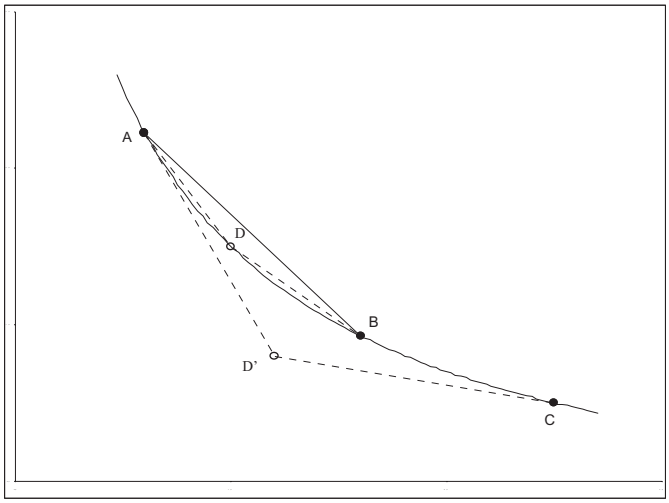

Figura 10 – Presença de *outliers*

- Supomos que a poligonal formada pelas DMUs **A**, **B**, **C** garantam a melhor representação disponível da curva envoltória, obtida da amostra original.
- A DMU **D'** é um *outlier*, deslocando a envoltória de sua real posição. Pode tratar-se de uma DMU de outra população, não sendo recomendável sua introdução na envoltória.

Trata-se de um problema ligado à análise de *outliers* em DEA, visto estarmos nos apoiando nesta técnica para definição da fronteira. A identificação de *outliers* em DEA é um problema abordado com bastante profundidade [Pastor, Ruiz and Sirvent, 1999].

[Forni, 2002] fez um longo e bem estruturado esforço dirigido à identificação de estudos efetuados com enfoque sobre *outliers*.

[Troutt et al,1996], em um estudo para concessão de crédito, através da técnica DEA, propôs que a fronteira fosse determinada com a orientação de especialistas financeiros, ponderando sobre a massa original de dados, evitando assim a introdução de *outliers*, em análise caso a caso das empresas da amostra.

"*Outliers*" muito fora do centro do agrupamento podem ser identificados através da distância de Mahalanobis [Duda et al, 2001], supondo em primeira aproximação que ocorra uma distribuição normal das DMUs, mesmo conhecendo a hipótese original de estarmos determinando a fronteira de uma distribuição truncada. Porém, "*outliers*" próximos da envoltória não serão identificados e, estes são os mais preocupantes, pois modificam a fronteira.

6.2 A possibilidade de "*feed back*"

Objetivamos, quando do desenvolvimento do modelo, atender situações de uso continuado, por exemplo, na análise de solvência de empresas para concessão de crédito, de modo que a introdução de novas informações seja inerente ao processo. Estas informações devem contribuir positivamente para ajustes, levando ao estabelecimento de soluções mais confiáveis em decisões subseqüentes.

O aprimoramento das fronteiras, com a introdução de novas DMUs, passa a ser fundamental, possibilitando a melhora de performance em novas decisões de classificação. A possibilidade de uma nova DMU produzir uma redefinição para melhor da fronteira, ou modificar para pior por se tratar de um *outlier*, deve levar ao aprofundamento de análises dirigidas a esta identificação.

Possibilidades de introdução de novas DMUs na fronteira foram apresentadas [Almeida, 2000], sugerindo que uma DMU pode ser considerada de fronteira, se sua introdução não elimina atuais DMUs de fronteira e ainda melhora o processo de classificação de uma massa de dados destinada à verificação do modelo, como a DMU **D** da Figura 10.

Sugerimos que a DMU **D** produz melhoria no delineamento da envoltória, sendo viável sua introdução. Assim, esta possibilidade de *feed back* constitui nossa resposta para garantir atualizações constantes nos parâmetros do modelo.

7 Um particular estudo de caso

Vamos aqui apresentar um particular problema envolvendo dois grupos onde, como base de dados, utilizamos um exemplo desenvolvido em livro [Hair et al, 1998]. Este exemplo, constituído de 14 variáveis, foi adotado ao longo do livro para apresentar diversas técnicas.

Para o problema de classificação baseado em análise discriminante, os autores procuraram discriminar a variável X_{11}, a partir das variáveis discriminatórias X_1, X_3 e X_7, identificadas como as mais adequadas para este fim.

Neste caso procura-se interpretar como empresas clientes efetuam suas compras, ou seja, se baseadas apenas nas características específicas do produto ou se empregam a análise do valor total da compra. A empresa fornecedora pode então alterar apresentações de vendas e benefícios oferecidos, conforme a característica de cada empresa compradora, melhorando a performance de vendas. Assim:

X_1 – Velocidade de entrega dos pedidos.

X_3 – Flexibilidade de preços. Como os representantes avaliam a política de preços da empresa.

X_7 – Qualidade do produto.

X_{11} – Variável com duas categorias. $X_{11} = 0$ se o comprador se baseia apenas nas características específicas do produto (identificado como comprador do grupo 0) e $X_{11} = 1$ se emprega a análise do valor total da compra (comprador do grupo 1).

Da amostra, constituída de 100 observações, os autores utilizaram parte (60 observações) para calibrar o modelo e outra parte para validar o modelo (40 observações).

Utilizando o mesmo conjunto de dados, podemos adotar, em nosso modelo, que as variáveis de classificação apresentam características opostas em relação aos dois grupos, ou seja, quando favoráveis em um grupo resultam desfavoráveis em outro e, vice versa. Tal não se deve necessariamente à relação existente entre os dois grupos, mas pela simples possibilidade de melhor explorarmos a característica discriminatória destas variáveis. Assim:

As variáveis X_1 e X_3 são de fato favoráveis ao grupo 1, pois os compradores deste grupo analisam a compra em seu todo e não apenas as especificações do produto. Quanto maior seu valor, mais atrai o comprador. Para o grupo 0 elas são de importância menor e, vamos considerá-las como "desfavoráveis", para melhor explorar a característica discriminatória destas pois, nos valores muito altos, se encontram os compradores do

grupo 1, podendo supor *outlier* a presença de um comprador do grupo 0. A variável X_7 é extremamente favorável ao grupo 0 e, pelas mesmas razões apresentadas para X_1 e X_3, vamos considerá-la como "desfavorável" ao grupo 1. A construção desta fronteira, melhor ilustrada no início da seção 7.1, não é uma restrição ao uso do modelo pois, de uma maneira geral, as variáveis são idealizadas probabilisticamente ilimitadas, mas na prática são limitadas superior e inferiormente. Porém, a fronteira assim obtida, pode ser conceitualmente criticada, face ao desenvolvimento proposto na seção 3, sendo defensável devido aos bons resultados práticos.

A Tabela 1 apresenta os resultados obtidos com os dados de calibração, para análise discriminante, conforme apresentado pelos autores.

A Tabela 2 apresenta os resultados obtidos com os dados de calibração, conforme modelo proposto.

Tabela 1 - Amostra de Calibração - Análise Discriminante

		Classificado			
		G0	G1	Total	% Acerto
Observado	G0	21	1	22	95,5%
	G1	4	34	38	89,5%
	Total	25	35	60	91,7%

Tabela 2 - Amostra de Calibração - Modelo proposto

		Classificado			
		G0	G1	Total	% Acerto
Observado	G0	21	1	22	95,5%
	G1	2	36	38	94,7%
	Total	23	37	60	95,0%

A Tabela 3 apresenta os resultados obtidos sobre a amostra de validação, ao utilizarmos nosso modelo. Verifica-se que uma observação do grupo 0 e outra do grupo 1 não puderam ser classificadas, pois apresentaram $h_G < 0$ relativamente às duas fronteiras, ou seja, se apresentaram fora dos dois grupos.

Tabela 3 - Amostra de Validação - Fronteiras originais

		Classificado			
		G0	G1	Sem Classificação	Total
Observado	G0	14	3	1 (*)	18
	G1	2	19	1 (*)	22
	Total	16	22	2	40

(*) DMU 57 do grupo 0 e DMU 76 do grupo 1

Este é o momento de analisarmos se estas duas observações tem característica *outlier* ou podem ser utilizadas para aperfeiçoamento das envoltórias. Deste modo, o próprio modelo evidencia quais novas DMUs devem ser utilizadas na análise.

No caso da DMU 57, a introdução na envoltória em nada alterou as DMUs originalmente definidoras da envoltória. Também em nada alterou a classificação das amostras de calibração e de validação, salvo o fato de agora, a mesma se mostrar incluída no grupo 0, no teste de validação.

Já a inclusão da DMU 76 colocou as DMUs 82 e 84, originalmente de envoltória, ligeiramente fora da mesma, com $h_{G82} = 0{,}02$ e $h_{G84} = 0{,}03$. Quanto ao resto, em nada alterou a classificação das amostras de calibração e de validação, salvo o fato de agora a mesma se mostrar incluída no grupo 1, no teste de validação.

Vamos considerar então a possibilidade de atualizarmos as fronteiras. O resultado na amostra de validação se encontra nas tabelas 4 e 5, evidenciando resultado positivo para o modelo proposto, frente análise discriminante.

Tabela 4 - Amostra de Validação - Análise Discriminante

		Classificado			
		G0	G1	Total	% Acerto
Observado	G0	15	3	18	83,3%
	G1	3	19	22	86,4%
	Total	18	22	40	85,0%

Tabela 5 - Amostra de Validação – Modelo proposto, fronteiras atualizadas

		Classificado			
		G0	G1	Total	% Acerto
Observado	G0	15	3	18	83,3%
	G1	2	20	22	90,9%
	Total	17	23	40	87,5%

Finalmente estamos apresentando, na Tabela 6, os resultados do modelo, quando aplicado a toda a amostra de validação com fronteiras atualizadas.

Tabela 6: Resultados sobre a amostra de verificação

				Classificação	
Amostra de Análise	Condição real	Front. Grupo 0 h_{G0}	Front. Grupo 1 h_{G1}	Critério: Maior h_G	Resultado
3	G0	0,33	< 0	G0	
4	G0	0,53	< 0	G0	
10	G0	0,21	< 0	G0	
27	G0	0,45	< 0	G0	
30	G0	0,05	< 0	G0	
34	G0	0,41	< 0	G0	
35	G0	< 0	0,48	G1	ERRO
37	G0	0,67	< 0	G0	
40	G0	1,07	< 0	G0	
41	G0	0,87	< 0	G0	
57	G0	0	< 0	G0	
60	G0	0,35	< 0	G0	
75	G0	0,54	< 0	G0	
83	G0	1,27	< 0	G0	
85	G0	< 0	0,48	G1	ERRO
87	G0	< 0	0,48	G1	ERRO
94	G0	1,11	< 0	G0	
98	G0	1,14	< 0	G0	
9	G1	< 0	0,43	G1	
16	G1	< 0	1,65	G1	
18	G1	< 0	0,70	G1	
19	G1	< 0	0,85	G1	
21	G1	< 0	0,75	G1	
22	G1	< 0	1,13	G1	
38	G1	< 0	0,48	G1	
44	G1	< 0	1,90	G1	
46	G1	< 0	0,99	G1	
55	G1	< 0	1,09	G1	
56	G1	0,35	< 0	G0	ERRO
62	G1	< 0	1,71	G1	
63	G1	< 0	0,49	G1	
64	G1	< 0	0,06	G1	
66	G1	< 0	0,66	G1	
69	G1	< 0	1,39	G1	
74	G1	< 0	0,48	G1	
76	G1	< 0	0,00	G1	
77	G1	< 0	1,11	G1	
78	G1	< 0	1,48	G1	
91	G1	0,29	< 0	G0	ERRO
100	G1	< 0	0,93	G1	

7.1 Considerações sobre as fronteiras no Estudo de Caso

Como todo modelo desenvolvido para classificação, existem limitações a serem observadas.

No modelo aqui desenvolvido, pode ocorrer que dois grupos apresentem todas as variáveis com as mesmas características, como na Figura 11, onde as duas variáveis são favoráveis aos dois grupos. Neste caso pode ocorrer que a fronteira de um grupo fique inserida na região domínio do outro grupo e, se a variância do grupo interno (grupo b) for maior ou igual a do grupo externo, o critério da probabilidade não se aplica, pois levaria à classificação de todas DMUs no grupo externo (grupo a). Com relação ao grupo interno, conseguiria apenas identificar as DMUs fora da região H_b, domínio deste grupo.

Uma alternativa seria considerar que, para o grupo externo, ocorre também uma limitação superior das variáveis, fornecendo outra fronteira (tracejada na Figura 11), generalizando a conceituação de envoltória do grupo externo, adotando que as variáveis são limitadas, inferior e superiormente. Trabalharíamos com esta nova fronteira para o grupo externo. Na seção 7, apesar de a situação não se ter configurado tão drástica, adotamos este procedimento, para obtermos a maior eficiência discriminatória das variáveis, conforme as justificativas que foram apresentadas.

Das considerações acima verificamos que o modelo, com as fronteiras originalmente propostas na seção 3, atinge eficiência máxima na discriminação de dois grupos, quando estes apresentam características opostas. Um exemplo desta situação são os casos de "*Credit Scoring*", onde as variáveis favoráveis ao grupo de empresas solventes se apresentam desfavoráveis ao grupo de empresas insolventes e vice-versa, dispensando considerações adicionais sobre construção de novas envoltórias.

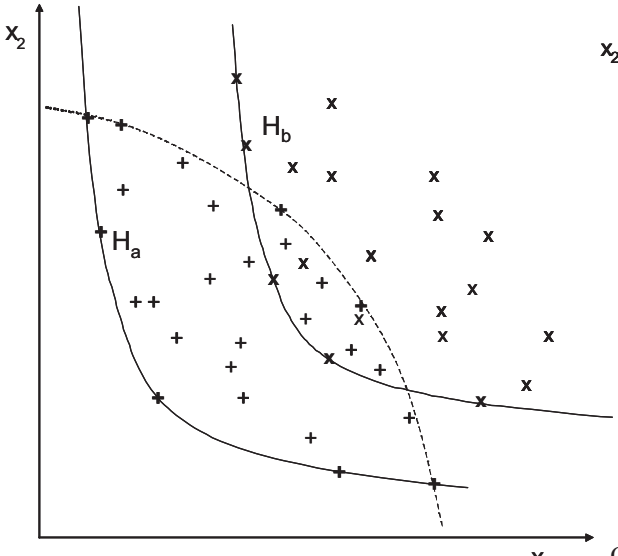

Figura 11 – Envoltória discriminante

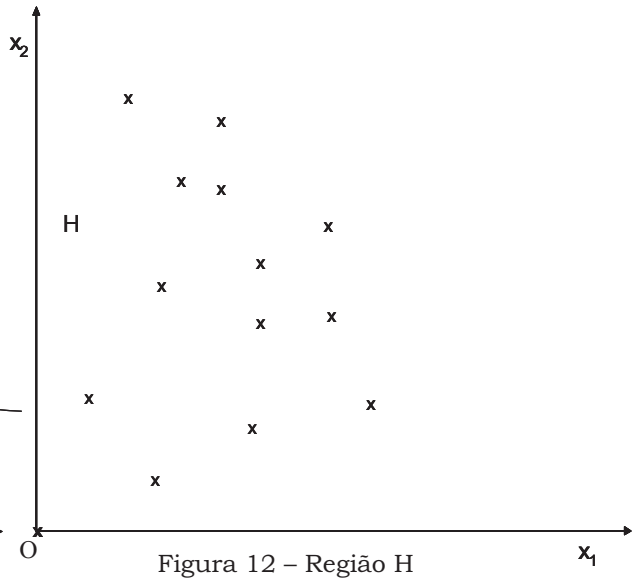

Figura 12 – Região H

Temos outra limitação, ao utilizamos a solução aproximada, quando o valor mínimo de todas as variáveis ajustadas ocorre para a mesma DMU. A região H incluirá a origem no processo de ajuste das variáveis, conforme Figura 12. Neste caso a medida h não pode ser definida na equação 7.

Porém pretendemos realçar que desenvolvemos um modelo para atender ao requisito de existência de fronteira, originada no fato de as variáveis serem limitadas. Nos preocupamos em evitar o erro de classificação de DMUs fora da fronteira, possibilitando a tomada de decisão em situações limites. Procuramos minimizar o erro Tipo 2 [Costa Neto,2002], ao estabelecermos um limite para tomada de decisão. Por exemplo, em casos de *Credit Scoring*, para uma empresa espacialmente localizada fora da fronteira de solvência, ou seja, tratando-se efetivamente de uma empresa condenada à insolvência

(caso a fronteira esteja corretamente definida), não estaríamos incorrendo em erro Tipo 2, ou seja, aceitarmos que a empresa seja solvente, sendo falsa esta hipótese, apoiando o fornecimento de crédito à mesma.

8 Conclusões

A apresentação de um caso envolvendo dois grupos, na seção 7, possibilitou a verificação de pontos importantes considerados durante o desenvolvimento do modelo, como:

- capacidade de identificação de *outliers*,
- independência dos parâmetros do modelo face ao tamanho das populações de cada grupo,
- evitar erro de classificação de DMUs fora da fronteira ($h < 0$).
  O erro Tipo 2, conforme exposto em 7.1, é minimizado, ao ficar estabelecido um limite real (as fronteiras dos grupos)para tomada de decisão,
- possibilidade de melhoria das fronteiras com introdução de dados novos.

A possibilidade de ajustes contínuos da fronteira originalmente estabelecida, conforme novas informações ocorram, melhorando a performance na tomada de novas decisões, é um ponto altamente positivo do modelo.

Deste modo, a possibilidade de levarmos as fronteiras em consideração, ou seja, efetuarmos as análises partindo da fronteira para o interior (e não do centro do agrupamento para fora), através da distância direcional h, é uma alternativa para classificarmos adequadamente as DMUs que estão próximas, mas fora das fronteiras. Estas DMUs são normalmente o foco para uma boa classificação, pois aquelas localizadas próximas aos centros dos diversos agrupamentos normalmente são corretamente classificadas por todas as técnicas de classificação.

9 Referências Bibliográficas

Almeida, H.R. [2000] Análise de Envoltória de Dados na Decisão de Concessão de Crédito, Anais do XXXII Simpósio Brasileiro de Pesquisa Operacional, Outubro, Viçosa MG.

Cooper W.W., Seiford L.M., Tone K. [2000] Data Envelopment Analysis – A Comprehensive Text with Models, Aplications, References and DEA-Solver Software, Boston: Kluwer Academic Publishers, Second Printing.

Costa Neto P.L. [2002] Estatística, Editora Edgard Blücher Ltda

Duda R.O., Hart P.E., Stork D.G. [2001] Pattern Classification, John Wiley & Sons Inc.

Forni A.L.C. [2002] On the Detection of Outliers in Data Envelopment Analysis Methodology, Tese de Mestrado, Instituto Tecnológico de Aeronáutica, São José dos Campos, Brasil

Hair J.F.J., Anderson R.E., Tatham L.T., Black W.C. [1998] Multivariate Data Analysis, Prentice Hall, Fifth Edition

Joro T., Korhonen P., Wallenius J. [1998] Structural Comparison of Data Envelopment Analysis and Multiple Objective Linear Programming, Manegement Science 44:7, 962-970

Meyer P.L. [2000] Probabilidade Aplicada à Estatística, 2ª Edição, Editora LTC SA

Pastor T. J., Ruiz L.J. and Sirvent I. [1999] A statistical test for detecting influential observation in DEA, European Journal of Operational Research 115 (1999) 542-554

Seiford L.M. and Zhu J. [1998] An Acceptance System Decision Rule With Data Envelopment Analysis, Computers Ops Res. Vol.25 Nº 4, pp 329-332

Troutt M. D., Rai A., Zhang A. [1996] "The Potential use of DEA for Credit Applicant Acceptance Systems, Computers Ops Res. Vol.23 Nº 4, pp 405-408